Understanding the Message of Images

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von
Lydia Weiland
aus Velbert

Mannheim, 2018

Dekan:	Dr. Bernd Lübcke, Universität Mannheim
Referent:	Prof. Dr. Simone Paolo Ponzetto, Universität Mannheim
Korreferent:	Prof. Dr. Wolfgang Effelsberg, Universität Mannheim

Tag der wissenschaftlichen Aussprache: 08. Juni 2018

Abstract

We investigate the problem of understanding the message (gist) conveyed by images and their captions as found, for instance, on websites or news articles. To this end, we propose a methodology to capture the meaning of image-caption pairs on the basis of large amounts of machine-readable knowledge that have previously been shown to be highly effective for text understanding. Our method identifies the connotation of objects beyond their denotation: where most approaches to image or image-text understanding focus on the denotation of objects, i.e., their literal meaning, our work addresses the identification of connotations, i.e., iconic meanings of objects, to understand the message of images. We view image understanding as the task of representing an image-caption pair on the basis of a wide-coverage vocabulary of concepts such as the one provided by Wikipedia, and cast gist detection as a concept-ranking problem with image-caption pairs as queries. Specifically, we approach the problem using a pipeline that: i) links detected object labels in the image and concept mentions in the caption to nodes of the knowledge base; ii) builds a semantic graph out of these 'seed' concepts; iii) applies a series of graph expansion and clustering steps on the original semantic graph to include additional concepts and topics within the semantic representation; iv) combines several graph-based and text-based features into a concept ranking model that pinpoints the gist concepts. Understanding the gist can be useful for tasks, such as image search and recommending images for texts.

As gist detection is a novel task, to the best of our knowledge, there is no dataset available. Thus, we create a dataset allowing for simultaneous evaluation of literal and non-literal image-caption pairs. The gold standard gist concepts are from a common knowledge base (Wikipedia) and the provided ranks are detailed with levels 0 to 5, which supports various benchmarking tasks, e.g., ranking according to different levels of granularity and classification. Furthermore, as our proposed gist detection pipeline touches on different research areas, we provide a detailed gold standard for each of our pipeline steps, such as entity linking or object detection in the images. Our gist detection pipeline is evaluated in a detailed ablation study, investigating aspects of twelve different research questions. These are elaborated in the evaluation section via human-assessment or cross-validation and provide detailed insights into the gist of image-caption pairs. Furthermore, we show in an end-to-end

setting the feasibility of state-of-the-art methods combined with our gist-detection pipeline and point to future research directions.

Our experiments show that the candidate selection and ranking of gist concepts is a more difficult problem for non-literal image-caption pairs than for literal image-caption pairs. Furthermore, we demonstrate that using features and concepts from both modalities (image and caption) improves the performance for all types of pairs – a finding which is in line with results from research on multimodal approaches for other related tasks. Additionally, a feature ablation study shows the complementary nature and usefulness of different types of features, which are collected from different kinds of semantic graphs of increasing richness. Finally, we experimented with a state-of-the-art image object detector and caption generator to evaluate the performance of an end-to-end solution for our task. The results indicate that using state-of-the-art open-domain image understanding provides us with an input that is good enough to detect gist concepts of image-caption pairs, with nearly half of the predicted gist concepts being relevant. However, it also demonstrates that improved object detectors could avoid a drop of 38% mean-average precision. Additionally, the caption contains useful hints especially for non-literal pairs.

Gist image identification is a small, yet arguably crucial part of the much bigger task of interpreting images beyond their denotation. Within a use case scenario of an established research problem, we show that gist detection in the form of concept ranking is useful for downstream tasks such as multimedia indexing, in that it outperforms shallow and deep approaches. Finally, we conclude that it could be useful also for image search and recommendation.

Kurzfassung

Wir untersuchen die Problematik des Verstehens der Kernbotschaft (Kern), die durch Bilder und ihren Bildunterschriften, wie sie z.B. auf Webseiten oder in Nachrichtenartikeln zu finden sind, vermittelt wird. Zu diesem Zweck präsentieren wir eine Methodik zur Erfassung der Bedeutung von Bild-Bildunterschriften-Paaren auf Basis von großen Mengen maschinenlesbaren Wissens, welches sich in der Vergangenheit für Textverständnis als sehr effektiv erwiesen hat. Unsere Methode identifiziert die Konnotation von Objekten jenseits ihrer Denotation: Während die meisten Ansätze zum Bild- oder Bild-Text-Verständnis sich auf die Benennung von Objekten, d.h. ihrer wörtlichen Bedeutung, konzentrieren, beschäftigt sich unsere Arbeit mit der Identifikation von Konnotationen, d.h. ikonischen Bedeutungen von Objekten, um die Botschaft von Bildern zu verstehen. Wir betrachten das Bildverstehen als die Aufgabe, ein Bild-Bildunterschriftenpaar auf Basis eines umfangreichen Vokabulars von Konzepten, wie sie in Wikipedia bereit gestellt werden, zu repräsentieren und gehen das Bild-Kern Verstehens als Konzept-Ranking Problem mit Bild-Bildunterschriften-Paaren als Abfrage an.

Konkret bedeuted dies, dass wir uns dem Problem mit Hilfe einer Pipeline nähern, die: i) die gefundenen Objektbeschriftungen im Bild und Konzepterwähnungen in der Bildunterschrift mit Knoten der Wissensbasis verknüpft; ii) die einen semantischen Graphen aus diesen 'Kern' Konzepten erstellt; iii) die eine Reihe von Graphen-Erweiterungs und -Clusterings Schritte auf dem originalen semantischen Graphen anwendet, um zusätzliche Konzepte und Themen in die semantische Repräsentation des Paares einzubeziehen; iv) die mehrere graph- und textbasierte Merkmale zu einem Konzept-Ranking-Modell zusammenfasst, welches die wesentlichen Konzepte, die die Kernaussage des Bildes aufzeigen, bestimmt. Das Verstehen der Kernaussage eines Bildes ist nützlich für die Bildersuche oder die Bildempfehlung zu Texten.

Da das Verstehen der Kernaussage ein neue Problemstellung ist, existieren - nach unserem besten Wissen - hierzu keine Datensätze. Folglich erstellen wir einen Datensatz, der die gleichzeitige Auswertung von literalen und nicht-literalen Bild-Bildunterschriften-Paaren ermöglicht. Die Goldstandard Kernkonzepte stammen aus einer allgemeinen Wissensbasis (Wikipedia) und ihre Rankinglevel sind detailliert mit den Stufen 0 bis 5 gesetzt. Dieser

Goldstandard unterstützen verschiedene Benchmarking-Aufgaben, z.B. Ranking nach verschiedene Stufen der Granularität und Klassifizierung. Außerdem, da unsere vorgeschlagene Pipeline zum Verstehen der Kernaussage, verschiedene Forschungsgebiete vereint, bieten wir einen detaillierten Goldstandard für jeden einzelnen unserer Pipeline-Schritte, wie z.B. Entity-Linking oder Objekterkennung in den Bildern.

Unsere Pipeline zum Verstehens der Kernaussage wird in einer detaillierten Ablationsstudie evaluiert und untersucht Aspekte von zwölf verschiedene Forschungsfragen. Die Evaluationen, auf die im Evaluationsteil näher eingegangen wird, werden entweder mit Hilfe von manueller Annotation oder Kreuz-Validierung durchgeführt. Sie geben detaillierte Einblicke in das Verstehen der Kernaussage von Bild-Bildunterschriften-Paaren. Darüber hinaus zeigen wir in einem End-to-End-Aufbau die Machbarkeit von State-of-the-Art Methoden in Kombination mit unserer Pipline zum Verstehen der Kernaussage und weisen auf zukünftige Forschungsrichtungen hin. Unsere Experimente zeigen, dass die Auswahl und das Ranking der Kandidaten für die Konzepte zur Repräsentation der Kernaussage ein schwierigeres Problem für nicht-literale Paare als für literale Paare ist. Dennoch demonstrieren wir, dass die Verwendung von Merkmalen und Konzepten aus beiden Modalitäten (Bild und Text der Bildunterschrift) die Performance für alle Arten von Paaren verbessert. Dies ist ein Befund, der mit Ergebnissen aus der Forschung zu multimodalen Ansätzen für andere verwandte Aufgaben übereinstimmt. Zusätzlich zeigt die Ablationsstudie, die komplementäre Natur und Nützlichkeit der verschiedenen Arten von Merkmalen, welche aus verschiedenen Arten von semantischen Graphen mit zunehmendem Informationsgehalt gesammelt werden. Schließlich experimentierten wir mit einem weitreichend akzeptierten Bildobjektdetektor und Bildunterschriften-Generator, um die Leistungsfähigkeit einer End-to-End-Lösung für unsere Aufgabe zu evaluieren.

Die Ergebnisse der Nutzung der modernen Open-Domain Verfahren zum Bildverständnis, deuten darauf hin, dass diese Informationen lieferen, die gut genug sind, um die grundlegenden Konzepte der Kernaussage von Bild-Bildunterschriften-Paaren zu erkennen. Hierbei sind fast die Hälfte der vorhergesagten Kernaussage Konzepte relevant. Darüber hinaus zeigen wir aber auch, dass verbesserte Objektdetektoren einen Rückgang der mittleren Genauigkeit um 38% vermeiden könnten. Zusätzlich enthält die Bildunterschrift nützliche Hinweise speziell für nicht-literale Paare.

Die Identifikation der Kernaussagen von Bildern ist ein kleiner, aber wohl entscheidender Teil des viel größeren Problems Bilder jenseits ihrer literalen Bedeutung zu interpretieren. Im Rahmen eines Anwendungsfalls eines etablierten Forschungsbereichs, zeigen wir, dass die Erkennung der Kernaussage in Form eines Konzept-Rankings sinnvoll für nachgelagerte Aufgaben, wie z.B. die Multimedia-Indizierung ist, da diese Methodiken aus dem Shallow

und Deep-Learning übertrifft. Abschließend kommen wir zu dem Fazit, dass das Verstehen der Kernaussage eines Bildes auch für die Bildsuche und die Empfehlung von Bildern zu beispielsweise Texten, nützlich sind.

Figure 1 "The mind loves the unknown. It loves images whose meaning is unknown, since the meaning of the mind itself is unknown." René Magritte

Contents

List of Publications

The work presented in this thesis has been published before in the proceedings of different conferences this may also include figures, tables, and algorithms. For all publications the authors of this thesis was the key contributor of the work presented in both the publications and this thesis.

- Weiland et al. [2014]: Weiland, L., Effelsberg, W., and Ponzetto, S. P. (2014). Weakly supervised construction of a repository of iconic images. In *Proceedings of the Workshop on Vision and Language 2014 (VL '14) at the 25th International Conference on Computational Linguistics (COLING '14)*.

- Weiland et al. [2015]: Weiland, L., Dietz, L., and Ponzetto, S. P. (2015). Image with a message: Towards detecting non-literal image usages by visual linking. In *Proceedings of the 2015 EMNLP Workshop on Vision and Language (VL'15)*, pages 40–47.

- Weiland et al. [2016]: Weiland, L., Hulpuş, I., Ponzetto, S. P., and Dietz, L. (2016). Understanding the message of images with knowledge base traversals. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016*, Newark, DE, USA, September 12- 6, 2016, pages 199–208.

- Weiland et al. [2017]: Weiland, L., Hulpuş, I., Ponzetto, S. P., and Dietz, L. (2017). Using object detection, nlp, and knowledge base to understand the message of images. In *Lecture notes in computer science MultiMedia Modeling : 23rd International Conference, MMM 2017*, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part II; 405-418, Springer International Publishing, Cham, 2017.

- Weiland et al. [2018a]: Weiland, L., Hulpuş, I., Ponzetto, S. P., Effelsberg, W., and Dietz, L. (2018). Knowledge-rich Image Gist Understanding Beyond Literal Meaning. In *Journal on Data & Knowledge Engineering, 2018*, Elsevier, 2018.

- Weiland et al. [2018b]: Weiland, L., Ponzetto, S. P., Effelsberg, W., and Dietz, L. (2018). Understanding the Gist of Images - Ranking of Concepts for Multimedia Indexing. In *arXiv preprint*, arXiv:1809.08593, 2018.

List of Figures

List of Tables

Chapter 1

Introduction

The goal of this thesis is to understand the message *(gist)* of images. Transmitting a meaning with the use of images is probably as old as humankind. The intention of conveying a meaning is very faceted and is strongly influenced by the answer to the question, whether one wants to convey a literal (a denotation of an object) or a non-literal (a connotation of an object) meaning. Furthermore, the intention of a sender of a meaning is not necessarily equivalent to the perception of the receiver of the meaning. To frame the image into context and to narrow the risk of potential misconceptions, we focus on images in the context of short texts, i.e., captions.

An object depicted in an image can be used in its *literal* meaning, where the intention is to refer to itself, e.g., an image of a coffee cup sleeve conveys the literal meaning coffee cup sleeve (cf. Figure 1.1a). Following a different intention, a coffee cup sleeve can be used in terms of a *non-literal* meaning, where the object triggers some association to a visually non-recognizable higher-level meaning, e.g., the coffee cup sleeve triggers the association of the amount of waste produced by disposable food packaging, where a coffee cup with its sleeve is an example of (cf. Figure 1.1b). This type of association is commonly used and learned, e.g., in that it is transmitted by media or appears in knowledge bases, such as Wikipedia. However, an image of a coffee cup can also trigger a personal association, which is not part of common knowledge, e.g., this coffee cup reminds me of my weekly coffee with my friend Anna. To decode the message of an image, besides common knowledge a person has also access to knowledge created from own experiences, on the cultural and social background.

Moreover, the intended meaning and the knowledge someone makes use of are strongly influenced by the time periods (e.g., during the industrial revolution smoking stacks are probably rather associated with productivity and useful inventions than with global warming). Referring to our example, based on the image alone it remains unclear, whether a literal

(a) COFFEE CUP SLEEVE:"Coffee cup sleeve on a coffee cup, sleeve makes it easy to hold hot drinks."

(b) DISPOSABLE FOOD PACKAGING:"According to the U.S. Environmental Protection Agency, paper and plastic food-service packaging discarded in the country's municipal solid waste stream accounted 1.3 percent in 2007 (by weight) of municipal solid waste."

Figure 1.1 Image and context, e.g., captions, convey a meaning. (a) Literal example about coffee cup sleeves. (b) Non-literal example about environmental concerns related to disposable food packaging. (a, b: https://en.wikipedia.org/wiki/File:Coffee-cup-sleeve.jpg, Nirzar, CC-BY-SA 4.0, last accessed: 08/30/2017).

or non-literal meaning applies. Additionally, whether one uses associations created from personal or common knowledge, cannot be disambiguated by the image alone. Consequently, it turns out, context helps to disambiguate the image. In our example, the image is used within paragraphs raising associations to different topics: coffee cup sleeves and environmental concerns related to disposable food packaging. Images can be contextualized by other images or by media of other types, e.g., by captions or surrounding text. Finally, the context can be given by several combinations of media types.

Within our research the images are framed in the context of short text. We leave the investigation of context given by other images to future research. Furthermore, we focus on associations that can be derived from commonly accessible knowledge, where we limit the knowledge to the time periods 2015-2017 (duration of this research) and to English speaking annotators from Europe.

What sounds like a typical problem from philosophy, communication scientists, or linguists, becomes obviously crucial with respect to computer science, e.g., cognitive systems that will be capable of understanding images.

1.1 Motivation

Image understanding is a major goal in the research areas of computer vision, multimedia indexing and modeling. However, the main emphasis has been put on the literal meanings of images or descriptive image-text pairs [Bernardi et al., 2016; Hodosh et al., 2013]. This focus becomes more clear, when reviewing the titles of publications in the area of multimodal modeling, i.e., indexing or classification, or cross-modal modeling and retrieval, where the terms 'descriptive' and 'descriptions' occur very often (e.g., [Elliott and Keller, 2013; Karpathy and Li, 2015; Kulkarni et al., 2011; Mitchell et al., 2012]). This focus is even more underlined by the characteristics of existing image or multimodal datasets, with their corresponding texts and gold standard annotations (cf. Section 3 for a detailed review on Data Resources). Most of the images are accompanied with descriptive texts and/or captions. The other datasets are collected from social media platforms, e.g., Flickr, provided with captions, without explicitly given classification of the type of affiliated caption or text. As the category of these platforms already suggests, the caption sometimes contains personal knowledge making it nearly unfeasible decoding without having access to this kind of information.

The gold standard annotations aim at the goals of generating or retrieving descriptive captions or retrieving images that best match textual descriptions. A classification of the type of pair is not necessarily needed to achieve these goals. The observation about the scope of the annotations holds true independent from the type of methodology (cf. Chapter 2.3, Table 2.1).

We outline the pipeline of our approach to understand the message of literal and non-literal images by the use of our introductory examples (cf. Figure 1.1), it starts with understanding (detection, recognition, and annotation) of objects and the scenery (interactions of objects with each other, background, etc.) depicted in an image, thus, a literal assessment, e.g., coffee-cup and table. Similarly, the caption text is processed with a standard natural-language processing pipeline to detect nouns and noun phrases. The object and scenery names from the image and the nouns and noun phrases of the caption are candidate concept mentions. The second step of our pipeline is the linking of these candidate concept mentions to candidates from the reference knowledge base. With graph-based traversal strategy, we expand the initial set of candidate concepts in step three. In step four, we apply a semantic relatedness-based strategy of additional concepts, which are in semantically close proximity of the previously collected and initial concepts, conduct a clustering and a final focused selection of concepts that potentially represent the gist of an image-caption pair. Finally, the potential gist concepts are ranked.

Consequently, before creating and interpreting any kind of association a precise output of state-of-the-art detectors are a preliminary. Before the recent advances in image under-

standing based on deep learning, increasing the precision in established vision understanding challenges, such as object detection, was one of the main objectives, thus, moving non-literal image meanings beyond the scope of researchers.

Beside this fact, only a few researchers mentioned the existence of non-literal pairs and gave reasons for their decision to focus on literal meanings [Hodosh et al., 2013]: non-literal meanings with their tendency to trigger associations, require knowledge that is often not encoded in the image-caption pair itself, which makes approaches dependent on additional data sources, e.g., lexical hierarchies like WordNet. Finally, and most importantly, going untrodden paths requires to create novel datasets and gold standard annotations, which are extensively human-labored tasks, often subject to discussions, optimizations, and changes.

The question we want to answer is: Can we actually teach machines to make sense of hard, complex use cases of image usages in context? Considering a use case such as search, one cannot computationally model the information need of a human query, while ignoring an important communicative means and without fully understanding human communication. Especially, with respect to the rapidly growing amount of multimodal data, which needs to be supplemented, complemented, indexed etc., the full range of type of intentions needs to be modeled to guarantee utilization of the data (and not only for the sake of collecting), e.g., in that a retrieval of abstract messages becomes possible. Furthermore, an improvement of tasks like caption generation, indexing, or image retrieval is expected, motivated by the fact that an image can be affiliated with a caption complementing the image-caption to a literal or even a non-literal message (cf. Figure 1.1).

This research was originally inspired by the question of understanding *media-iconic images* from the domain of global warming. Media-iconic images are a subclass of non-literal image caption pairs and per definition known so well by the people, because these have been published and used so widely that a textual context does not necessarily have to be given. Examples of media-iconic images are the tank man [1], the fall of the Berlin Wall (cf. Figure 1.2a), death of Carlo Giuliani [2], a polar bear on a melting ice-floe [Perlmutter, 1997], or the Fukushima Daiichi nuclear disaster (cf. Figure 1.2b) etc.

Even though we investigate the step which is before an *established* media-iconic image, thus, the creation, development, and computationally understanding of non-literal pairs, we put special focus to media-iconic pairs. These can be decoded with common knowledge and they are publicly accessible. Furthermore, the theme coverage of our data in this research rather focuses on the domain of global warming. It is a very diverse, controversial, and often differently conveyed theme, so that it is rich in media-iconic image-caption pairs.

[1] https://en.wikipedia.org/wiki/Tank_Man, due to copyright an image cannot be shown here

[2] https://en.wikipedia.org/wiki/Death_of_Carlo_Giuliani, due to copyright an image cannot be shown here

(a) People atop the Berlin Wall near the Brandenburg Gate on 9 November 1989.

(b) The Fukushima I Nuclear Power Plant after the 2011 Tōhoku earthquake and tsunami.

Figure 1.2 Media-iconic image examples (a: https://en.wikipedia.org/wiki/File:Thefalloftheberlinwall1989.JPG, Lear21 at English Wikipedia, b: https://en.wikipedia.org/wiki/File:Fukushima_I_by_Digital_Globe.jpg, Digital Globe, both: CC BY-SA 3.0, last accessed: 10/20/2017).

Finally, one must note that the distinction between literal and non-literal is a rather gradually than a binary classification task. Even though we are talking as if there exists a binary choice, we are completely aware that sometimes just one element of an image or a caption changes a pair from one type to the other. Especially non-literal pairs contain literal elements or encode several aspects in one, e.g., cf. Figure 1.2a, "People atop the Berlin Wall" is literal and non-literal at the same time: it is literal because it describes what can be seen on the image, however, knowing that standing on top of the Berlin Wall was life threatening, before the fall of the Berlin Wall, makes it non-literal. This gradient-like transition from one type to another and the ambiguity in parts of the pairs are one of the reasons why we are following a ranking and not a classification approach.

1.2 Contribution

The core contributions of this thesis build upon initial exploratory studies that focused on the viability of creating datasets of non-literal image usages with minimal supervision [Weiland et al., 2014], and a mixed-method analysis of the problem combining computational and qualitative methods [Ponzetto et al., 2015]. The result of these researches have shown that media-iconic images are of very diverse nature, consequently, hard to detect and to understand by purely vision-based approaches, without providing additional context [Ponzetto et al., 2015; Weiland et al., 2014]. The context frames an image in such a way that it becomes clear(er), which kind of knowledge one has to apply to decode the respective message of an (iconic) image. If no further context is given (cf. Fig 1.1, our introductory example of a paper cup), the image can have multiple meanings, e.g., coffee cup sleeve and disposable food

packaging, where the intended meaning cannot be finally pinpointed. Yet, the results provided a foundation to the definition of non-literal images, thus, insights for the computational modeling of non-literal and literal images.

The work is summarized in:

- Weiland et al. [2014]: Weiland, L., Effelsberg, W., and Ponzetto, S. P. (2014). Weakly supervised construction of a repository of iconic images. In *Proceedings of the Workshop on Vision and Language 2014 (VL '14) at the 25th International Conference on Computational Linguistics (COLING '14)*.

- Ponzetto et al. [2015]: Ponzetto, S. P., Wessler, H., Weiland, L., Kopf, S., Effelsberg, W., and Stuckenschmidt, H. (2015). Automatic classification of iconic images based on a multimodal model : an interdisciplinary project. In Wildfeuer, J., editor, *Sprache - Medien - Innovationen- Building bridges for multimodal research : international perspectives on theories and practices of multimodal analysis*, 7, pages 193–210, Frankfurt am Main ; Bern; Wien. Peter Lang Edition.

Taking advantage of our preliminary research, the contribution to the task of understanding the gist of images in context of short texts is three-fold:

First, a novel dataset of image-caption pairs - containing literal and non-literal examples - is presented. Furthermore, extensive human-labeled annotations are provided, e.g., concept ranking to represent the message of a pair, entity linking to represent the initial pairs as entities, and labeled bounding box image annotations. Second, an approach to understand the messages, with definitions that are necessary for the research, thus, the computational model, and comprehensive evaluation of the approach, is provided. Third, the usefulness of understanding the gist of image-caption pairs is shown in an established use case, i.e., multimedia indexing.

1.2.1 Multimodal dataset of literal and non-literal image-caption pairs

In Chapter 3 we give a comprehensive overview of image-only and multimodal datasets. As there is no dataset satisfying all requirements to allow for a study of non-literal pairs and the understanding of the message of images, we provide a novel dataset. We describe the gist dataset and all required human labeled and evaluated gold standard annotations (also in Chapter 3).

- **Novel dataset:** 328 non-literal and literal image-caption pairs in one dataset. This feature allows the first time for a direct comparison of an approach for both types of communicative means.

- **Human-labeled annotations:** The dataset provides different annotations, which are assessed or created by human annotators.
  - **Ranked gist annotations:** The gist of an image-caption pair is a ranked list of concepts from a knowledge base. Human annotators have assigned concepts from the knowledge base to a pair and for each of the concepts assigned a ranking according to how well the concepts represent the gist.
  - **Entity linking annotations:** Information retrieval has shown the benefits of entity linking. The image-caption pairs are transferred to a list of entities best representing the pair. The gold standard annotation allows for a comparison of different entity linking methods.
  - **Image annotations:** Each of the images is provided with textual image object annotations from a pre-defined list of concepts. Additionally, the position of an object in an image is marked with a bounding box.
  - **Visual linking annotations:** Correspondences between image objects and nouns or noun phrases in the text are set manually. This annotation helps to investigate which parts of the two modalities are descriptive, literal, and in line with each other - contrasting to the complementary parts of a pair.
- **State-of-the-art system annotations:** The amount of multimedia data is rapidly growing, thus, the required human input for initial image or pair annotation also grows. Finally, text generation is a barrier as it needs to be conducted by human annotators. Deep and neural network technologies are promising directions towards overcoming or at least lowering this barrier. To allow for an evaluation of the performance of state-of-the-art systems the dataset provides also the results of a deep learning approach. These results are neither created with additional human labor, nor checked for validity.
  - **Image annotations:** Each image is associated by the system with tags from a vocabulary of around 2,000 recognizable objects. A recognized object is accompanied by a confidence value.
  - **Caption generations:** The tags of recognized objects in an image are used as input for a caption generation approach. As a result one image is described by one caption accompanied with an overall confidence value, too.

The work is summarized in:

- Weiland et al. [2016]: Weiland, L., Hulpuş, I., Ponzetto, S. P., and Dietz, L. (2016). Understanding the message of images with knowledge base traversals. In *Proceedings*

of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, September 12- 6, 2016, pages 199–208.

- Weiland et al. [2017]: Weiland, L., Hulpuş, I., Ponzetto, S. P., and Dietz, L. (2017). Using object detection, nlp, and knowledge base to understand the message of images. In *Lecture notes in computer science MultiMedia Modeling : 23rd International Conference, MMM 2017*, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part II; 405-418, Springer International Publishing, Cham, 2017.

- Weiland et al. [2018a]: Weiland, L., Hulpuş, I., Ponzetto, S. P., Effelsberg, W., and Dietz, L. (2018). Knowledge-rich Image Gist Understanding Beyond Literal Meaning. In *Journal on Data & Knowledge Engineering, 2018*, Elsevier, 2018.

1.2.2 Understanding the Message of Images

Chapter 2 starts with the definition of non-literal image-caption pairs, media-iconic images, and their differentiation to literal pairs. Paying special attention to these definitions and the important differences between non-literal to literal pairs, our dataset with gold standard is created. Both, data and gold standard, are used within all contributions of the thesis. Furthermore, a pipeline combining entity linking, leveraging from external knowledge, and using a learning to rank approach to detect the message of images is created and evaluated (cf. Chapter 4 and Chapter 5, respectively). The contributions are:

- **Novel task definition:** We formulate the novel task of detecting the gist of image-caption pairs using the vocabulary and topics provided by a reference external resource, i.e., a knowledge base.

  - **Definition of Gist:** the preliminary to computationally approaching the understanding of the gist of images is a definition about what is a message of an image in the context of a short text. Gist can be expressed by concepts in a knowledge base.

  - **Depictable vs. non-depictable:** Differentiation of depictable and non-depictable objects. As a preliminary to the definition of literal and non-literal one needs to distinguish concepts that can in principle be depicted, e.g., visible objects, recognizable by humans and object detectors, from those that are non-depictable concepts, i.e., in that they are abstract, such as global warming, thus, not recognizable by object detectors.

  - **Literal vs. non-literal image caption pairs:** Differentiation of literal and non-literal image usages. Literal image caption pairs are self-contained, often descriptive. Non-literal pairs are often complementary, conveying an abstract message.

- **Approach:** A pipeline to detect and rank potential messages for an input pair as query. This pipeline addresses several characteristics of using images in the context of short text as a query, such as enhancing the query with external knowledge.

  - **Benefit from external knowledge:** The usage of external knowledge, i.e., the knowledge base Wikipedia, addresses the characteristic of image-caption pairs that are not self-contained but pointing towards common knowledge and connections between facts. Entries in the knowledge base are referred to as concepts.
  - **Benefit from the knowledge graph structure:** Considering that facts and knowledge is connected with each other, the knowledge is represented by a graph. Leveraging from the structure, thus graph connectivity measures, supports to derive these connections between query and gist concepts.
  - **Benefit from textual content measures:** Knowledge and especially more complex matters are dependent of textual content, which explains the details and gives explanations especially to complex themes. The proposed approach considers also the textual content of concepts, in that the article texts from the knowledge base corresponding to a concept are used to generate content-based measures.

- **Evaluation:** An extensive analysis of our approach is conducted, considering different types of features and features collected at different stages of the proposed methodology.

  - **Benefit of different components of our approach:** Analysis of what strategy and what features reveal the best gist concepts. This evaluation provides detailed insights into the proposed approach and each of its stages.
  - **Benefit from combining the modalities (multimodal approach):** Analysis whether using the combination of signals from both modalities (image and text) performs better than using one of the modalities (image or text) as query.
  - **End-to-end gist detection:** (assessing the quality of Deep Learning in context of image message understanding). A more realistic scenario is investigated by the use of a Deep Learning API, which is capable of detecting objects in an image and generating a caption for a given image.

  - **Benefit of visual linking:** Addressing the descriptive nature of literal pairs and the complementary nature of non-literal pairs, the visual linking provides useful information about the initial query pair.

The work is summarized in:

- Weiland et al. [2015]: Weiland, L., Dietz, L., and Ponzetto, S. P. (2015). Image with a message: Towards detecting non-literal image usages by visual linking. In *Proceedings of the 2015 EMNLP Workshop on Vision and Language (VL'15)*, pages 40–47.

- Weiland et al. [2016]: Weiland, L., Hulpuş, I., Ponzetto, S. P., and Dietz, L. (2016). Understanding the message of images with knowledge base traversals. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016*, Newark, DE, USA, September 12- 6, 2016, pages 199–208.

- Weiland et al. [2017]: Weiland, L., Hulpuş, I., Ponzetto, S. P., and Dietz, L. (2017). Using object detection, nlp, and knowledge base to understand the message of images. In *Lecture notes in computer science MultiMedia Modeling : 23rd International Conference, MMM 2017*, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part II; 405-418, Springer International Publishing, Cham, 2017.

- Weiland et al. [2018a]: Weiland, L., Hulpuş, I., Ponzetto, S. P., Effelsberg, W., and Dietz, L. (2018). Knowledge-rich Image Gist Understanding Beyond Literal Meaning. In *Journal on Data & Knowledge Engineering, 2018*, Elsevier, 2018.

1.2.3 Using Gist Detection for Multimedia Indexing

In Chapter 6 we modify the presented pipeline for understanding the gist, to demonstrate its benefit when applied to an established research problem, such as multimedia indexing.

- **Use Case:** We apply the approach of understanding the gist to an established research problem. On a benchmarking dataset with 25,000 instances it can be shown that also established research tasks, such as multimedia indexing benefits from gist detection in an end-to-end approach.

- **Evaluation:** We again conduct an extensive analysis of our approach using the benchmarking dataset, where we analyze the impact of each pipeline component and the proposed modifications. Furthermore, we study the benefits of the different gist candidates collected from the two modalities and the different graph expansion steps.

The work is summarized in:

- Weiland et al. [2018b]: Weiland, L., Ponzetto, S. P., Effelsberg, W., and Dietz, L. (2018). Understanding the Gist of Images - Ranking of Concepts for Multimedia Indexing. In *arXiv preprint*, arXiv:1809.08593, 2018.

1.3 Outline

In Chapter 2 we review the related work from the perspective of communication scientists and linguists as a fundamental to build a computational model. Furthermore, the perspective of the computer scientists shows the novelty and difficulty of our work. In Chapter 3 datasets from the image processing and multimodal domain are revised. Owed to the fact that there is neither a dataset covering both types of pairs (at least classified as such), nor gist concept annotations, a dataset with detailed gold standard annotations satisfying this need is created and described. The main contributions of this thesis can be found in Chapter 4 and Chapter 5, where the approach is explained in detail and where the extensive evaluation of the proposed approach can be found, respectively. Chapter 6 is about the use case: within a comparative study we show the usefulness of understanding the gist to standard questions from the multimodal domain, e.g., multimedia indexing. Finally, in Chapter 7 the thesis is concluded with a discussion of the objectives and the outlook.

Chapter 2

Related Work

The understanding of images is a complex and faceted field of research. To allow for a contextualization and classification of the research, the discussion of the related work is grouped according to the three different levels representing the development of the proposed approach. First, the definition for image understanding with the focus on media-iconic images and the delimitation to other multimodal research areas need to be set. Second, an overview of approaches from computer science on how to handle and benefit from multimodal data, which mainly focuses on the combination of image and text, is given. Third, as the proposed approach facilitates methodologies from information retrieval and entity linking, these domains are further discussed.

The goal of this chapter is to study the different definitions of iconic images and to merge those to one definition, which is used within the proposed approach. From the computer science perspective it will be discussed, how to benefit from existing multimodal approaches, however, it will be shown that our research is more difficult due to the lack of direct related work. Furthermore, the importance and novelty of this research domain will be demonstrated.

2.1 Iconic Images and Semiotic

This section starts with a perspective of the communication scientists and linguists to the understanding of images in context of texts, a definition of news iconic images, and their approaches to dissect and analyze images. The goal of the subsection is to leverage one definition of news icons that the research works with, furthermore, to delimit from image interpretation aspects that are currently not considered. Additionally, within a perspective of computer science about iconic images we show the novelty of the research branch.

2.1.1 A Perspective from Communication Science and Linguistics

From the perspective of communication scientists and linguistics, there are a lot of related works. These works are not free from complexity or controversy. Furthermore, these are influenced different schools, so that in the following we try to give an overview about related works, which help to understand iconic images. The goal is to end up with a definition of iconic images, that can be computationally modeled.

When talking about the understanding of iconic images, one stumbles the sooner than later, over the semiotic theory written by Charles Sanders Peirce in the 1860s. Semiotic, defined as the theory and study of signs, is not invented by Peirce, as signs are nearly as old as human mankind, however, for his thorough works on semiotic he is considered to be one of the most famous representatives. Due to this detailed nature, we do not discuss Peirce's theories and definitions, but instead review works that are closer to our purpose and domain of media-iconic images.

Tight coupling of images and texts. Within this research we focus on images in the context of short texts, i.e., captions. This focus is motivated by the fact that the combination is well studied in linguistic and communication science. Furthermore, creating meaning by the use of the combination of text and images, is not an invention of modern times, if one thinks about visually stunning illustrations of Biblical texts [Harrison, 2003], but it is undergoing a revival. One of the most relevant aspects is given by Horn, who calls this multi-modal mix visual language [Harrison, 2003]:

> ".. the tight coupling of words, images, and shapes into a unified communication unit. "Tight coupling" means that you cannot remove the words or the images or the shapes from a piece of visual language without destroying or radically diminishing the meaning a reader can obtain from it [Horn, 1999, p. 27].

Social semiotics. From social semiotics three important principles can be extracted (names of principles are adapted from [Harrison, 2003, p. 48]), which should be considered when studying visual social semiotics, thus, iconic images. Additionally, these principles motivate our choice to conduct the research on one language (English) with data from one semiotic system (USA) which is close to our European semiotic system.

1. Social conventions.

   > Although things may exist independently of signs we know them only through the mediation of signs. We see only what our sign systems allow us to see. [...] Semioticians argue that signs are related to the signified by social conventions which we learn. We become so used to such conventions

> in our use of various media that they seem "natural," and it can be difficult for us to realize the conventional nature of such relationships [Chandler, 1994].

2. Social and cultural bias.

> Meaning of signs is created by people and does not exist separately from them and the life of their social/cultural community. Therefore, signs have different meanings in different social and cultural contexts - meanings can range from very different [...] to subtle and nuanced [...]. [...] The growing number of books and articles on this subject attests to the difficulties writers face when trying to create messages for people whose semiotic systems are different from theirs. [Harrison, 2003, p. 48]

3. Affect and alter meanings.

> Semiotic systems provide people with a variety of resources for making meaning. Therefore, when they make a choice to use one sign, they are not using another. [...] The ability to choose gives communicators a certain amount of power to use signs in unconventional ways and, therefore, affect and even alter meanings [Lemke, 1990].

Visual social semiotics. Semiotics is generally described as the "study of signs" [Harrison, 2003, p. 47]. A sign represents or conveys a meaning, which is named the signified. A derivative of semiotics is the visual social semiotics. Jewitt and Oyama defines the visual social semiotics as

> the description of semiotic resources, what can be said and done with images (and other visual means of communication) and how things people say and do with images can be interpreted. [Jewitt and Oyama, 2001, p. 136]

Categories of images. According to [Irene Hammerich, 2002, pp.140-142] there are three categories of images, which are necessary to be aware of to understand that images are connected to a meaning in different ways and as such need to be approximated differently. **Icons** are used to refer to something which is similar or a resemblance of what is depicted in the image. An **index** is given when a reader can obtain the meaning of the index due to its relationship to its meaning. An index does not have any similarity or resemblance. **Symbols** do neither have a similarity nor have a direct relationship to its meaning, it is something established, something we've learned.

The goal of our research is to understand the meaning of iconic images, especially those presented in news and media (news icons according to Perlmutter and Wagner [2004]).

Definition of news icon. In general Perlmutter defines a news 'icon' (interchangeably used with photo-journalistic icon) as a celebrated product of photojournalism [Perlmutter and Wagner, 2004, p. 91]. More strict than the definition of an image being iconic, which is given "[...] if it bears a similarity or resemblance to what we already know or conceive about an object or person" [Harrison, 2003, p. 50] (adapted definition from [Irene Hammerich, 2002, pp. 140-142]), an image of a news icon functions metonymically [Perlmutter and Wagner, 2004, p. 100]. Special attention needs to be payed to the fact that images can have a metonymical or metaphoric function - two language constructs that should not be interchanged. Recall, a metonym is something that is used to stand for something else, e.g., the famous image of 'tank man', has become a news icon for the student demonstrations in Tiananmen in 1989. Different from that a metaphor is used to explain something having the same functionality, e.g., depicting burnout, which can be shortly explained by an unbalanced inner life, as an unbalanced scale. Perlmutter defined a typology, about the making of an icon [Perlmutter, 1998; Perlmutter and Wagner, 2004]. Therefore, he has defined the standard elements of a photo-journalistic icon. These include: (1) importance of the event depicted, (2) metonymy, (3) celebrity, (4) prominence of display, (5) frequency of use, and (6) primordality.

Each of these typologies need to be examined for an image to allow it to be used as an encapsulation and exemplification of complex issues.

Furthermore, from the theory on the analysis of iconic images in general, we learn that according to Kress and van Leuven [Gunter R. Kress, 1996] an image performs three kinds of meta-semiotic tasks to create meaning. These tasks are called the representational metafunction ("what is the picture about?"), interpersonal metafunction ("How does the picture engage the viewer?"), and compositional metafunction ("How do the representational and interpersonal metafunctions relate to each other and integrate into a meaningful whole?") [Harrison, 2003]. We are mostly interested in the representational and compositional metafunction, and less in the interpersonal perspective.

Nevertheless, there are also discrepancies in the definition of news icons. Thus, according to Perlmutter [Perlmutter and Wagner, 2004, p.95] news professionals use as the most standard analogue the description of natural 'windows' onto the world to define news icons. However, he disagrees and claims that these images (photojournalism's output) are manufactured and framed for consumption as any other news product.

Ambiguity and bias of news icon. Ambiguity is resolved for the viewer when the respective mainstream media sources frame the image for reader consumption **through captions** [Perlmutter and Wagner, 2004, p.102].

The linguistic and communication science perspective on semiotics and the definition of non-literal and iconic images is very faceted and complex. With respect to the computational modeling of non-literal and iconic images, the complexity of these definitions need to be reduced in that some restrictions and delimitations need to be set. Thus, when investigating iconic images, we use images with their affiliated captions as input. This is in line with the observation that image and text build a communicative unit, which in turn can contain complementary or similar information. Furthermore, as the channels and ways of transmitting news and how news reach the people have changed a lot, we refer to *media-iconic images* [Drechsel, 2010], instead of the term news icons. However, the characteristic of an image being rather a metonymy than a metaphor still applies. Additionally, to build a counterpart to the descriptive nature of literal image-caption pairs, we use the terminology of non-literal images as the parent-class (hypernym) of media-iconic images.

2.1.2 A Perspective from Computer Science

Just recently computer scientists from the domain of computer vision moved away from only considering things that can be seen in an image. Jiang et al. [2017] allow for a learning and retrieval of abstract concepts that are inevitably connected to concepts that are in principle depictable and in their training data represented as visible objects in the images. Examples of depictable concepts are, cake, presents, and kids. These visually recognizable concepts are connected to the more abstract concept of birthday (example from [Jiang et al., 2017]). The translation from concrete to abstract concepts and their retrieval is a step towards the understanding of images, but it does not tackle the even harder problem of understanding non-literal images. Most related to the understanding of non-literal images and their messages is the research area of understanding metaphors in texts. This connection is given by the fact that non-literal images can be a metonymy or a metaphor. Similar to the semiotics, the multimedia view on these kinds of focus areas has only recently become attractive. Shutova et al. [2016] benefit from a fusion of text and visual features to identify textual metaphors. They compare different stages of fusion (middle and late fusion) for the learned textual and visual embeddings and evaluate the 'metaphoricity' of an input text based on several cosine measures, where the thresholds of a text being a metaphor are learned from two annotated corpora, containing metaphors. Additionally, their approach relies on external resources, whereas textual embeddings are trained with a Wikipedia corpus and the visual embeddings are trained on images retrieved from Google image search with keyword-based queries.

However, to the best of our knowledge there is no research about tackling the problem of automatically understanding non-literal image usages conveying abstract topics as meanings.

2.2 Multimodal Modeling

Recent years have seen a growing interest for interdisciplinary work which aims at bringing together processing of visual data such as video and images with NLP and text mining techniques. This is no surprise, since text and vision are expected to provide complementary sources of information, and their combination is expected to produce better, grounded models of natural language meaning [Bruni et al., 2012], as well as enabling high-performing end-user applications [Aletras and Stevenson, 2012].

However, while most of the research efforts so far concentrated on the problem of image-to-text and video-to-text generation – namely, the automatic generation of natural language descriptions of images [Feng and Lapata, 2010a; Gupta et al., 2012; Kulkarni et al., 2011; Yang et al., 2011], and videos [Barbu et al., 2012; Das et al., 2013b; Krishnamoorthy et al., 2013] – few researchers focused on the complementary, yet more challenging, task of associating images or videos to arbitrary texts – [Feng and Lapata, 2010b] and [Das et al., 2013a] being notable exceptions. However, even these latter contributions address the easier task of generating literal descriptions of depictable objects found within standard, news text, thus disregarding other commonly used, yet extremely challenging, dimensions of image usage such as media icons [Drechsel, 2010; Perlmutter and Wagner, 2004]. The ubiquity of non-literal usages has not received much attention yet in the fields of automatic language and vision processing: researchers in Natural Language Processing, in fact, only recently started to look at the problem of automatically detecting metaphors [Shutova et al., 2013] (cf. Subsection 2.1.2), whereas research in computer vision and multimedia processing did not tackle, to the best of our knowledge, the problem of (media) iconic images at all. However, there are a lot of related works in broader related domains. These related works are reviewed, as they might be able to solve the challenge of understanding the message of images or at least parts of the problem. Since non-literal image-caption pairs (in the following also referred to as **non-literal images** or **non-literal pairs**), at least according to the definition and data we use in our research, are multimodal due to their tight coupling between caption and image, we mainly review approaches and insights from the domain of multimedia modeling and multimedia analysis in the following. Surveys, which give broad while detailed overviews about different approaches to a specific domain or field of problem, typically group approaches according to the existence or absence of labeled data, thus, group according to supervised, semi-supervised (a mix between labeled and unlabeled data), and

unsupervised approaches. We review all three types of data, however, in Table 2.1, which is an extended version of the table provided by Bernardi et al. [2016], we do not group according to the presence or absence of labeled data during training, but according to the type of approach. These approaches are very diverse and often inspired by works from single-modality relatives. We review in Table 2.1 the following groups of approaches: statistical and probabilistic, (shallow) machine learning, neural network, deep learning, common, latent, and hamming space, tree and (scene) graph, knowledge base, and mathematical optimization. The latter approaches, such as knowledge base approaches, are less popular.

Furthermore, in Table 2.1, we indicate (marked with x), which tasks are conducted. We review the tasks of annotation, indexing, retrieval, and generation. We review approaches working with the modalities text, image, video, and their combinations. The column headers I, V, and T are abbreviations for the respective modalities, e.g., [Jeon et al., 2003] proposes a probabilistic model to conduct the tasks of image annotation and image retrieval (cf. Table 2.1, ID 1).

In the following we distinguish between multimedia indexing and classification and cross-media retrieval and generation, where the objectives are to represent and to retrieve or to generate the data, respectively. The approaches follow in principle two different ideas of representing the training data, either learning one space for each modality type (single space), e.g., an image and a text space, or jointly learning a common space, representing both modalities in one space. The results have shown that a common space outperforms single space approaches. In turn, multimedia modeling usually refers to the combination of modalities as fusion and the approaches are distinguished by the level of fusion. The levels are distinguished according to an early, a late, and a hybrid fusion strategy [Atrey et al., 2010]. An early fusion combines the features from the different modalities, a late fusion combines the decision of a decision unit for each modality separately (e.g., a classifier assigns a feature vector to a class), and a hybrid approach uses a mixture between these two. The retrieval and generation tasks can further be grouped according to the modality that is retrieved or generated, thus, whether a text or an image is retrieved or generated. Our focus is put on cross-media approaches, which uses one modality as a query, e.g., an image, to retrieve the other modality, e.g., a text.

If the approach uses single spaces for the representation of the training data, the retrieval or generation is conducted via a detour: first, similarities between the query instance and the training instances are exploited. Second, the affiliated instances of the other modality of the most similar training instance(s) are used or concatenated as result, e.g., for a query image the most similar images in the image space are collected, the captions of those images are presented as the result, ranked by the similarity score between the query and training image.

2.2.1 Multimedia Indexing and Classification

The task of multimedia indexing is mainly about representing the respective multimedia data. Representing multimedia data is approached as a classification task, where the goal is to assign the multimedia instances to one or several classes best representing the content of the instance. In the context of multimedia data, which consists of image and text, i.e., the best representing class(es) is given as the class that best describes the most salient object(s) in the image [Huiskes and Lew, 2008]. The originators of the MIR Flickr benchmarking dataset presented two different approaches for classification of the data. A Support Vector Machine (SVM) and a Linear Discriminant Analysis (LDA) based classification. Both approaches are trained with different feature sets: one using uni-modal features (low-level features from the vision domain) and one combining the low level visual features with the textual tags as features [Mark J. Huiskes and Lew, 2010]. There are several successors using various types of approaches. These are shallow, such as the mentioned SVM or LDA [Mark J. Huiskes and Lew, 2010], or vector space representations, e.g., high-dimensional vector space representations based on a cross-lingual latent semantic indexing [Hare and Lewis, 2010]. Furthermore, because of the recent advances in deep learning, there are several deep approaches. [Srivastava and Salakhutdinov, 2012] present a multimodal deep Boltzman machine (DBM) for representing multimodal data and for generating for one given modality the other missing modality (e.g., caption generation for a query image). The approach benefits from the simpler Restricted Boltzman Machine (RBM) that represents the texts as binary vectors and from a Gaussian-Bernoulli RBM that represents the image with real-valued vectors. The DBM is compared to a multimodal version of a Deep Belief Network (DBN), which is a directed network and where the multimodal modeling takes place in the joint layer (which is not the case for DBMs, here this responsibility is spread across the entire network) [Srivastava and Salakhutdinov, 2012]. In [Wang et al., 2016] the problem is approached with a regularized deep neural network (RE-DNN) making use of the same feature set as in [Srivastava and Salakhutdinov, 2012] (PHOW, GIST, MPEG-7) compared with CNN features (features extracted from the last layer before the classification layer of a Convolutional Neural Network). Finally, in [Chen et al., 2016] the multimodal representation and image retrieval is addressed with a multi-label hashing approach.

2.2.2 Cross-Media Retrieval and Generation

Similar to multimedia indexing, the cross-media retrieval and generation approaches represent the multimodal data. As opposed to indexing, the cross-media tasks are evaluated with respect to the quality of the retrieval and generation tasks, often human annotators are consulted to

assess the result quality or to build a gold standard, e.g., [Elliott and de Vries, 2015]. In the following we review related works according to the type of modality that is used for querying. We separately review both types of modality as the approaches often do not present a bi-directional approach or do only claim, but not evaluate the bi-directionality; [Chen and Zitnick, 2015; Verma and Jawahar, 2014] being notable exceptions (cf. Table 2.1, IDs 14 and 19, respectively). Furthermore, we want to get a detailed notion about the ability of each approach to address literal and non-literal aspects for each modality.

Retrieval or generation of a text for a query image. If approaches make use of single modality spaces, the image is taken as the query. The method consists of learning the visual representations and the retrieval of a ranked list of similar images based on detected objects, detected interactions between objects, detected attributes, and the detected scene. Then often the ranked list is re-ranked according to visual or textual information [Bernardi et al., 2016].

To learn the visual representations low-level features (e.g., color, texture, and contrast), higher level features, or descriptors (e.g., SIFT, HOG, Gabor, Haar) are extracted from the training images. A ranked result list is either retrieved by the similarity to the query image [Gupta et al., 2012; Ordonez et al., 2011; Patterson et al., 2014] or by a combination of separate rankings retrieved according to the similarity between image regions of the query image and training images [Kuznetsova et al., 2012]. The re-ranking is then conducted by the use of additional visual information [Gupta et al., 2012; Ordonez et al., 2011; Patterson et al., 2014] or textual information [Gupta et al., 2012; Mason and Charniak, 2014]. The final caption(s) for the query image are either also presented in a ranking [Hodosh et al., 2013] or combined to one caption [Kuznetsova et al., 2012].

To summarize the reviewed related works, the focus is clearly on the retrieval and generation of descriptions or (literal) captions. The objective of these works is to understand the interconnection of images and texts when both form a communicative unit to convey a meaning. Consequently, these methodologies are evaluated for their performance to generate or retrieve literal captions or to understand the pure visually recognizable content of an image, by making use of the output of object detectors and visual features. This focus is underlined by the mostly literal datasets that are used in the related works to train and test the methodologies. Their gold standard annotations which are consulted to evaluate the respective approaches also address the literal focus (cf. Chapter 3 for further details). In principle, the presented approaches would be able to retrieve or generate non-literal captions to a given query image. However, as the data lacks the annotation of non-literal and literal pairs, it is open for speculation how the models perform if both types of texts are represented in such a model. Compared to our novel task of gist detection, the objective of generating or retrieving texts takes place before our actual gist detection, thus, it can serve as input for

our approach, but it cannot solve the task of understanding the gist of non-literal images or image-caption pairs (cf. Chapter 5, RQ8 and 9).

Retrieval of an image for a query text. Similar to the research area described in the previous subsection, there are different tasks that can be solved using a textual query within a multimodal training data approach, e.g., image annotation, image retrieval, or image generation. Also similar to the previously described research area there are different ways to handle the multimodal data. The most relevant difference is the change between query and target spaces, if the data is represented in single modality spaces. The query is conducted in the textual space, whereas the target is selected from the vision space. Nevertheless, we want to review also this domain, as it closes the circle of the question on how image and text interact with each other.

Some of the works used for text retrieval and generation have also shown to perform well when the type of modality for query and result are exchanged. Bidirectional mapping of image and text are often represented in deep learning models [Karpathy et al., 2014] and compositional semantics, using e.g., KCCA [Hodosh et al., 2013; Socher and Fei-Fei, 2010], however these approaches annotate the images as a whole, without considering detailed expressiveness when the relations of objects are changed. Approaches representing images and texts in scene-graphs [Johnson et al., 2015; Schuster et al., 2015] or so-called visual dependency representations (VDR) [Elliott et al., 2014] allow for such an exchange as they better represent the details of the image-text interplay, e.g., image regions do have a tight coupling to textual labels, attributes, etc.

Similar to the visual query approaches, the textual query approaches of cross-modal research focus on literal image-text pairs - this focus is again due to the nature of the datasets. Also similar to the previous conclusion, it remains open for speculation how the models perform when trained on non-literal or both types of pairs. Besides the fact that non-literal pairs are out of scope of the focus areas, the question whether the literal meaning is different from the task that has been approached and solved so far remains unanswered. However, understanding the meaning of literal image-text pairs is not yet finally solved and thus remains an active research area. Finally, - and even more important - one can conclude that understanding the meaning of non-literal image-text pairs is a novel task.

Table 2.1 Overview of the different multimodal approaches and their tasks (extended version of Bernardi et al. [2016]). Ordered by the type of approach to represent data.

ID	Name	Task												Methodology
		(I)mage, (V)ideo, (T)ext												
		Annotation			Index			Retrieval			Generation			
		I	V	T	I	V	T	I	V	T	I	V	T	
		Statistical/Probabilistic Approaches												
1	Jeon et al. [2003]	x						x						Probabilistic Model
2	Feng and Lapata [2008]							x						Joint Probabilistic Model, Latent Variable
3	Farhadi et al. [2010]									x				Probabilistic Model: Markov Random Fields (MRF)
4	Mark J. Huiskes and Lew [2010]				x		x							LDA
5	Kulkarni et al. [2011]												x	Probabilistic Model: Conditional Random Fields (CRF)
6	Yang et al. [2011]												x	Probabilistic Model: Hidden Markov Model (HMM)
7	Srivastava and Salakhutdinov [2012]				x		x							Deep Boltzman Machine (DBM)
8	Gupta et al. [2012]									x				Probabilistic Model
9	Hodosh et al. [2013]									x				Kernel Canonical Correlation Analysis
10	Mason and Charniak [2014]									x				Probabilistic Model

Table 2.1 Overview of the different multimodal approaches and their tasks (extended version of Bernardi et al. [2016]). Ordered by the type of approach to represent data.

ID	Name	Task												Methodology
		(I)mage, (V)ideo, (T)ext												
		Annotation			Index			Retrieval			Generation			
		I	V	T	I	V	T	I	V	T	I	V	T	
11	Yatskar et al. [2014]												x	Probabilistic Model
12	Johnson et al. [2015]							x						CRF and scene graphs
13	Schuster et al. [2015]							x						CRF and scene graphs (Johnson et al. [2015])
14	Verma and Jawahar [2014]							x		x				LDA and CCA
	Deep Learning/Neural Network Approaches													
15	Karpathy et al. [2014]							x		x				Common Embedding Space (Image Object-Sentence Tree Embeddings)
16	Socher et al. [2014]							x		x				Dependency Tree RNN (DT-RNN)
17	Mao et al. [2015]									x			x	Multimodal-RNN (m-RNN)
18	Vinyals et al. [2015]									x			x	Neural Image Caption (CNN and RNN)
19	Xu et al. [2015]									x			x	Neural Image Caption and Attention Model (Stochastic and Deterministic)
20	Chen and Zitnick [2015]							x		x			x	RNN with latent variables
21	Donahue et al. [2017]									x			x	Long-term Recurrent Convolutional Networks (LRCNs)
22	Devlin et al. [2015]									x				Maximum Entropy (ME) and RNN

Table 2.1 Overview of the different multimodal approaches and their tasks (extended version of Bernardi et al. [2016]). Ordered by the type of approach to represent data.

ID	Name	Task												Methodology
		(I)mage, (V)ideo, (T)ext												
		Annotation			Index			Retrieval			Generation			
		I	V	T	I	V	T	I	V	T	I	V	T	
23	Fang et al. [2015]												x	ME, CNN, and Multiple Instance Learning (MIL)
24	Jia et al. [2015]									x			x	Long Short Term Memory (LSTM, extended)
25	Karpathy and Li [2015]									x			x	Multimodal RNN
26	Kiros et al. [2015]									x			x	Joint embedding and LSTM
27	Lebret et al. [2015]									x			x	Common space CNN
28	Yagcioglu et al. [2015]									x				FC-7 and Distributional Semantics
	Common Embedding/Latent/Hamming Space Approaches													
29	Gong et al. [2014]									x				Common Latent Space (Image-Sentence Embeddings)
30	Ushiku et al. [2015]									x				Common Subspace for Model and Similarity (CoSMoS)
31	Chen et al. [2016]				x		x							Multi-label hashing
	Machine Learning Approaches													
32	Mark J. Huiskes and Lew [2010]				x		x							SVM

Table 2.1 Overview of the different multimodal approaches and their tasks (extended version of Bernardi et al. [2016]). Ordered by the type of approach to represent data.

ID	Name	Task												Methodology
		(I)mage, (V)ideo, (T)ext												
		Annotation			Index			Retrieval			Generation			
		I	V	T	I	V	T	I	V	T	I	V	T	
33	Ordonez et al. [2011]									x				Linear regression, SVM
34	Patterson et al. [2014]									x				Im2Text [Ordonez et al., 2011]
	Knowledge Base Approaches													
35	Altadmri and Ahmed [2009]				x	x		x	x					VisualNet (Knowledge base)
	Tree and Scene Graph Approaches													
36	Mitchell et al. [2012]												x	Syntactic Trees (Description Generation)
37	Elliott and Keller [2013]												x	Visual Dependency Representations (VDR)
38	Elliott et al. [2014]							x						VDR
39	Elliott and de Vries [2015]												x	VDR
40	Lin et al. [2015]												x	Scene/Parse Graphs and Semantic Trees
41	Ortiz et al. [2015]												x	Statistical Machine Translation Model and VDR
	Mathematical Optimization													
42	Kuznetsova et al. [2014]												x	Tree Composition, Integer Linear Programing (Description Generation)

Table 2.1 Overview of the different multimodal approaches and their tasks (extended version of Bernardi et al. [2016]). Ordered by the type of approach to represent data.

ID	Name	Task												Methodology
		(I)mage, (V)ideo, (T)ext												
		Annotation			Index			Retrieval			Generation			
		I	V	T	I	V	T	I	V	T	I	V	T	
43	Kuznetsova et al. [2012]									x				Visual Similarity (Candidate retrieval), Integer Linear Programing (Description Generation)
44	Li et al. [2011]												x	Phrase fusion n-grams

2.3 Gist Pipeline-Specific Preliminaries

To the best of our knowledge there is no explicit related work for the task of gist detection. However, this work touches on different research communities evolving around the fields of object detection from images, entity linking and retrieval. Furthermore, it benefits from graph structure and content of knowledge bases, which we discuss in detail in the following.

Object detection from images. There are a lot of related works about object detection in images as it has been in the scope of researchers for decades. In the following we review a short selection of more recent works. Triggered through benchmark collections for image retrieval [Thomee and Popescu, 2012] and benchmarking tasks [Russakovsky et al., 2015], a large body of works focuses on how to detect objects in images [Everingham et al., 2010; Lin et al., 2014; Ordonez et al., 2011, *inter alia*]. These either train object detectors from images with bounding box annotations, use captions to guide the training, or generate captions for images, based on an unsupervised model from the spatial relationship of such bounding boxes [Elliott and de Vries, 2015].

Since many images are accompanied by captions, approaches have been devised so that the usage of text passages from the captions aid the detection of objects and actions - whereas actions refer to activities that can be detected from still images, such as sitting, riding a bike, walking - depicted in the image. This idea is exploited using supervised ranking [Hodosh et al., 2013], using entity linking and WordNet distances [Weegar et al., 2014], and using deep neural networks [Socher et al., 2014]. One application is image question answering [Ren et al., 2015]. Research to this end has thus far focused on literal image-caption pairs, where the caption enumerates the objects visible in the image. In contrast, the emphasis of this work is on non-literal image-caption pairs with media-iconic messages, which allude to an abstract gist concept that is not directly visible.

Even though datasets such as ImageNet provide over 14 million images, only 8% have bounding boxes, which are crucial for training object detectors. The lack of such training material is the only barrier for application in our domain. For this reason and to facilitate reproducibility of our research, we simulate object detection or rely on an external system such as the Microsoft API. While this work builds on object detection tags, it has been shown that object classes available in ImageNet are insufficient to capture objects found in images on topics of global warming [Weiland et al., 2015].

Knowledge Bases. DBpedia [Lehmann et al., 2015] is a structured knowledge base, which extracts knowledge from Wikipedia. It extracts information from categories, the category hierarchy, or infoboxes. As DBpedia uses a single ontology to represent classes and properties, it can map content from different language versions of Wikipedia. Furthermore, it uses

Semantic Web and Linked Data technologies, to provide the structured information. The ontology is rich in classes and properties (relation types), e.g., currently there are 685 classes and 2,795 properties (as of 2017).

BabelNet [Navigli and Ponzetto, 2012a] is a multilingual semantic lexicon (network), which combines information from several different resources, such as WordNet, Wikipedia, and ImageNet. Similar to DBpedia it provides a SPARQL endpoint and a linked data interface to provide access to its content and structure. BabelNet reports its statistic on an instance level: it contains over 13 mio. Babel synsets (concepts) and over 380 mio. lexico-semantic relations [1].

Similar to DBpedia, Yet Another Great Ontology (YAGO [Suchanek et al., 2007]) is a knowledge base making benefit of information and content from Wikipedia, where the information instances are organized according to an ontological structure. Different to DBpedia, YAGO focuses on, e.g., the spatial and temporal dimension [2].

Entity linking. Detecting entity mentions in text and linking them to nodes in a knowledge base is a task well studied in the TAC KBP venue. Most approaches include two stages. The first stage identifies candidate mentions of entities in the text with a dictionary of names. The second stage disambiguates these candidates using structural features from the knowledge graph, such as entity relatedness measures [Ceccarelli et al., 2013; Hulpuş et al., 2015] and other graph walk features [Talukdar et al., 2008]. A prominent entity linking tool is the TagMe! system [Ferragina and Scaiella, 2010]. A simpler approach, taken by DBpedia spotlight [Mendes et al., 2011], focuses on unambiguous entities and breaks ties by popularity. We evaluate both approaches in Chapter 5.

Entity retrieval. We cast our gist detection task as an entity retrieval and ranking task, with an image-caption pair as the query. As we are using articles and categories as candidates we will refer to concept retrieval instead of entity retrieval in later chapters. Entity retrieval tasks have been studied widely in the IR community in INEX and TREC venues [Balog et al., 2010; Demartini et al., 2009]. The most common approach is to represent entities through textual and structural information in a combination of text-based retrieval models and graph measures [Zhiltsov et al., 2015].

Different definitions of entities have been explored. Recently, the definition of an entity as "anything that has an entry in Wikipedia" has become increasingly popular. Using entities from a knowledge base that are (latently) relevant for a query for ad-hoc document retrieval has lead to performance improvements [Dalton et al., 2014; Raviv et al., 2016].

[1] http://babelnet.org/stats, last accessed 10/26/2017

[2] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/faq/, last accessed 10/26/2017

Moreover, using text together with graphs from article links and category membership for entity ranking has been demonstrated to be effective on free text entity queries such as "ferris and observation wheels" [Demartini et al., 2008]. In contrast to this previous work, our paper focuses on a graph expansion and clustering approach.

In order to facilitate a robust ranking behavior, clustering is often combined into a back-off or smoothing framework. This has been successfully applied for document ranking by Raiber and Kurland [2013], and our approach adopts it for the case of concept ranking.

Entity relatedness. The purpose of entity relatedness is to score the strength of the semantic association between pairs of concepts or entities. Similar to concept retrieval, we will refer to concept relatedness instead of entity relatedness in later chapters. The research on this topic dates back several decades [Zhang et al., 2013], and a multitude of approaches have been researched. Among them, we place particular emphasis on measures that use a knowledge base for computing relatedness. We distinguish two main directions: (i) works that use the textual content of the knowledge base [Gabrilovich and Markovitch, 2007; Hoffart et al., 2012], particularly Wikipedia, and (ii) works that exploit the graph structure behind the knowledge base, particularly Wikipedia or Freebase hyperlinks [Milne and Witten, 2008], DBpedia [Hulpuş et al., 2015; Schuhmacher and Ponzetto, 2014].

Topic and document cluster labeling. Other research directions that are closely related to ours are concerned with labeling precomputed topic models [Hulpuş et al., 2013; Mei et al., 2007] and with labeling document clusters [Carmel et al., 2009]. Topic model labeling is the task of finding the gist of a topic resulting from probabilistic topic modeling. Solutions to this related problem make implicit or explicit use of knowledge about words and concepts harvested from a document corpus. Such knowledge is not available for our problem, rendering most of these approaches inapplicable.

2.4 Conclusion

The perspective of linguists and communication scientists has provided a detailed analysis of the message of images and their usage, where the focus of our review is put on the special form of media-iconic images. Furthermore, these studies have shown that framing images in the context of text helps to solve ambiguity in the meaning and the perception of images. Finally, they have spotted several sources of ambiguity, e.g., different semiotic systems due to a cultural and social bias. Considering the principle how meaning is created and the different sources of ambiguity, one might notice that the distinction of media-iconic (non-literal) images from literal images is often not a binary decision. Much more it is a gradient-like

transition from one to the other, as especially non-literal pairs contain literal elements to convey a message. With our concept ranking approach we are addressing this fact, however, to approach the problem of these untrodden paths of understanding the gist, we are talking about the two classes as of a binary decision problem. These findings serve as foundation when creating the gold standard and while developing the approach.

Research with multimodal data has gained a lot of attention, especially due to the great success of deep learning-based approaches in this domain. Despite this trend, multimodal approaches vary a lot. However, they have some important commonalities. There is a strong focus on the literal aspects of the data. Thus, objects and activities that can be seen in an image are annotated with textual labels or described in more detail in a caption. Rarely, the provided self-contained information of the data is enriched with additional knowledge, beyond the insights provided by the training data, e.g., Altadmri and Ahmed [2009] aims at a higher recall in a retrieval process with supplementing the list of annotations with additional synonyms or hyponym/hypernym relations, i.e., a relation of the type *isA* representing child-parent relations, e.g., a Shire horse is a horse. To the best of our knowledge there is no work representing an image-caption pair as entities of a knowledge base to benefit from common sense knowledge in a computationally accessible way. Furthermore, none of the previously mentioned related works have gist annotations, where the information gain is devised a step further. However, all the multimodal related works show to outperform single modality approaches. This observation indicates - at least for the literal data - a complementary nature of images and captions, which results in one of our working hypotheses.

As a novel task is often accompanied by the need of creating at least a corresponding gold standard annotation, which is conducted and/or assessed by humans, we want to study existing datasets and gold standards in detail in the next chapter.

Chapter 3

Data Resources

Compared to other research domains, the field of multimodal modeling, especially the fusion of visual and textual modalities, is relatively young. However, there have already been several datasets created and a lot of benchmarking challenges conducted. Similar to the related work chapter the datasets can be grouped according to their different purposes of generating and retrieving either descriptions or captions.

3.1 Image Datasets

Before starting with an overview of multimodal datasets, unimodal datasets, which as their name already suggests, consider only one modality, will be reviewed first. They have shown to be a good foundation when creating multimodal datasets. Finally, their additional material and own structure (e.g., annotations from lexical hierarchies) already indicate the tight coupling between vision and text.

ImageNet [Russakovsky et al., 2015] is the visual equivalent to WordNet. WordNet [Fellbaum, 1998] itself is a hierarchy of lexical concepts allowing for a computational querying and modeling of linguistic hierarchies and connections. Meaningful concepts, referred to as synsets, where the majority are nouns, are represented with a collection of images in ImageNet, e.g., synset: police dog, definition: any dog trained to assist police, especially in tracking. According to the statistics [1], there are more than 14 million images distributed over nearly 22,000 synsets, following the aim of providing 1,000 images per synset (current average number of images per synset is 645). Around 8% of the images have precomputed descriptors, such as SIFT [Lowe, 2004], and nearly the same amount of images have bounding boxes. The bounding box frames the object depicting the synset itself or higher level

[1] http://image-net.org/about-stats, latest update: 04/30/2010, last access: 04/30/2017

concepts of the synset, e.g., police dogs have bounding boxes referring to the higher level concept dog.

Between 2005 and 2012 the **PASCAL Visual Object Classes (VOC)** challenge has taken place annually. The project had started with two challenges about object detection and classification of images. Similar to the growing number of challenges, the original dataset with 4 object classes and 1,578 images from 2005 has grown with each year. Thus, the final dataset consists of 20 different object classes represented by nearly 12,000 images. Additionally, 27,450 annotated regions of interest (ROIs) are provided, where, comparable to bounding boxes, a rectangle frames an object and assigns a corresponding label to the object frame. Finally, nearly 7,000 segmentations of objects are provided, which aims at a more detailed understanding of object contours and object-to-object delimitation. The PASCAL VOC dataset has been fully or partially used and extended with the missing textual modality in several works from multimodal modeling, e.g., VDR [Elliott et al., 2014], Visual Phrases [Sadeghi and Farhadi, 2011].

The amount and diversity of available image training data has massively grown. Nevertheless, if one does not want to calculate and search for similarities, but wants to train and detect objects or phrases (cf. [Sadeghi and Farhadi, 2011]) in an image instead of global annotations, bounding boxes or segmentations are required to delimit the object from the rest of the image. These annotations as produced or at least checked by a human, thus, being very expensive, are often just provided for a limited vocabulary size. Consequently, there is a lack of adequate training data for less common or more complex object classes, e.g., solar panels.

3.2 Multimodal Datasets

In the following two different classes of multimodal datasets are reviewed. First, datasets where images are accompanied with descriptions or *descriptive* captions. In general a descriptive caption aims at everything that can be seen in the image and that is self-contained, thus, being understandable without prior information or external knowledge. Second, datasets where images are accompanied by *captions from newspapers and media*. These captions can be descriptive, but they can also be non-descriptive (non-literal). In general a non-literal caption is a caption that often complements the information conveyed by the image and that often refers to abstract and not visually recognizable concepts, as found, for instance, in an external knowledge base.

1. A large wild cat is pursuing a horse across a meadow.
2. A lioness chases a black animal with horns.
3. A lioness closes in on its prey.
4. a lioness is chasing a black bison across a grassy plain.
5. The lioness attacks a wildebeest on a plain.

Figure 3.1 An example image with five descriptive captions from the Flickr8k/30k datasets. The captions are created by human annotators and sometimes not completely correct, e.g., wrong object names, spelling errors. (image by Steve Johnson (CC BY-NC-SA 2.0), http://bit.ly/2wBfeol, last accessed 09/21/2017.)

3.2.1 Images and Descriptive Texts

The **Flickr8k** [Hodosh et al., 2013] and its extension **Flickr30k** have 8,092 and 30,000 images, respectively, collected from six different groups from flickr.com. These groups, e.g., Wild-Child (Kids in Action), contain images that show humans and animals performing an activity. However, to guarantee diversity across the dataset the images are manually selected. Within an extensive crowd-worker project for each of the images five different human written descriptive captions were collected. On average the captions have 11.8 words. Only 21% of the images do not have any verb or common verbs (i.e., sit, stand, look) in their descriptions, which reflects the focus of the actions.

Similar to the previous dataset, the **MIR flickr** [2] dataset [Huiskes and Lew, 2008] collects images and texts from Flickr as a resource. In 2008 the MIR flickr consisted of 25,000 images, in 2010 the dataset increased to 1 million. However, instead of complementing the images with the descriptions from Flickr, the tags and exif metadata are provided. Both text types are given in the original version from Flickr and in an edited version, where spaces and punctuations are removed and upper-cases are substituted by lower-cases. The dataset targets at different challenges, such as visual concept/topic recognition, tag propagation, and tag suggestion. Thus, besides the tags a user has assigned to an image, the images are assigned to at least one topic. The annotators can select from a pre-defined list of 10 general and 19 subtopics. Furthermore, they assess whether a topic is relevant or potentially relevant for an image, e.g., to address the fact that the annotator is not fully confident about whether or not a topic is relevant for an image [Huiskes and Lew, 2008].

[2] http://press.liacs.nl/mirflickr/, last accessed 09/22/2017

<TITLE>Salvador - Pelourinho</TITLE>
<DESCRIPTION>A narrow, rising street with colourful houses on both sides, among them a green house with balconies and a white car parked in front of it, and a blue-and-white church on the right; one car is going up the street; there are a few people in the street and a large thundercloud in the background;</DESCRIPTION>
<NOTES>The old town of Salvador (Pelourinho) has been announced world cultural heritage by the UNESCO; the name of the church is Nossa Senhora do Rosário dos Pretos;</NOTES>
<LOCATION>Salvador, Brazil</LOCATION>
<DATE>March 2002</DATE>
<IMAGE>images/00/42.jpg</IMAGE>

Figure 3.2 An example image from the IAPR TC-12 dataset with English annotation, such as title and description. (available free of charge and without any copyright restrictions, last accessed 09/22/2017.)

The **IAPR TC-12** [Grubinger et al., 2006] dataset is also an upstart from an evaluation campaign initiated by ImageCLEF (Cross Language Evaluation Forum). It contains 20,000 images, each with a descriptive caption in three different languages (in fact, not all images have a caption in all languages). Image titles pin down the name of the location or the event. As a part of the images is provided by a travel agency, beside sports, actions, animals and people, also landscapes and cities are represented in the collection. Additionally, to the diversity of the photo motifs, there is a variety in the scenery conditions, e.g., cities in different lighting or different seasons. All textual information is generated by humans and proofed within an additional iteration for correctness according to rules, e.g., cardinality of named objects. The segmented and annotated IAPR TC-12 (**SAIAPR TC-12**) dataset takes a step further and provides segmented images and masks with object labels of those for all 20,000 images. The dataset is supplemented with image features and relationship information between regions in one image.

The **SBU Captioned Photo Dataset** [Ordonez et al., 2011] resulted from a research about automatically generating descriptive captions. It contains 1 million images with corresponding descriptive captions. As the image-caption pairs are collected using a text-based query from Flickr, where noise is a major issue, the pairs are manually and automatically proofed for certain criteria. Thus, in each caption at least one prepositional word, which describes visible relationships or interactions between objects, and at least two terms from a pre-defined term must be used to be accepted. The pre-defined list of terms includes objects,

Images of the ruins of Patras castle on the hill overlooking the city. A view over the city.

Figure 3.3 An example image from the SBU Captioned Photo Dataset. (image by Automatomato (CC BY-SA 2.0), http://bit.ly/2hnf89g, last accessed 09/22/2017.)

attributes, actions, scenes, and other things. Based on this pre-defined list also the initial queries to collect the pairs are manually created.

The **BBC News Database** [Feng and Lapata, 2008] has 3,361 items, each consisting of a triple of an image, a caption, and an article text. These triples are collected from the news website of BBC (http://news.bbc.co.uk/). The dataset aims at the task of automatic image annotation. Instead of using expensive manual annotation of the images, three annotation baselines are proposed. Thus, the baseline annotations consist of top-k annotations according to tf-idf, the document title, and the output from the continuous relevance model [Lavrenko et al., 2004]. For the first and second baseline (tf-idf and title, respectively), only nouns, verbs, and adjectives are considered. The later baseline is trained on image captions.

The **SVCL Cross-Modal Multimedia Retrieval** [3] dataset [Rasiwasia et al., 2010] consists of a selection of around 2,800 featured articles from Wikipedia, which are grouped according to categories, i.e., arts, music, sports. The dataset consists of pairs of one image and one text, where the text is the paragraph text in which the image appears in the original Wikipedia article. All other aspects of a Wikipedia article, e.g., links to other articles, the paragraph title, or the image caption, are discarded in the dataset. Instead, for all images the SIFT [Lowe, 2004] features are provided.

Compared to the days when data was only available on CDs, it is now more than ever feasible to collect large amounts of data. The proof of the data itself and the generation of gold standards or baselines are most often conducted by humans. Summarizing the focus of the presented datasets, the majority of datasets considers literal, thus, descriptive texts (captions or descriptions). Only the BBC News Database is collected from a source (news), which typically plays with metaphoric language, associations derived from additional knowledge, and a stronger dependency on the interplay between the image and the caption.

[3] http://www.svcl.ucsd.edu/projects/crossmodal/, last access: 04/30/2017

Waste bins 'an ID theft goldmine'

Householders are still throwing out too many documents that help criminals steal their identity a survey suggests.

To help solve the problem police and other consumer organisations have launched their second national identity fraud prevention week.

Be careful what you throw out, say police

A bin-raiding test in London found nearly half of the 120 tested homes had thrown away enough information for their identity to be stolen.

The government has estimated that ID fraud cost the UK £1.7bn last year.

Figure 3.4 Example article from the BBC dataset with image and caption (excerpt). No further annotation is provided. (http://bbc.in/2yhEXhV, last accessed 09/22/2017.)

Illustration 16: The Grimaldi's casino created the family's wealth but by the 1880s, Monaco had acquired a reputation as a decadent playground. The contemporary writer Sabine Baring-Gould described its habituées as: "The moral cesspool of Europe."[23]

By the time of Charles III's death in 1889, Monaco and Monte Carlo were synonymous as one and the same place, and had acquired, through gambling, a reputation as a louche and decadent playground of the rich. It attracted everyone from Russian grand dukes and railway magnates, often with their mistresses, to adventurers, causing the small country to be derided by many including Queen Victoria.[24] In fact so decadent was Monaco considered that from 1882, when she first began visiting the French Riviera, Queen Victoria refused to make a courtesy social call at the palace.[25] The contemporary writer Sabine Baring-Gould described the habituées of Monaco as "the moral cesspool of Europe."[23]

The successive rulers of Monaco tended to live elsewhere and visit their palace only occasionally. Charles III was succeeded in 1889 by Albert I. Albert married Lady Mary Victoria Douglas-Hamilton, a daughter of Scotland's 11th duke of Hamilton, and his German wife, a princess of Baden. The couple had one son, Louis, before their marriage was annulled in 1880. Albert was a keen scientist and founded the Oceanographic Institute in 1906; as a pacifist he then founded the International Institute of Peace in Monaco. Albert's second wife, Alice Heine, an American banking heiress who was the widow of a French duke, did much to turn Monte Carlo into a cultural centre, establishing both ballet and the opera in the city. Having brought a large dowry into the family she contemplated turning the casino into a convalescent home for the poor who would benefit from recuperation in warm climes.[26] The couple, however, separated before Alice was able to put her plan into action.

In 1910, the palace was stormed during the Monegasque Revolution. The prince pronounced an end to absolute monarchy by promulgating a constitution with an elected parliament the following year.

Albert was succeeded in 1922 by his son Louis II. Louis II had been brought up by his mother and stepfather, HSH Prince Tasziló Festetics de Tolna, in Germany, and did not know Monaco at all until he was 11. He had a distant relationship with his father and served in the French Army. While posted abroad, he met his mistress Marie Juliette Louvet, by whom he had a daughter, Charlotte Louise Juliette, born in Algeria in 1898. As Prince of Monaco, Louis II spent much time elsewhere, preferring to live on the family estate of Le Marchais close to Paris. In 1911 Prince Louis had a law passed legitimising his daughter so that she could

Figure 3.5 The SVCL Cross-Modal Multimedia Retrieval dataset contains images of featured articles and the affiliated paragraphs in which the images appear. Note: Some elements, such as the caption or titles, of the article are not included in the dataset. The image shows an excerpt of one of the contained featured Wikipedia articles PRINCE'S PALACE OF MONACO (https://en.wikipedia.org/wiki/File:Monte_Carlo_Casino.jpg, Wigulf commonswiki, CC BY 2.5, last accessed 09/22/2017).

Table 3.1 Overview of the different datasets and their additional material, e.g., image object segmentation

Name	Images		Text	
	# of Images	Object Location given?	# per Image	Type
BBC News (Feng and Lapata [2008])	3,361	-	1	Article and Caption
Déjà-Image Caption (Chen and Zitnick [2015])	4,000,000	-	Varies	Descriptive Caption
IAPR-TC12 (Grubinger et al. [2006])	20,000	-	1-5	Descriptive Caption
SAIAPR-TC12 (Escalante et al. [2010])	20,000	Annotated Segments	1-5	Descriptive Caption
Pascal1K (Rashtchian et al. [2010])	1,000	Bounding Boxes	5	Descriptive Caption
VLT2K (Elliott and Keller [2013])	2,424	Region Annotations	3	Descriptive Caption
Flickr8K (Hodosh et al. [2013])	8,108	-	5	Descriptive Caption
Flickr30K (Young et al. [2014])	31,783	-	5	Descriptive Caption
Abstract Senses (Zitnick and Parikh [2013])	10,000	Annotated Segments	6	Descriptive Caption
MS COCO (Lin et al. [2014])	164,062	Annotated Segments	5	Descriptive Caption
SBU1M Captions (Ordonez et al. [2011])	1,000,000	-	1	(Descriptive) Caption
MIR Flickr 25k (Huiskes and Lew [2008])	25,000	-	avg 8.94	Tags
SVCL Cross-Modal (Rasiwasia et al. [2010])	2,866	-	1	Article (Main Section)
ImageNet (Russakovsky et al. [2015])	14,197,122	Bounding Boxes	-	-

None of the mentioned datasets provides gist annotations to pinpoint the message of an image-caption pair, nor do the datasets contain both types of pairs, texts, or images (literal and non-literal).

3.3 Gist Dataset

To conduct the experiments we use the dataset for understanding the message of images covering the topic of non-literal and literal image-caption pairs first introduced in [Weiland et al., 2016]. Previously, there has not been any dataset aiming at the detection of non-literal pairs, consequently, none of the existing datasets included neither non-literal, nor both (literal *and* non-literal) types of pairs [4].

We define the understanding of image-caption pairs as a concept retrieval and ranking task with image-caption pairs as query. The image-caption pair queries are represented by concepts in the knowledge base. We detail the corresponding entity linking gold standard in subsection 3.3.1. As the message of an image is represented by several concepts assigned with a ranking value, we discuss the gold standard for the concept ranking in subsection 3.3.2.

From the newspaper The Guardian, Our World magazine, and the website of the organization 'Union of Concerned Scientists' [5] we collect image-caption pairs related to the topic

[4] Our previous work in [Weiland et al., 2014] provides us only with images.

[5] https://www.theguardian.com; https://ourworld.unu.edu; http://www.ucsusa.org

Table 3.2 Overview of the themes and their amount of instances in the dataset.

Topic	Non-literal	Literal	Instances
Sustainable energy	Wind energy	Windmill	15
Sustainable energy	Solar power	Solar panel	17
Endangered species	Endangered species (rainforest)	Orangutan	18
Endangered places	Coral bleaching	Coral reef	17
Climate change	Flash flood and flooding	Rain and river	17
Climate change	Heat waves and drought	Ground and paddock	18
Deforestation	Deforestation	Rainforest	20
Endangered species	Endangered species (arctic)	Polar bear	19
Pollution	Air pollution	Smokestacks and smog	19
Pollution	Waste	Disposable cups	4

of global warming. We consider six related narrower topics: sustainable energy, endangered places, endangered species, climate change, deforestation, and pollution. Whereas each of the narrower topics have sub-themes used to collect dataset instances, e.g., solar power and wind energy are sub-themes of sustainable energy. Each of the sub-themes have a literal pendant. In Table 3.2 we give a complete overview of the instance per sub-theme distribution.

The collected pairs are non-literal pairs. As these pairs convey gists based on common knowledge, they satisfy the requirement for a realistic, but challenging dataset. Alternative descriptive captions are created for each image to obtain literal image-caption pairs. The result is a balanced collection of 328 image-caption pairs (164 unique images). Compared to other benchmarking datasets it is a small dataset, but the pure number of images is misleading, as one has to consider the ranking of gist nodes (over 8,000 in total) and a diverse concept coverage used in the images and captions of around 800 different entities. Furthermore, this is the first test collection for literal and non-literal image-caption pairs with gold standard gist annotations and simulated object tags [6]. In order to benchmark the proposed approach for selecting and ranking the gist nodes, and to narrow the potential risk of noise given by automatic object detection, we let annotators assign bounding boxes and object labels to the image from a predefined list of concepts. To arrive at a baseline, that is comparable with automatic image object detectors, the list of selectable concepts has to be limited to objects that are in principle depictable. The annotators used a set of 43 different concepts (e.g., Windmill, Solar panel, Orangutan) to annotate the visible objects in the images. The images are annotated with an average of 3.9 concepts per image, which is a total of 640 annotated bounding boxes for the complete dataset (cf. Table 3.3 for the number of occurrences per concept).

[6] https://github.com/gistDetection/GistDataset

Table 3.3 Overview of the number of image objects according to their label, 43 labels with 640 bounding boxes in total, avg. 3.90 per image.

Label	**Amount**	**Label**	**Amount**	**Label**	**Amount**	**Label**	**Amount**
Windmill	72	Coral	23	Bus	3	Crane	1
Vegetation	59	Smoke	17	Cow	3	Kite	1
Person	54	Snow	15	Sign	3	Mountain	1
Floor	53	Fish	13	Smog	3	Roof	1
Building	44	Tree	13	Backhoe	2	Street	1
Solar Panel	38	Car	8	Monument	2	Sun	1
Ocean	31	Cup	8	Straw Bale	2	Water	1
Polar Bear	30	Cooling Tower	6	Traffic Sign	2		
Orangutan	28	Coral Reef	6	Wood	2		
Grass	27	Trunk	5	Bird	1		
Smokestack	27	Bicycle	3	Cloud	1		
Sky	25	Boat	3	Construction	1		

3.3.1 Query Representation: Entity Linking

The set of labels for the images are provided to the annotators. These were selected from the knowledge base. Consequently, for every image label the corresponding entity link is already existing.

The captions, which are collected from the web and created from human annotators, require - due to their complexity and the tendency of language of being ambiguous - another approach to create the initial gold standard entity links. The captions are therefore pre-processed with a typical pipeline from natural language processing containing part-of-speech extraction and lemmatization, resulting in a set of nouns and noun phrases in their lemmatized form. For each of the nouns and noun phrases we provide candidates generated with three different methodologies. For the first method a concept from the knowledge base, which matches the string of the noun or noun phrase, is retrieved. If the result is a page of an ambiguous term, each of its concepts is collected. An annotator decides for a given image-caption for each of its nouns and noun phrases with candidate concepts, whether the concept represents the mentioned noun (binary decision). For ambiguous pages, the question to be answered is which of the candidate concepts best represent the entity mention in the caption. The same approach is conducted for categories in the knowledge base. For the second method the candidates for the entity mentions are generated based on a query likelihood model on the texts associated with the concepts. Again, annotators have to decide whether the concept represents the entity mention in the caption. The third methodology relies on TagMe! [Ferragina and Scaiella, 2010], an external API with the purpose of entity linking. Here, instead of linking already extracted nouns and noun phrases, the whole captions are

used to generate the entity links. Annotators decide about whether the proposed entities represent the marked entity mention. The results of the annotation process of all three methodologies build the gold standard. We are aware of the fact that in the gold standard entity mentions can be linked to multiple concepts in the knowledge base, e.g., one method resolves the orangutan to the main page of ORANGUTAN and another resolves it to the page of BORNEAN ORANGUTAN. However, an entity linking method is considered correct if one of the gold standard links is found. This way, we allow for diverse set of entities, which are more complex, while producing a common sense for those which are less ambiguous. Entity linking strategy optimization is not in the scope of our research, we rather focus on taking advantage of the knowledge base.

3.3.2 Gist Ranking

In order to evaluate the results of our gist selection and ranking, experts select concepts from the knowledge base which best represent the message. Additionally, they grade the concepts based on relevance levels ranging from 0 (non-relevant) to 5 (most relevant). In the following we will refer to concepts with grade 5 as **core gists** and to concepts with level 4 or 5 as **relevant gists**. A pair can only have one concept graded with level 5. This concept represents the most relevant aspect of the gist.

Of the 8,191 non-zero gist node annotations in total (≈25 per pair), 3,100 obtain a grade of 4 or higher. The list of gold standard gists in the dataset are grouped by topical gists, which can be seen as some sort of core gist.

For the non-literal pairs it is often the case, that the core gist corresponds to one of the before mentioned six aspects of the domain of our testbed, such as *Endangered Species*. A corresponding literal core gist is *Orangutan*. Among all relevant nodes in this study: 54.6 % of all gist nodes are entities and 45.4% are categories.

3.3.3 Visual Linking

Finally, in order to provide us with a ground truth to evaluate whether the visual linking as a feature (RQ11) improves the performance of the concept ranking, annotators separately assessed links between nouns and noun phrases from the caption to labels of the objects in the image.

Multiple linking is allowed in an *NxM* manner: As one caption noun can link to several image objects, e.g., an image showing two polar bears and the caption just mentions "polar bears", both of the image objects are linked to "polar bears". In turn, as image objects are not labeled according to their parts, it is allowed to link several caption nouns to one image

object, e.g., the caption nouns "man" and "shoulder" are linked to the corresponding "person" image object. In the dataset there are 640 image labels, with 43 different object categories, 1433 entity mentions for the non-literal captions, and 849 entity mentions for the literal captions, with 801 different entity mentions. 23 of those cannot be mapped to any entity, however, they are considered as candidates for the visual linking.

Annotators are provided with the image, the caption, and an overlay of image objects with their label and the nouns and noun phrase passages in the caption we have previously extracted via the NLP pipeline. The task is to find for the labels in the image a correspondence in the caption, with the caption concept itself being necessarily a depictable (visually recognizable) object. To help the annotators with this assessment they are provided with concepts from ImageNet (cf. Section 3.1 for further details on ImageNet). Consequently, links can be a string match of the lemma forms between image object label and caption noun - a principle that is quite comparable to the entity linking procedure. However, a link can also be something which has a concept hierarchy-based hyponym/hypernym relation, e.g., plant and tree. Therefore, the WordNet hierarchy is also provided to the annotators.

For the literal pairs 523 of the 640 image labels can be connected via manual visual linking to correspondences in the caption. 84 of them are multi-links, e.g., a caption noun is linked to two or more correspondent objects in the image. For the non-literal pairs 297 of such correspondences can be found, with 57 being multi-links.

3.4 Conclusion

Compared to other test collections in computer vision, our gist dataset of 328 “queries” is a rather small collection. However, this is the first test collection for literal and non-literal image-caption pairs with gold standard gist annotations, labeled image object bounding boxes, entity linking of entity mentions in the text and the image, and visual linking. This makes this dataset a valuable source to foster further research on the topic of gist detection and understanding. Finally, it allows for an analysis with respect to the distinction between non-literal and literal image-caption pairs.

Chapter 4

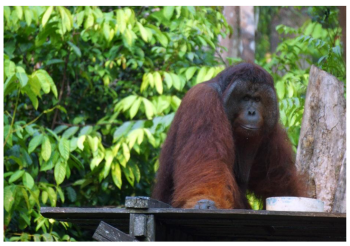
Understanding the Message of Images

4.1 Introduction

Newspaper articles and blog posts are often accompanied by figures, which consist of an image and a caption. While in some cases figures are used as mere decoration, more often figures support the message of the article in stimulating emotions and transmitting intentions. This is especially the case in matters of controversial topics, such as, for instance, global warming, where emotions are conveyed through so-called media icons [Drechsel, 2010; Perlmutter and Wagner, 2004]: images with a high suggestive power that illustrate the topic. A picture of a polar bear on melting shelf ice is a famous example cited by advocates stopping carbon emissions [O'Neill and Smith, 2014]. As such, many image-caption pairs are able to broadcast abstract concepts and emotions [O'Neill and Nicholson-Cole, 2009] beyond the physical objects they illustrate.

As mentioned, previous research in image understanding has focused on the identification and labeling of objects that are visible in the image (e.g., PascalVOC [Everingham et al., 2010], MS COCO [Lin et al., 2014], Im2Text [Ordonez et al., 2011], to name a few, cf. also Section 2.3). Recently, the captionbot system [Tran et al., 2016] was proposed to generate captions for a given image. However, all these approaches focus on the description of what can be explicitly found, i.e., *is depictable*, within pictures [1]. For the example in Figure 1b, captionbot generates the caption: "I think it's a brown bear sitting on a bench.". But despite many research efforts having focused on the so-called problem of bridging the semantic gap in both automatic image and text analysis – namely, the process of replacing low-level (visual and textual) descriptors with higher-level semantically rich ones – few people looked at the

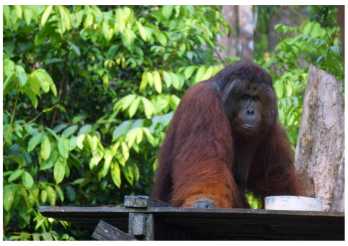
[1] Throughout the paper, we use the terms *depictable* and *non-depictable* to refer to concrete and abstract aspects of image-caption pairs and their gists, respectively.



A male orangutan waits near a feeding station at Camp Leakey in Tanjung Puting National Park in Central Kalimantan province, Indonesia.

Mammals of Southeast Asia, Trees, Orangutans, Botany, Plants

(a) Literal Pairing



Fight to save Indonesia's wildlife corridors key for endangered orangutan.

Habitat Conservation, Biodiversity, Extinction, EDGE species, Deforestation

(b) Non-literal Pairing

Fight to save Indonesia's wildlife corridors key for endangered orangutan.

Habitat Conservation, Biodiversity, Extinction, Politics, Protest

(c) Non-literal Pairing

Figure 4.1 Example image-caption pairs sharing either images or captions with their respective gist nodes (a, b: http://reut.rs/2cca9s7, REUTERS/Darren Whiteside, c: http://bit.ly/2bGsvii, AP, last accessed: 10/20/2016.)

complementary, even more challenging problem of bridging the *intentional* gap, namely understanding the intention behind using a specific image in context [Kofler et al., 2016].

We take a first step towards addressing this hard problem by presenting a method to identify the message that an image conveys, including also abstract (i.e., *non-depictable*) topics. Specifically, we look at the task of identifying and ranking *concepts* that capture the message of the image, hereafter called *gist*. Starting from the visible objects in the image and entity mentions in the caption, we study the use of external knowledge bases for the identification of concepts that represent the gist of the image. Thus, we cast the problem of gist detection as a concept retrieval and ranking task with the following twist: Given an image-caption pair, retrieve and rank concepts from Wikipedia according to how well they express the gist of the image-caption pair.

4.2 The Problem of Image Gist Understanding

Our goal is to understand the gist conveyed by a given image-caption pair. In this work we make a first step in this direction by algorithmically identifying which concepts in a knowledge base describe the gist best. We cast the task of gist detection as a *concept ranking problem* – namely, to predict a ranking of concepts (i.e., Wikipedia articles and categories) from a knowledge base ordered by their suitability to express the gist of a given image-caption pair.

Task Definition: Predict a ranking of concepts from the knowledge base ordered by their relevance to express the gist of the message.

Given: An image instance and its associated textual caption, and knowledge base, viewed as a graph consisting of a vocabulary of concepts or entities (i.e., nodes), associated textual descriptions, and semantic relations between them (i.e., edges).

Output: A ranked list of concepts expressing the message conveyed by the image.

Terminology: seed vs. gist knowledge base nodes. In order to leverage the content and structure of the knowledge base, a link between the objects that are visible in the image, the linguistic expressions (i.e., nouns, proper names) found in the caption and their corresponding concepts in the knowledge base is established using so-called *linking* methods: we call the corresponding nodes in knowledge graph *seed nodes*. Here, we aim to rank the nodes of the knowledge graph based on their relevance to the seed nodes, thus, the initial query. The highly ranked ones then become the gist of the image-caption pair. A node that corresponds to the gist of an image-caption pair is referred to as *gist node*: as a consequence, in this work a gist concept can refer to any node in a given knowledge graph, envisioning any of the general-purpose knowledge bases that are congruent to Wikipedia (including DBpedia [Bizer et al., 2009], YAGO [Hoffart et al., 2013], etc.).

Beyond literal meaning: literal vs. non-literal image-caption pairs. We define an image and its affiliated textual caption as *image-caption pair.* We define further two types of pairs, depending on the kind of message they convey. *Literal* pairs are those in which the caption describes or enumerates the objects depicted in the image. Figure 4.1a) exemplifies a literal image-caption pair. *Non-literal* pairs, of which media icons are an example, are those conveying an abstract message and where images and captions often contain complementary information. Figure 4.1b) shows a non-literal pair with the gist HABITAT CONSERVATION.

We claim that in order to understand the gist of images, both the image and the caption are needed. As they together form a union, an image-caption pair can encode a different gist by changing the caption. Vice versa, combining a caption with a different image can shift the focus of the gist or change the semantics. To illustrate the effects of different image-caption pairs, we show three different pairs as examples in Figure 4.1. Two of the pairs are media icons commonly used to convey the message of species threatened by deforestation. One is a literal pair (cf. Figure 4.1a), which lacks the connection to threat, extinction, and deforestation. The caption describes the image showing an orangutan in what seems to be a national park. The gist of the pair is BORNEAN ORANGUTAN.

By exchanging the caption it becomes apparent that the gist is habitat conservation to save an endangered animal (cf. Figure 4.1b). Considering the corresponding caption thus

helps in the disambiguation of the gist. On the other hand, captions alone are often brief, and when taken out of the context of the image, they fail to convey the entire gist. For instance, by inspecting only the caption Fight to save Indonesia's jungle corridors key for endangered orangutan, it is not clear whether the focus is on endangered species as victims of deforestation, as depicted Figure 4.1b, or on people who fight for habitat conservation, as depicted in Figure 4.1c. That is, only an image can disambiguate the gist. We consequently consider an image-caption pair as the targeted *query* for which gist concepts are ranked.

4.3 Preliminaries

The main idea behind our approach is that a general-purpose knowledge base such as Wikipedia can aid the algorithmic understanding of the message conveyed by image and caption. We hypothesize that the way articles and categories are connected in Wikipedia can be exploited to identify nodes that capture the gist of the image-caption pair. In the following, we shortly explain preliminaries that are applied within our pipeline, but that are not part of our contribution.

4.3.1 Knowledge Graphs

Given a knowledge base, we define a **knowledge graph** as the directed or undirected graph $\mathbf{KG}(V,E,T)$ such that the set of nodes V contains all nodes representing concepts in the knowledge base, every edge $e_{ij} \in E$, $E \subseteq V \times T \times V$ corresponds to a relation in the knowledge base between two nodes v_i and v_j, and the set T contains the relation types in the knowledge base [2].

Node properties. One of the most commonly used properties of nodes is their *degree* [Griffiths et al., 2007]. The degree of a node is the count of all edges that are adjacent to it. Another property of nodes is their tendency of being part of triangles called *local clustering coefficient* [Griffiths et al., 2007]. It is computed as the probability that any two random neighbors of a node are connected themselves.

Our intuition is that these measures help to find a balance between specific and trivial nodes, and thus, the correct gist nodes. The degree and clustering coefficient of nodes are local measures that describe the nodes only in their closest vicinity.

Graph centrality measures. In the domain of network analysis, a wide range of graph centrality measures have been used with the purpose of locating the most important or influential

[2] We consider the knowledge graph to be undirected, unless specified otherwise. Additionally, we denote labeled edges in the graph as (v_i,t,v_j), which is assumed to imply: i) $v_i,v_j \in V$, ii) $(v_i,t,v_j) \in E$, iii) $t \in T$.

nodes in the network. The PageRank [Brin and Page, 1998] scores nodes based on their stationary probability that a random surfer will visit them. Betweenness centrality [Freeman, 1978] defines a node as more important the more often it lies on the shortest path between any two nodes in the graph.

Given a knowledge graph **KG**, as we detail later, our approach makes use of a distance metric $\sigma^{(-1)} : V \times V \rightarrow \Re^{+}$ between two nodes. This metric captures the inverse of a similarity, relatedness, or semantic association measure between the concepts that are represented by the nodes. There are two of the main classes of measures: (i) those based on textual content associated with nodes and (ii) those based on a graph measure. In this work we are interested in using both content and graph structure.

4.3.2 Entity Relatedness

Content-based relatedness. Additionally we incorporate a content-based measure of relatedness. As each node in the DBpedia knowledge graph has a corresponding article in Wikipedia, we leverage a retrieval index on Wikipedia articles.

For a given entity mention, an object tag, or textual representation of the whole image-caption pair, we can use a retrieval model, which uses a query likelihood, to associate a measure of relevance with each node.

Graph-based relatedness. A great variety of semantic relatedness measures have been studied [Zhang et al., 2013]. We follow Hulpuş et al. [Hulpuş et al., 2015] who introduce the exclusivity-based measure, which we use as a node metric $\sigma^{(-1)}$. The authors found that it works particularly well on knowledge graphs of categories and article membership (which we use also) for modeling concept relatedness. It was shown to outperform simpler measures that only consider the length of the shortest path, or the length of the top-k shortest paths, as well as the measure proposed in [Schuhmacher and Ponzetto, 2014].

The exclusivity-based measure assigns a *cost* for any edge $s \xrightarrow{r} t$ of type r between source node s and target node t. The cost function is the sum between the number of alternative edges of type r starting from s and the number of alternative edges of type r ending in t, as shown in Formula 4.1.

$$\text{cost}(s \xrightarrow{r} t) = |\{s \xrightarrow{r} *\}| + |\{* \xrightarrow{r} t\}| - 1, \tag{4.1}$$

where 1 is subtracted to count $s \xrightarrow{r} t$ only once.

The more neighbors are connected through the type of a particular edge, the less informative that edge is, and consequently the less evidence it bears towards the relatedness of its adjacent concepts. By summing up the costs of all edges of a path p, one can compute

the cost of that path, denoted cost(p). The higher the cost of a path, the lower its support for relatedness between the nodes at its ends. Thus, given two nodes, s and t, their relatedness is computed as the inverse of the weighted sum of the costs of the top-k shortest paths between them (ties are broken by cost function). Each path's contribution to the sum is weighted with a length-based discounting factor α:

$$\sigma(s,t) = \sum_{i=1}^{k} \alpha^{\text{length}(sp_i)} \times \frac{1}{\text{cost}(sp_i)} \tag{4.2}$$

where sp_i denotes the ith shortest path between s and t. $\alpha \in (0,1]$ is the length decay parameter and k is a number of shortest paths to consider.

4.4 Methodology

The main idea behind our approach is that a general-purpose knowledge base such as Wikipedia can be used to understand the message conveyed by an image and its caption. To this end, we develop a framework for gist detection based on the following pipeline: First, detected objects in the image and entity mentions in the caption are linked to a reference machine-readable repository of knowledge, i.e., a knowledge base such as the one provided by Wikipedia [Hovy et al., 2013]. Our hunch is to exploit the content and connectivity of the knowledge base, i.e., Wikipedia, which we view as a graph (hence a 'knowledge graph'), in order to identify relevant topics that capture not only the content of the image-caption pair, but also its intended meaning. By using the knowledge base as a graph, we can represent the concepts collected through object detection (in the image) and entity detection (in the caption), as nodes. Next, the neighborhood of these projected nodes in the knowledge graph is inspected to provide a set of candidates of possible gists. Finally, we combine (1) content-based features, extracted from the analysis of Wikipedia text, and (2) graph-based features obtained by analyzing Wikipedia's underlying article-category graph. These features are combined into a node ranking model that pinpoints the gist concepts for a given image-caption pair.

At the heart of our method lies the idea that we can leverage the content and structure of a knowledge base to identify concepts (i.e., Wikipedia pages in our case) that capture the gist of the image-caption pair. Our hunch is that, given a knowledge graph that covers the subject of the image-caption pair, the gist concepts lie in the proximity of those mentioned in the caption or depicted in the image, namely the seed nodes. We define features of candidate gist nodes based on their graph relations or textual content according to their corresponding concept page in the knowledge base. These, in turn, are used to build a supervised ranking

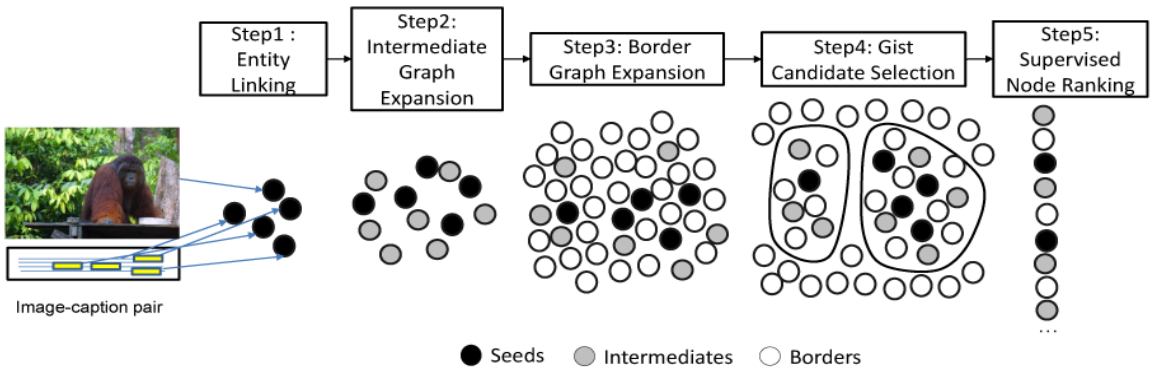


Figure 4.2 Our gist extraction and ranking pipeline (edges between nodes removed for simplicity).

model that is able to rank concepts on the basis of their relevance for the image-caption pair. By using the knowledge base, our method is able to identify gist concepts that are neither visible in the image nor explicitly mentioned in the caption. We expect this hypothesis to be true especially for pairs with an abstract gist, e.g., media icons. Examples of such concepts transmitting the message referred to as gist are GLOBAL WARMING, ENDANGERED SPECIES, BIODIVERSITY, or SUSTAINABLE ENERGY [3]. Despite not being depictable and consequently identifiable by image recognition, the gist nodes will likely be in close proximity in the knowledge graph to the objects in the image that are visible, as well as to the concepts mentioned in the captions.

We present our approach as a pipeline (Figure 4.2). For explanatory purposes, we make use of the media icon of Figure 4.1b to provide us with a running example to illustrate each step of our pipeline.

4.4.1 The Wikipedia and DBpedia Knowledge Graphs

Wikipedia provides a large general-purpose knowledge base [Hovy et al., 2013]. Furthermore, and even more importantly for our approach, the link structure of Wikipedia can be exploited to identify topically associative nodes. In this work, our knowledge graph contains as nodes all articles and categories from the English Wikipedia. As for edges, we consider the following types of relations T, named by their DBpedia link property, which have been previously found to provide useful information for topic labeling [Hulpuş et al., 2013]:

- **Page-category links**: The category membership relations that link an article to the categories it belongs to (e.g., page Wildlife corridor is categorized under WILDLIFE

[3] We use Sans Serif for words and queries, SMALL CAPS for gists, Wikipedia pages and categories.

CONSERVATION). These relations provide topics for different aspects of the concepts described by the Wikipedia page.

- **Super- and sub-category links** The relationship between a category and its parent category (e.g., WILDLIFE CONSERVATION is a sub-category of CONSERVATION), as well as its children categories (e.g., CONSERVATION is a super-category of EDGE SPECIES). Relations between categories can be taken to capture a wide range of topical associations between concepts of different granularities [Nastase and Strube, 2012], including semantic generalizations and specializations [Ponzetto and Strube, 2011].

4.4.2 Step 1: Image and Caption Node Linking

Initially, we project objects depicted in the image and concepts mentioned in the caption onto nodes in the knowledge base. That is, given an image-caption pair (C,I), we want to collect a set of *seed nodes* S in the knowledge graph (Section 4.2). Here, we view the caption as a set of *textual mentions* $C = \{m_1,\ldots,m_n\}$, namely noun phrases that can be automatically extracted using a standard NLP pipeline (in this work, we use the StanfordNLP toolkit [Manning et al., 2014]). Next, each mention $m \in C$ is linked to a set of corresponding concepts from the knowledge graph $V_C \cup \lambda \subset V$, namely the **caption nodes**, or λ, a conventional symbol to represent the 'undefined concept'. The image is viewed as consisting of a set of *object labels* $I = \{l_1,\ldots,l_n\}$ that are either manually given, or are taken from the output of an automatic object detector. Similarly to the captions' mentions, each of the labels $l \in I$ needs to be linked to a set of knowledge graph concepts $V_I \cup \lambda \subset V$, namely the **image nodes**. Finally, we take the union of mapped textual mentions and object labels, namely caption and image nodes as the set of **seed nodes** – i.e., the concepts from the knowledge base that correspond to the entities and objects found in the image-caption pair:

$$S = V_C \cup V_I, \quad S \subset V \tag{4.3}$$

There exist many different ways to link string sequences such as textual mentions (from the caption) and object labels (from the images) to concepts in a knowledge base – i.e., the so-called problem of *entity linking*, which has received much attention in recent years (cf. Chapter 2). Here, we opt for a simple iterative concept linking strategy that is both applicable to captions' mentions and images' object labels, and is particularly suited for short object labels for which no textual context is available to drive the disambiguation process. First, we attempt to link mentions and labels to those Wikipedia articles whose title matches lexicographically, e.g., INDONESIA. Additionally, whenever we find a title of a disambiguation page ORANGUTAN (DISAMBIGUATION), we include all redirected articles

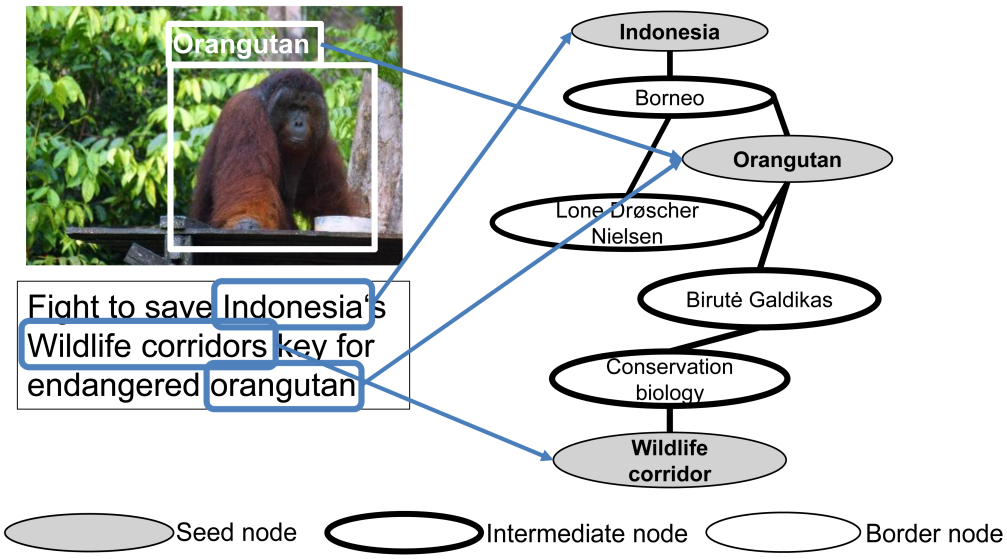


Figure 4.3 Example of intermediate graph for the image-caption pair in Figure 4.1b.

that can be reached with two hops at most from previously linked nodes along the Wikipedia graph. In our experiments (cf. Chapter 5), we demonstrate that this simple approach is, for the purpose of our task, as good as TagMe [Ferragina and Scaiella, 2010], a state-of-the-art entity linking system.

Example. In our working example (Figure 4.1b), objects in the image have been associated with labels like orangutan, sign, trunk, tree, ground, and vegetation. The NLP pipeline, instead, extracted mentions from the caption like fight, Indonesia, jungle, corridor, key, and orangutan. These, in turn, are linked to seed nodes such as INDONESIA and WILDLIFE CORRIDOR, among others (Figure 4.3, depicted in grey).

4.4.3 Step 2: Intermediate Graph Expansion

Especially for media-iconic pairs, one cannot assume that the gist corresponds to any of the concepts found among those obtained by linking either the image labels or the textual captions. For instance, in the case of our example (Figure 4.1b), we cannot find the gist node EDGE SPECIES among any of the seed nodes identified in Step 1 (i.e., those highlighted in gray in Figure 4.3). That is, Step 1 may not be sufficient to identify such gists by simple entity linking, especially in the case of abstract, non-depictable concepts that are rarely mentioned explicitly in the caption.

We operationalize our hypothesis that gist nodes will be found in the knowledge base on paths between seed nodes as follows. We start with the seed nodes from Step 1 and build

a query-specific knowledge graph by extracting all the paths that connect pairs of seeds – similar in spirit to previous approaches for knowledge-rich lexical [Navigli and Lapata, 2010] and document [Schuhmacher and Ponzetto, 2014] understanding.

To produce our semantic graphs, we start with the seed nodes S and create a labeled directed graph $G_I = (V_I, E_I)$ as follows: a) first, we define the set of nodes V_I of G_I to be made up of all seed concepts, that is, we set $V_I = S$; b) next, we connect the nodes in V_I based on the paths found between them in Wikipedia. Nodes in V_I are expanded into a graph by performing a depth-first search (DFS) along the Wikipedia knowledge graph (Section 4.4.1) and successively adding all simple directed paths $v, v_1, \ldots, v_k, v'$ ($\{v, v'\} \in S$) of maximal length L that connect them to G_I, i.e., $V_I = V_I \cup \{v_1, \ldots, v_k\}$, $E_I = E_I \cup \{(v, t_1, v_1), \ldots, (v_k, t_k, v')\}$, $t_i \in T$. As a result, we obtain a subgraph of Wikipedia containing the initial concepts (seed nodes), together with all edges and intermediate concepts found along all paths of maximal length L that connect them. In this work, we set $L = 4$ (i.e., all paths with a length shorter than 4), based on a large body of evidence from previous related work [Hulpuş et al., 2013; Navigli and Ponzetto, 2012a; Schuhmacher and Ponzetto, 2014, *inter alia*]. We call the nodes along these paths, except the seed nodes, **intermediate nodes**, $I = V_I \setminus S$. The graph resulted from combining all the nodes on these paths (including the seeds) as well as the edges of the paths, is what we call the **intermediate graph**:

$$\mathbf{KG}_I(V_I, E_I, T), \quad V_I = S \cup I \tag{4.4}$$

Example. The graph shown in Figure 4.3 is obtained by connecting three concepts, namely ORANGUTAN, INDONESIA and WILDLIFE CORRIDOR with connecting paths found in Wikipedia.

4.4.4 Step 3: Border Graph Expansion and Node Relatedness

The intermediate graph can be used to identify the region of the reference knowledge graph (i.e., Wikipedia) that covers the topics of the image-caption pair. However, while graphs of this kind have been extensively shown to be useful for text lexical understanding [Navigli and Lapata, 2010; Navigli and Ponzetto, 2012b; Ponzetto and Navigli, 2010, *inter alia*], it might still be the case that they do not contain relevant gist nodes – e.g., in the graph in Figure 4.3 we cannot find any of the gists HABITAT CONSERVATION, BIODIVERSITY, EXTINCTION, EDGE SPECIES or DEFORESTATION from our example (Figure 4.1b). We additionally expand the intermediate graph to include all neighbors and their connecting paths that can be reached within two hops from the nodes it contains.

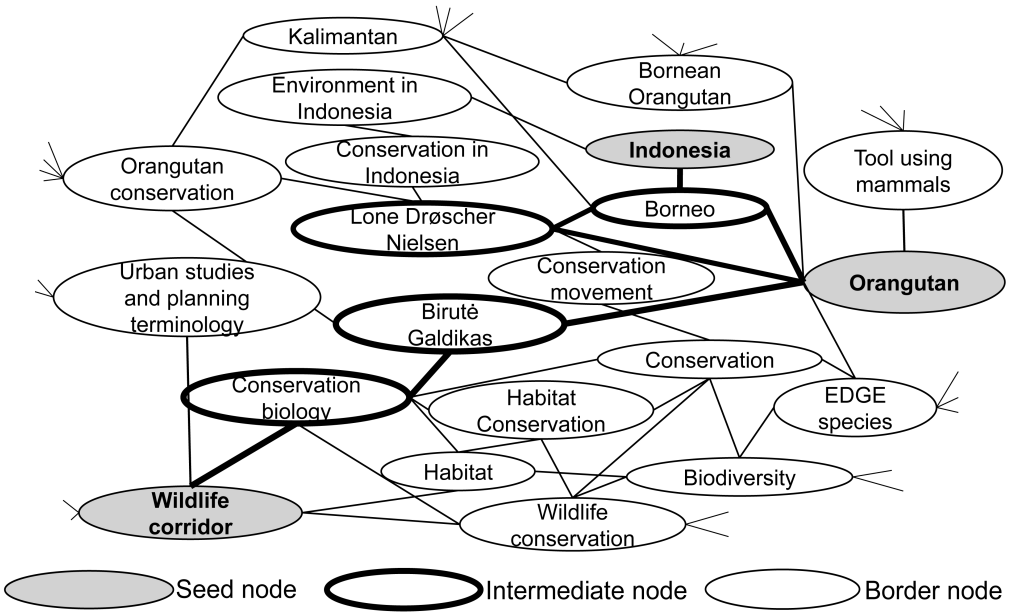


Figure 4.4 Border graph example for the image-caption pair in Figure 4.1b. For simplicity, the image does not make the distinction between article nodes and category nodes, and it also omits edge directions and edge costs.

To expand our semantic graphs we use a procedure similar in spirit to the one we used to create intermediate graphs. We start with the nodes from the intermediate graph V_I and create a labeled directed graph $G_B = (V_B, E_B)$ as follows: a) first, we define the set of nodes V_B of G_B to be made up of all nodes from the intermediate graph by setting $V_B = V_I$; b) next, we expand the set of nodes in V_B using a DFS along Wikipedia, such that for all paths $v, v_1, \ldots, v_k, v'$ ($v \in V_I$, $v' \in V$) of maximal length 2 we set $V_B = V_B \cup \{v_1, \ldots, v_k, v'\}$ and $E_B = E_B \cup \{(v, t_1, v_1), \ldots, (v_k, t_k, v')\}$, $t_i \in T$. The nodes that are added to the graph by the expansion are called **border nodes**, as they lie between the seeds, intermediates, and the rest of the knowledge graph, i.e., $B = V_B \setminus V_I$. We name the resulting graph the **border graph**:

$$\mathbf{KG}_B(V_B, E_B, T), \quad V_B = S \cup I \cup B \tag{4.5}$$

Figure 4.4 shows a part of the border graph obtained from the intermediate graph of Figure 4.3.

Given a knowledge graph **KG**, our approach makes use of a distance metric $\sigma^{(-1)} : V \times V \to \mathbb{R}^+$ between two nodes: this metric captures the inverse of similarity, relatedness, or semantic association measure between the concepts that are represented by the nodes in the knowledge graph. A great variety of semantic relatedness measures have been studied [Zhang et al., 2013]. We follow Hulpuş et al. [2015], who introduce the exclusivity-based measure

that we use here as a node metric. The authors found that it works particularly well on knowledge graphs of categories and article membership (which we use also) for modeling concept relatedness. It was shown to outperform simpler measures that only consider the length of the shortest path, or the length of the top-k shortest paths, as well as the measure proposed in [Schuhmacher and Ponzetto, 2014].

This metric is used in two ways when extracting the top gist candidates:

- Step 4a) clustering the seed and intermediate nodes;
- Step 4b) selecting border nodes close to clusters as gist candidates.

Example. According to Figure 4.4, this leads to a richer semantic graph that includes the concepts HABITAT, BIODIVERSITY, and EDGE SPECIES. Some of these constitute good candidates for gist nodes. Besides, their inclusion affects the pairwise metric distance between the seed nodes. From the sparse information in the intermediate graph (Figure 4.3), it is already clear that INDONESIA is much closer than WILDLIFE CORRIDOR. Now, the border graph encloses a semantic relatedness, which might prefer structural further away concepts or penalize structural closer concepts, such as the mentioned INDONESIA and ORANGUTAN, just because they are connected over paths that are semantically less relevant. However, in our example, the structural short distance is in line with the semantic relatedness: we find that INDONESIA and ORANGUTAN are much closer than WILDLIFE CORRIDOR and ORANGUTAN.

4.4.5 Step 4a: Cluster Seed and Intermediates

After the previous step, we obtain a graph that contains all the concepts from the image and its caption, as well as other concepts from the knowledge graph that lie in close proximity. As previously stated, our assumption is that the gist nodes are part of this graph, and that graph properties will make them identifiable. However, a challenge is that often, an image-caption pair covers multiple sub-topics. These sub-topics represent different aspects of the core topic (cf. core gist) of an image-caption pair, e.g., the core gist HABITAT CONSERVATION has the aspects of habitat conservation in general and region-specific habitat conservation (cf. Fig 4.6, expressed by the clusters in dashed lines). Applying the border graph strategy directly on the seed and intermediate graph in the presence of multiple topics will most often result in a semantic drift and low-quality results.

Consequently, we identify weakly related sub-topics of an image-caption pair by clustering the set of seed and intermediate nodes: for this, we apply Louvain clustering [Blondel et al., 2008], a nonparametric network clustering algorithm, to the border graph $\mathbf{KG}_B$,

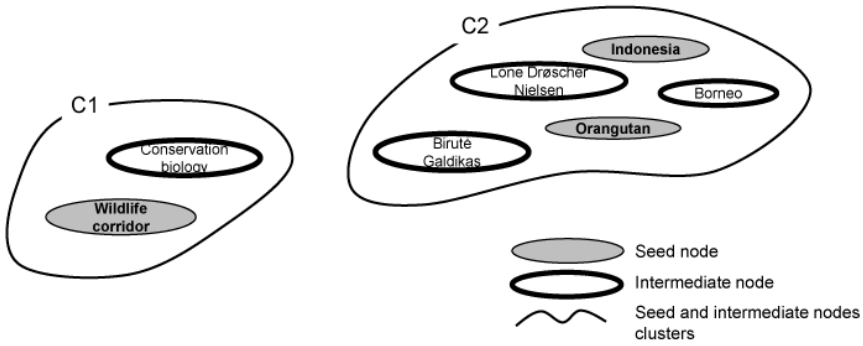


Figure 4.5 Example of seeds and intermediates, clustered based on their pairwise metric.

weighted using our metric σ. The clustering results in groups of seed and intermediate nodes $C = \{C_1, \ldots, C_n\}$ ($C_i \subseteq S \cup I$) that broadly correspond to different sub-topics of the image-caption pair.

Example. Figure 4.5 shows two clusters identified for our example (Figure 4.4): C_1 is about wildlife conservation, containing the seed node WILDLIFE CORRIDOR, and C_2 covers instead topics about Indonesia, including both ORANGUTAN and INDONESIA.

4.4.6 Step 4b: Selecting Gist Candidates

In the next step, we identify suitable border nodes that make good gist candidates. We hypothesize that these are the border nodes that are close to any of the clusters according to the metric σ. We therefore compute for every border node $x \in B$ its average distance $\bar{\sigma}$ to each cluster $C_i \in C$:

$$\bar{\sigma}(x, C_i) = \frac{1}{|C_i|} \sum_{y \in C_i} \sigma(x, y) \tag{4.6}$$

For each cluster C_i, we select its candidate border nodes as the top-k scoring concepts $Gist_{C_i}$. The final set of candidate gist nodes is built as the union of the top-k border nodes across all clusters, together with the set of seed and intermediate nodes:

$$Gist = \bigcup_{C_i \in C} Gist_{C_i} \cup S \cup I \tag{4.7}$$

These nodes constitute the candidate node set which is ranked in the following step.

Example. The association of top-border nodes with the two example clusters is illustrated in Figure 4.6. For instance, the extended wildlife cluster C_1 includes HABITAT and BIODIVERSITY, whereas both ORANGUTAN CONSERVATION and the geographic region KALIMANTAN

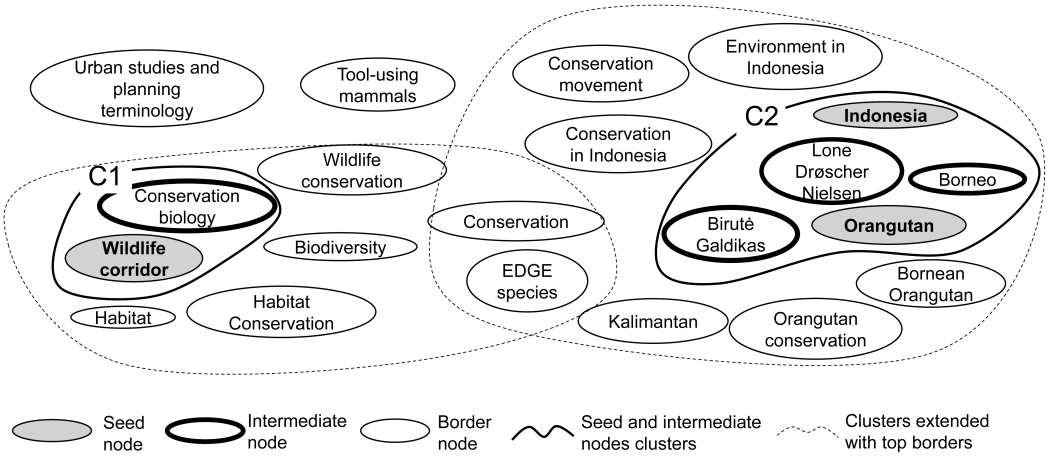


Figure 4.6 Example of clusters of seeds and intermediates, extended with their most related border nodes (top borders). The border nodes that have only weak semantic associations to the clusters are filtered out (e.g., TOOL-USING MAMMALS and URBAN STUDIES AND PLANNING TERMINOLOGY).

are associated with the extended cluster C_2. The border node CONSERVATION is associated with both clusters. These border nodes are included in the candidate set, in contrast to border nodes with a high distance such as URBAN STUDIES AND PLANNING TERMINOLOGY or TOOL-USING MAMMALS which are left out.

Important to note is that the clusters are generated according to the current intermediate graph with the current seed and intermediate nodes. Thus, concepts, which are left out for one image-caption pair, however, can be included for another pair, by changing the caption or the image, e.g., another caption might lead the gist candidate selection method to include TOOL-USING MAMMALS and to leave out CONSERVATION.

4.4.7 Step 5: Supervised Node Ranking

For our task, we train a supervised learning model on labeled data, a method which has been shown to provide robust performance across a wide range of information retrieval and natural language processing tasks [Li, 2011]. Moreover, it provides us with a clean experimental setting to evaluate the contribution of different information sources (i.e., relevance indicators).

The objective of the learning-to-rank method is then to learn a retrieval function such that the computed ranking scores produce the best possible ranking according to some evaluation or loss function. For each of the candidate nodes among those found in the set *Gist*, a feature vector x is created and ranked for relevance with supervised learning-to-rank. Many of our features rely on the topography of the graphs we built as part of our pipeline, including node

degree and *local clustering coefficient* [Griffiths et al., 2007], as well as graph centrality measures like PageRank [Brin and Page, 1998] and betweenness centrality [Freeman, 1978]. Consequently, the feature vector consists of the features listed in Table 4.1 collected from the various steps of the pipeline:

Table 4.1 Features for supervised re-ranking

	Feature	Pipeline	Feature set type				
		Step	seed	intermediate	border	other	baseline
1.	is seed node?	1	✓				
2.	is intermediate node?	2		✓			
3.	Page Rank on intermediate graph	2		✓			
4.	Betweenness centrality on intermediate graph	2		✓			
5.	is border node?	3			✓		
6.	max node-cluster relatedness	4			✓		✓
7.	avg node-cluster relatedness	4			✓		
8.	sum node-cluster relatedness	4			✓		
9.	is member of cluster with most seed nodes?	4			✓		
10.	is member of cluster with most seeds/intermediates?	4			✓		
11.	fraction of seeds in cluster	4			✓		
12.	fraction of seeds and intermediates in cluster	4			✓		
13.	query likelihood on KB text	-				content (text)	✓

Table 4.1 Features for supervised re-ranking

	Feature	Pipeline	Feature set type				
		Step	seed	intermediate	border	other	baseline
14.	Jensen-Shannon divergence on KB text	-				content (text)	
15.	in-degree of node	-				global (KB)	
16.	clustering coefficient	-				global (KB)	

Seed and intermediate features (#1–4, Steps 1–2). Seed and intermediate nodes are distinguished by two binary features. For all the nodes in the intermediate graph, we compute and retain their betweenness centrality and their PageRank score as features.

Border features (#5–12, Steps 3–4). We introduce a feature indicating the border nodes. We leverage information from the clustering step by associating each node with its average proximity $\bar{\sigma}(x, C_i)$ (Equation 4.6) to the nearest cluster C_i. This feature is also used as an unsupervised baseline in the experimental evaluation. Since border nodes can be associated with more than one cluster (cf. CONSERVATION in Figure 4.6) we additionally add features capturing the sum (and average) proximity to all clusters, i.e., $\sum_{C_i \in C} \bar{\sigma}(x, C_i)$.

We assume that the more seed nodes are members of a cluster, the more relevant this cluster is for expressing the gist of the image-caption pair. This assumption is expressed in two features: a binary feature indicating nodes which are members of the cluster containing the highest number of seed nodes; and a fraction of all seed nodes that this node is sharing a cluster with (summing fractions for nodes with multiple cluster memberships). Exploiting the potential benefit of the joint set of seed and intermediate nodes, we further indicate membership of the cluster with the highest number of nodes that are seed or intermediate nodes, as well as the fraction of all seed or intermediate nodes in shared clusters.

Content features (#13–14). We include two content-based similarity measures for image-caption pairs. For this we concatenate all (distinct) entity mentions from the caption and all object annotations from the image as a keyword query. We use the query to retrieve textual content associated with article and category nodes using a query likelihood model with Dirichlet smoothing (cf. study of smoothing methods in language models [Zhai and Lafferty, 2004]). The retrieval model is then used to rank nodes in the candidate set relative to each other. We use this ranking as a baseline for the experimental evaluation and include the reciprocal rank as a node feature. Complementary, in later experiments (cf. Chapter 5, from RQ 8) we also add the Jensen-Shannon divergence, which is calculated between the Wikipedia article texts associated with concepts (category concepts are associated with the text of the equivalent article). The former addresses the topical relevance [Croft et al., 2009] and the later addresses the textual similarity between the texts of query and candidate concepts. As the texts of category pages are very short or consist only of link names to Wikipedia articles, we use a substitution strategy for the categories also in these later experiments: if there is a corresponding article to a category, we use the article text for the content features instead.

Global features (#15–16). Finally, we include global node features that are independent of the image-caption pair. These include the in-degree of a node in the knowledge base (i.e., the number of incoming links for a Wikipedia page or category), as well as its clustering coefficient.

Learning-to-rank model. Our generated feature vectors serve as input for a list-wise learning-to-rank model [Li, 2011]. In a learning-to-rank setting, the image-caption pairs are the set of documents D, with d_i as the i-th document (following the notation of [Li, 2011]). The themes described in Table 3.2 of Section 3.3 are the set of queries, denoted by Q, where q_i is the i-th query. The j-th image-caption pair for a query q_i is represented as a feature vector $x_{i,j} = \phi(q_i, d_{i,j})$ of feature functions ϕ, such as Betweeness Centrality. Finally, $S' = (x_i, y_i)$ represents the training data for q_i, with y denoting the set of labels $\{1, 2, ..., 5\}$. The objective is to find a parameter setting $\widetilde{\phi}$ that maximizes the scoring function: $\widetilde{\pi} = \text{argmax}_{\pi_i \in \Pi_i} S(x_i, \pi_i)$. Specifically, we use RankLib [4], trained with respect to the target metric Mean-Average Precision (MAP). For optimization we use coordinate ascent with a linear kernel [Metzler and Bruce Croft, 2007].

4.5 Conclusion

We presented a knowledge-rich approach to discover the message conveyed by image-caption pairs. We focused on a heterogeneous dataset of literal image-caption pairs – whose topic is described through objects and concepts found in either the picture or the accompanying text – as well as non-literal ones – i.e., referring to abstract topics, such as media-iconic elements found in news articles. Using a manually labeled dataset of literal and non-literal image-caption pairs, we cast the problem of gist detection as a ranking task over the set of concepts provided by an external knowledge base. Specifically, we approached the problem using a pipeline that: i) links detected object labels in the image and entity mentions in the caption to nodes of the knowledge base; ii) builds a semantic graph out of these 'seed' concepts; iii) applies a series of graph expansion and clustering steps of the original semantic graph to include additional, non-depictable concepts and topics within the semantic representation; iv) combines several graph-based and text-based features into a node ranking model that pinpoints the gist nodes.

[4] https://sourceforge.net/p/lemur/wiki/RankLib

Chapter 5

Experiments

In the following, we investigate our proposed approach according to several aspects formulated in twelve different research questions (RQs). Our concept ranking task is benchmarked in RQs 1 through 4, where we first evaluate our entity linking strategy to create the seed nodes (RQ1), look at the suitability of different semantic graphs (RQ2), perform extensive feature analysis (RQ3), and conduct an analysis of the clustering (RQ4). RQ5 and RQ6 evaluate different aspects related to the benefit of filtering gist candidates. RQ7 studies whether gist concepts are in principle depictable (visually recognizable). The role of automatic object detection and caption generation is addressed in RQ8 to RQ11. Finally, in RQ12 we investigate the impact of visual linking between caption and image concepts.

Gold standard. We use the dataset and the gold standard for understanding the gist of non-literal and literal pairs, as described in Chapter 3. In order to benchmark the proposed approach for selecting and ranking the gist nodes, and to narrow the potential of noise given by automatic object detection, we use the manually assigned object labels for the experiments in RQ1–5. For the evaluation of RQ1, additionally, the query representation via entities is used (cf. Section 3.3.1). All learning-to-rank experiments are evaluated on the Gist Ranking described in detail in Section 3.3.2. RQ11 uses the visual links as described in Section 3.3.3. Similar to the simulated object detector RQ11 uses the 'perfect' visual links to evaluate the impact of this feature on the overall gist ranking without additional noise from imperfect linking between image and caption concepts.

Experimental setup. We use a combined knowledge base aligning Wikipedia (WEX dump from 2012), Freebase (from 2012), and DBpedia (from 2014). This knowledge base is used for concept linking, deriving edges for the graph, and the content-based retrieval methods. Concepts in DBpedia are referred to by an URI. The last part of the URI is the same as the last part of the URL of an Wikipedia article, e.g., the URI http://dbpedia.org/

resource/Orangutan and the URL https://en.wikipedia.org/wiki/Orangutan denote the same concept. The URLs of categories in Wikipedia have the prefix 'Category:', DBpedia uses the same prefix, e.g., https://en.wikipedia.org/wiki/Category:Mammals_of_Indonesia and http://dbpedia.org/resource/Category:Mammals_of_Indonesia, respectively. As articles and categories are concepts, and concepts are nodes in the knowledge graphs, the previously described mapping between URLs and URIs can be used to align the nodes from one knowledge graph to the equivalent node of the other knowledge graph. We benefit from the fact that the different knowledge bases provide us with different edges, e.g., DBpedia provides us with typed edges such as "rdf:type", whereas Wikipedia provides us with Wiki links (e.g., links in the Wikipedia article text to another Wikipedia article). These different edges and thus, the different graph structures allow us to calculate a diverse set of features (cf. Table 4.1).

As relatedness measure (Section 4.4.4), we use the metric $\sigma^{(-1)}$ from Hulpuş et al. [2015]. We use their settings for hyperparameters $\alpha = 0.25$ and take the $k = 3$ shortest paths.

Evaluation metrics. We evaluate with five-fold cross validation using standard retrieval metrics such as Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG), and Precision (P), which are calculated until a given rank indicated by @k, e.g., @10. Unless noted otherwise we binarize the assessments to relevant and non-relevant gists.

Baselines. We compare our approach with four different baselines. The first is a content-based method using the texts of Wikipedia article and category pages to construct a query likelihood model with Dirichlet smoothing [Zhai and Lafferty, 2004]. A query likelihood model ranks documents of a collection according to their likelihood of being relevant for the query. The likelihood can be calculated in different ways, e.g., based on the term probabilities (unigram language model). Smoothing is a method to regularize, in that the probability 0 for unseen data is avoided (thus, the chance of overfitting the data is regularized). For a given image-caption pair used as the query, we evaluate the resulting ranking of gist candidates according to the ranking of the probabilities given by the query likelihood model (cf. Table 5.5, Baseline Wikipedia). The second baseline generates a ranking according to the relatedness measure computed in Step 4b. As a candidate node can be a member in several clusters, we consider the maximum relatedness score for the ranking (cf. Table 5.5, Baseline Max relatedness node-cluster). A third baseline randomly ranks the seed nodes instead (Table 5.6, Baseline Random Seeds), so as to assess the need for external knowledge. Finally, as a fourth baseline the confidence values of each detected object from the Microsoft Services API (in the following referred to as **MS tag**) are used to generate a ranking (Table 5.7, Baseline MS tag confidence). This baseline is compared to the experiments using the state-of-the-art object detector.

Table 5.1 Number of image and caption nodes after entity linking.

	Non-Literal		Literal		Overall
	Image	Caption	Image	Caption	
Unique nodes	43	674	43	298	806
Total occurrences	640	1612	640	894	3780

5.1 RQ1: Seed node linking (Step 1) – Which strategy finds the best seed nodes?

We first evaluate the entity linking performance of the simple string-match method used in Step 1 to produce a set of image nodes and caption nodes. These together form the set of seed nodes. We use a separate gold standard to evaluate the correctness of the established links (i.e., not the same gold standard used in RQ2 and RQ3). That is, in order to provide us with a ground truth to evaluate RQ1, annotators separately assessed links between entity mentions from the caption and objects of the image to nodes in the knowledge base, which were validated for correctness.

Image and caption nodes. Table 5.1 shows that a total of 806 different Wikipedia concepts (i.e., pages or categories) are linked across all pairs of images and captions for a total of 3,780 links. Overall, only five noun phrases in captions could not be linked to the knowledge base (e.g., underwater view). Images make use of an object vocabulary of 43 different nodes with a total of 640 links across all images (since each image is manually paired with a literal and a non-literal caption, there is no difference between the columns). We observe a much wider range of nodes when linking entity mentions in the caption. In particular we notice a smaller vocabulary for literal image-caption pairs (298 unique nodes) compared to non-literal pairs (674 unique nodes), where each concept is mentioned about three times on average. However, we find that the caption nodes from literal versus non-literal pairs nearly have no overlap.

Entity linking. The set of seed nodes is given by the union of image and caption nodes. In Step 1 we link object labels and entity mentions to article nodes (String2Article). However, the same procedure could have been applied to category names as well (String2Category). We first compare these two methods to entity links produced by TagMe, a state-of-the-art system [Ferragina and Scaiella, 2010]. Furthermore, we use the retrieval index of texts associated with nodes and output the top ranked node (Wikipedia index). Table 5.2 presents

Table 5.2 Correctness of different entity linking methods for image and caption nodes.

Linking Method	P	R
String2Article	0.9	0.97
String2Category	1.0	0.27
TagMe	0.7	0.83
Wikipedia index	0.81	0.98

precision and recall achieved by these four methods on the set of all 806 unique image/caption nodes. We find that all methods perform reasonably well, where only the category-based linking strategy cannot associate a vast majority of 581 objects/mentions. In particular, we find that our heuristics in Step 1 outperform TagMe and are better in precision than retrieving from the Wikipedia index.

Discussion. The TagMe system poorly performs on our dataset despite being a strong state-of-the-art entity linking system. Manual inspection revealed that TagMe is particularly strong whenever interpretation and association is required, for instance to disambiguate ambiguous names of people, organizations, and abbreviations. In contrast, the concepts we are linking in this domain are mostly common nouns, for which Wikipedia editors have done the work for us already. In the remaining cases that need disambiguation, our heuristic is likely to encounter a disambiguation page. At this point, we are using a well-known disambiguation heuristic by using graph connections to unambiguous contextual mentions/objects. We conclude that our simple entity linking method on articles works much better than on categories and better than TagMe.

Summary of findings. We propose to use objects that have been manually extracted from the image and entity mentions from the caption of the pair, and apply a simple string-matching strategy for linking those objects and entity mentions to nodes, which we call the seed nodes, without direct disambiguation. The actual disambiguation is then implicitly achieved in the subsequent steps of graph traversal and re-ranking. We show that this straightforward “lazy” linking strategy provides comparable results to state-of-the-art algorithms.

Table 5.3 Quality of the gist candidate selection method. Significance is indicated by * (paired t-test, p-value ≤ 0.05).

	Avg cands.	P	R	F1	ΔF1%
Seeds	8.6	0.18	0.16	0.17	0.0
Intermediates	11.4	0.19	0.22	0.21	+19.0%*
Top Borders	31	0.09	0.30	0.14	-21.4%*

5.2 RQ2: Distribution of relevant gist nodes (Steps 2–4) – Which graph is best?

Benefits of graph expansion. We first investigate whether good gist nodes are found in close proximity to the depicted and mentioned seed nodes. To this end, we distinguish proximity in the three expansion layers of seed, intermediate, and top-k border nodes (Step 1, 2, and 4b, respectively), and evaluate the benefits of each graph expansion by studying how precision and recall change with respect to the selection of relevant gist nodes for each expansion step. The results are presented in Table 5.3, where we provide precision, recall, and balanced F-measure, together with the number of average candidates per image-caption pair. In order to judge the significance of improvement for F1 we evaluate the relative increase in precision, on a per-image-caption-pair basis, and report the average (denoted Δ). Significance is verified with a paired-t-test with level 0.05.

We find that especially the expansion into the intermediate graph increases both recall and precision. While the increase in F1 is relatively small, it is statistically significant across the image-caption pairs, where it yields an average increase of 19%. The expansion into the border graph of Step 3 and its contraction to the closest border nodes in Step 4b yields the new set of top border nodes. While it increases recall quite drastically, the loss in precision leads to a significant loss in F1 (over the seed set).

Distribution of high-quality gists. We next change perspective and ask in which expansion set the majority of high-quality gists are found. Initially, we hypothesized that especially for non-literal image-caption pairs, fewer good gists will be found in the seed set, which motivated the graph expansion approach. Accordingly, we separately report findings on literal and non-literal subsets. We study two relevance thresholds in Table 5.4, for relevant gists (grade 4 or 5) as well as a stricter threshold including the core gists (grade 5 only).

Focusing on the distribution of relevant gists, we notice that more than half of the gists are already contained in the seed set, and about 20% are found in the intermediate set. The

Table 5.4 Statistics about proportion of relevant (grade 4 and 5) and core gists (grade 5).

	Grade 4 or 5			Grade 5	
	All	Non-Lit.	Literal	Non-Lit.	Literal
Seeds	53.79%	53.46%	53.96%	57.89%	70.75%
Intermediates	21.05%	21.70%	20.73%	07.89%	17.92%
Borders	25.16%	24.84%	25.32%	34.21%	11.32%

much larger border set still contains a significant portion of relevant gists. Focusing on the differences between literal and non-literal pairs, we find that there are no significant differences between the distributions. Where gists with grade 4 or 5 are highly relevant, they still include the most important visible concepts for non-literal image-caption pairs. However, regarding the distribution of gists with grade 5, we notice that 71% of the high-quality gists in literal pairs are found in the seed set, which is in contrast to only 58% for non-literal pairs. Also, for non-literal image-caption pairs we found the most useful gists in the set of border nodes with a high cluster proximity.

Discussion. We confirm that many relevant and high-quality (grades 4 and 5) gists are found in the seed set and the node neighborhood. The large fraction of nodes available in the border set (compared to the intermediate set) suggests that limiting the intermediate graph expansion in Step 2 to be between seed nodes is too restrictive. We see our initial assumption confirmed in that literal image-caption pairs, which is where most of the related work is focusing on, contain more visible gists, and those are directly visible in the image or mentioned in the caption. For non-literal pairs, the high-quality gists are not only invisible, but also more often only implicitly given. Nevertheless, the graph-based relatedness measures are able to identify a reasonable candidate set.

Summary of findings. We study the distribution of highly relevant gist nodes and whether good gist nodes are found in close proximity to the depicted and mentioned seed nodes. We distinguish proximity in the three expansion layers of seed, intermediate, and top-k border nodes and evaluate the benefits of each graph expansion. We show that while the gist nodes for about half of the studied image-caption pairs are among the seed nodes, for the other half one must look for the gist further away from the seeds, especially for non-literal pairs.

5.3 RQ3: Learning to rank image gists (Step 4–5) – Which features reveal the gist nodes?

We next evaluate the overall quality of our supervised node ranking solution (Section 4.4.7). We further inspect the question of whether features generated by global and local graph centrality measures, especially those derived from border graph expansions, enhance the overall gist node ranking. Moreover, we use our supervised learning-to-rank approach to evaluate the benefit of the feature sets collected over the various steps of our pipeline. To this end, we train a learning-to-rank model using the ground-truth judgments of relevant gist nodes (cf. the previous description of our gold standard). Due to the limited amount of image-caption pairs, we opt for a 5-fold cross validation using each image-caption pair as one "query": in this way we are able to predict 328 node rankings for all image-caption pairs, while keeping training and test data separate.

We study the research question with respect to both, non-literal, and literal pairs and report ranking quality in terms of mean-average precision (MAP), NDCG@10, and precision (P@10) of the top ten ranks. We train and compare four models based on our feature set (Table 4.1): (i) all features, (ii) all features except for the border features, (iii) all features except for the intermediate features, (iv) the subset of border features only. This helps us to understand and assess the different aspects of content and graph-based semantic relatedness. Moreover, we implemented two baselines (using the features highlighted in Table 4.1): One retrieves Wikipedia text using the query likelihood model on all entity mentions and object annotations concatenated, the other is based on an unsupervised ranking according to the maximal node-cluster relatedness measure σ described in Step 4. The results, presented in Table 5.5, are tested for significance (p-value ≤ 0.05).

Table 5.5 Entity ranking results (grade 4 or 5) of supervised learning-to-rank. Significance is indicated by * (paired t-test, p-value ≤ 0.05).

	Both				Non-Literal				Literal			
	MAP	Δ%	NDCG @10	P @10	MAP	Δ%	NDCG @10	P @10	MAP	Δ%	NDCG @10	P @10
All Features	0.69	0.0	0.73	0.7	0.56	0.0	0.6	0.56	0.82	0.0	0.87	0.84
All But Borders	0.66	-4.4*	0.7	0.67	0.54	-6.9*	0.57	0.55	0.78	-4.9*	0.83	0.8
All But Interm.	0.69	-0.3	0.71	0.7	0.56	+0.7	0.57	0.57	0.81	-1.0	0.85	0.83
Only Borders	0.63	-8.7*	0.64	0.64	0.52	-10*	0.54	0.52	0.73	-11*	0.74	0.76
No Clusters	0.70	+1.4	0.74	0.70	0.55	-1.7	0.59	0.54	0.83	+1.2	0.87	0.83
Baselines												
Wikipedia	0.43	-38*	0.48	0.37	0.43	-24*	0.46	0.37	0.44	-46*	0.37	0.49
Max relatedness node-cluster	0.27	-57*	0.57	0.30	0.24	-57*	0.59	0.31	0.31	-62*	0.31	0.54

Overall results. Our approach achieves a relative high ranking performance of 0.69 MAP across all image-caption pairs. As expected, ranking non-literal image-caption pairs is much harder (MAP: 0.56) than for literal pairs (MAP: 0.82). Yet, even in the non-literal case, more than half of the nodes in the top-10 are relevant. Thanks to our approach, we are able to beat the baselines by a large margin. The baseline which ranks nodes by the query likelihood model on all entity mentions and objects achieves a MAP of 0.43 (being 38% worse). The baseline which just includes the max node-cluster relatedness obtains an even worse performance of 0.27 for MAP, even though both achieve the same P@10 performance.

Feature analysis. We next look at the contribution of different types of features (cf. Table 4.1), and compare the performance changes in an ablation study (cf. Table 5.5). When only the border features are used the ranking quality drops significantly, by up to 11%. This indicates the importance of the global graph, intermediate graph, and content-based features. The maximum quality drop is for literal pairs, which indicates that the literal pairs benefit less from the graph expansion to two hops around seeds and intermediates than the non-literal pairs. When we use all the features except the border ones, the performance drops by up to 7%. This drop is stronger for non-literal pairs, reinforcing the fact that non-literal pairs benefit more from the border features. This performance drop cannot be detected when the intermediate features are not considered (not significant), the results are more or less comparable to the results of the complete feature set.

Summary of findings. We thoroughly analyze global and local graph features, as well as content features for image gist ranking using a learning-to-rank approach. We show that the combination of the two types of features achieves the highest accuracy. Furthermore, we show the superiority of our solution in comparison with both supervised and unsupervised baselines. The fact that our full re-ranking pipeline improves so drastically over both a retrieval and a cluster-relatedness baseline demonstrates the benefit of our approach.

5.4 RQ4: What is the impact of clustering the candidate nodes?

As the pipeline is already very complex and consists of different steps, we ask whether we can simplify it further. In this research question we evaluate the need for a clustering of the candidate nodes and investigate if this step changes the gist candidates and the final ranking. We use the same set of features in the same learning-to-rank approach. However, skipping the clusters has a direct impact to the border features (#5-–12), in that they might not differ in their sub-groups, e.g., the features maximum, average, and sum of the relatedness values for each node become the same (#6–8): the sum over all relatedness values for a node.

The MAP for both types of pairs without clustering the candidates is 0.70 (cf. Table 5.5, No Clusters). The MAP results compared to the clustering are better for literal pairs (MAP: 0.82 vs. 0.83) and worse for non-literal pairs (MAP: 0.56 vs. 0.55). However, none of the results are significant.

Summary of findings. There can be no significant difference reported between clustering the candidates or considering all candidates at once. However, the clustering can have a positive impact to the calculation performance, because the calculation of all shortest paths between candidates evolves quadratic in worst case, whereas the clusters reduces the number of pairs to consider. Consequently, we keep the clustering in our pipeline and evaluate all subsequent research questions with respect to features that might be different because of the clustering.

5.5 RQ5: Ranking different sets of candidate gists – Which node types reveal the gist?

The statistics from Table 5.4 indicate that gist nodes are scattered across all sets of concepts gathered throughout our pipeline (i.e., seeds, intermediate and border nodes). Consequently, we next investigate the ability of our supervised model in detecting gists across these different regions: we benchmark this by conducting an ablation study and comparing different sets of candidate gists as input, which are collected from the different regions of our semantic graphs. We evaluate the performance of our learning-to-rank approach on four different node sets: (i) seed nodes, (ii) seed and border nodes, (iii) seed and intermediate nodes, and (iv) all three node types. Across all combinations we only consider the top-k nodes (k = 20). We compare this against a baseline that uses a random subset of the seed nodes.

The results, shown in Table 5.6, indicate that the best MAP scores can be achieved with the complete set of candidate nodes (S, I & B MAP: 0.68), that is, by providing candidate gists as found among all seed, intermediate and border nodes. This observation holds for both - non-literal and literal - types of pairs. Throughout both types of pairs, the candidate set provided by seed and intermediate nodes performs better than the one provided by seed and border nodes. An additional interesting aspect is the performance comparison of the seed nodes with respect to the literal and non-literal pairs, where the MAP for the non-literal pairs is half (MAP: 0.42 vs. 0.21).

Table 5.6 Evaluation of different candidate sets, abbreviated as seeds (S), intermediate (I), and border (B) nodes. Entity ranking results (grade 4 or 5) of supervised learning-to-rank. Significance is indicated by * (paired t-test, p-value ≤ 0.05).

	Both				Non-Literal				Literal			
	MAP	Δ%	NDCG	P	MAP	Δ%	NDCG	P	MAP	Δ%	NDCG	P
Top 20												
S, I & B	0.69	0.00	0.73	0.7	0.56	0.00	0.6	0.56	0.82	0.00	0.87	0.84
S, I	0.57	-17*	0.71	0.68	0.46	-18*	0.58	0.54	0.67	-16*	0.83	0.81
S, B	0.48	-30*	0.65	0.58	0.31	-45*	0.47	0.38	0.64	-20*	0.83	0.78
S	0.31	-54*	0.61	0.52	0.21	-63*	0.43	0.33	0.42	-48*	0.80	0.72
All												
S, I & B	0.56	-18*	0.62	0.61	0.43	-23*	0.50	0.50	0.68	-15*	0.74	0.72
Baseline												
Random Seeds	0.17	-75*	0.41	0.35	0.14	-75*	0.35	0.26	0.2	-76*	0.49	0.35

Summary of findings. We investigate which node types help reveal the gists, and evaluate the performance on four different node sets. We show that the best results are achieved with the complete node set. Furthermore, we confirm our previous finding that it is harder to detect the gist of non-literal image-captions than literal ones. This effect can especially be observed for the seed-only candidate set: even though the amount of relevant gists for non-literal and literal pairs are nearly equal, the non-literal pairs have less core gist within the seeds (cf. Table 5.4), which directly influences the quality of the ranking. This is because the gist of non-literal pairs cannot be found explicitly among the entity mentions.

5.6 RQ6: Filtering candidate gists – Do related concepts better reveal the gist?

Although results for RQ4 show that the best results can be achieved by considering all node types as input to the supervised model, we are still left with the question of whether some candidates are better than others. Consequently, we next look at whether considering only the top-k nodes from the candidate set of seed, intermediate, and border nodes helps improve the results – i.e., by assuming that there are only a few nodes related to the initial query (the seed nodes). We propose to use our relatedness measure [Hulpuş et al., 2013] as an indicator to select the top-k most relevant nodes and filter out distracting ones.

In Table 5.6 we compare the performance using the complete set of candidate nodes (S, I & B, for seed, intermediate, and border nodes in line 5) with a subset obtained by selecting its top-20 elements (line 1). The performance loss of nearly 20% for both types of pairs (MAP: 0.56) indicates the usefulness of using a relatedness measure as a pre-filtering step. The non-literal pairs benefit more from the relatedness-based selection than the literal pairs (Δ%: -23 vs. -15), arguably the hardest subset of data.

Summary of findings. We propose to identify the gists by ranking only the top-k candidates, obtained using a relatedness measure: the results indicate that relatedness-based filtering helps for both image-caption pair types, literal and non-literal.

5.7 RQ7: Finding relevant gist types – Are image gists depictable concepts?

One of the main objectives of our work is to develop a framework to identify the message (gist) conveyed by images and their captions, when used either literally or non-literally (Section 4.2). Consequently, we next investigate the type of concepts that humans find suitable as gists, that is, whether the gist concepts as selected by annotators tend to be depictable or non-depictable. Note that here, we are not looking for the visibility of concepts in a specific image [Dodge et al., 2012], but rather investigate whether the message of the image-caption pair can in general be depicted. Our hypothesis is that the problem of gist detection is particularly challenging for image-caption pairs whose gist is a concept that is not depictable.

For the gold standard pairs, annotators labeled each relevant gist concept as *depictable, non-depictable, or undecided*. An example concept, where the annotators assigned 'undecided', is ARCTIC SEA ICE DECLINE. On average the fraction of depictable core gists is 88% for literal pairs versus only 39% for the non-literal pairs. On the larger set of all relevant gists, 83% are depictable for literal pairs versus 40% for the non-literal pairs. The annotation task, in practice, tends to be rather difficult for humans themselves, as reflected in an inter-annotator agreement (Fleiss' kappa [Fleiss et al., 1971]) of $\kappa = 0.42$ for core gists and $\kappa = 0.73$ for relevant gists.

Summary of findings. We study whether gists are in principle depictable or not. The results of our annotation study are in line with our initial assumption that literal pairs tend to have depictable concepts as gist, whereas the message of non-literal pairs is conveyed through a predominant amount of non-depictable concepts. Generally, this indicates that the core message of images does not necessarily correspond to objects that are depicted, i.e., explicitly to be found within the image: as such, it motivates semantic approaches like ours that aim at going beyond what is found explicitly in the image and accompanying text, to detect the *purpose* for which an image is used.

Table 5.7 Ranking results (grade 4 or 5) according to different input signal and their combination (automatically generated and single signals). 'M' and 'A' indicate manually and automatically produced object labels and caption text, respectively. Significance is indicated by * (paired t-test, p-value ≤ 0.05).

			Both				Non-Literal				Literal			
#	object labels	caption text	MAP	Δ%	NDCG @10	P @10	MAP	Δ%	NDCG @10	P @10	MAP	Δ%	NDCG @10	P @10
Image + caption														
1	M	M	0.74	0.0	0.78	0.74	0.64	0.0	0.69	0.62	0.84	0.0	0.87	0.86
2	M	A	0.48	-35*	0.63	0.56	0.36	-44*	0.45	0.38	0.61	-27*	0.80	0.73
3	A	M	0.43	-42*	0.58	0.53	0.40	-38*	0.49	0.44	0.46	-45*	0.68	0.61
4	A	A	0.14	-81*	0.28	0.23	0.09	-86*	0.17	0.14	0.20	-76*	0.39	0.32
Image only														
5	M	–	0.48	-37*	0.65	0.57	0.28	-47*	0.40	0.33	0.68	-28*	0.89	0.82
6	A	–	0.13	-84*	0.24	0.20	0.06	-90*	0.13	0.11	0.20	-80*	0.35	0.29
Caption only														
7	–	M	0.38	-50*	0.54	0.49	0.31	-52*	0.40	0.35	0.45	-48*	0.67	0.63
8	–	A	0.07	-92*	0.15	0.12	0.05	-93*	0.10	0.08	0.09	-89*	0.19	0.16
Baseline														
9	MS tag confidence		0.02	-97*	0.26	0.05	0.01	-98*	0.15	0.03	0.03	-98*	0.36	0.07

5.8 RQ8: Manual vs. automatic object detection – Do we need manual object labeling?

All experiments carried out so far relied on a gold standard where human annotators manually assigned bounding boxes and object labels to image objects in our dataset. Consequently, we now investigate the performance of our system when the objects in the image are automatically detected with state-of-the-art image annotation tools. We make use of the Computer Vision API [1] from Microsoft Cognitive Services [Fang et al., 2015] – a Web service that provides a list of detected objects and is also capable of generating a descriptive caption of the image. This experiment provides an evaluation of our method in a realistic, end-to-end setting, where images are given with accompanying captions but without manually labeled image object tags.

We first compare the tagging output of the automatic versus manual object labeling. The manual gold standard is based on a vocabulary of 43 different object labels used to annotate 640 instances over the complete dataset. The automatically labeled data amounts to 171 unique object labels used to tag 957 instances. There are 131 overlapping instances between manual and automatic tags, which amounts to less than one shared tag per image, and 20% overlap over the complete dataset.

We next compare the performance on our concept ranking gold standard (Table 5.7, lines 1 vs. 3). A higher performance is achieved with manual tags (MAP: 0.74), as the automatic approach suffers from a mild yet clear decrease in performance (MAP: 0.43). Thus, our experiments show that while there is a certain quality loss in the output predictions, our approach is stable enough to provide useful gists even when applied to the more noisy output of an automatic image annotation system.

Summary of findings. The overlap between automatic and manual image tags is rather low (20%), and the detected objects are not always correct (e.g., a polar bear is detected as a herd of sheep). However, the automatic tags in combination with the human captions lead to a mild drop in performance on gist detection. Thus, the results indicate the viability of framing the gist detection as the proposed concept ranking task in an end-to-end setting.

[1] https://www.microsoft.com/cognitive-services/en-us/computer-vision-api

5.9 RQ9: Manual vs. automatic caption generation – Do we need human captions?

The state-of-the-art image understanding system provided by Microsoft's Computer Vision API is able not only to tag images, but also to generate descriptions of the content of the images. As such, it provides us with a high-performing system to generate image captions [Fang et al., 2015]. Consequently, we next investigate a research question complementary to the previous one, namely how the performance of our method is affected when the caption is automatically generated, as opposed to having been manually produced.

Similarly to RQ7, we first compare the tagging output of the automatic versus manual captions. Performing the entity linking (Step 1, Section 4.4) with the manually created captions results in around 300 and 700 different entities (seed nodes) for the literal and non-literal pairs (Table 5.1). When using the automatically generated captions, these numbers shrink to 130 different detected entities only – thus indicating that the automatic captions are less heterogeneous in meaning than the manual ones. Arguably, this is due to the fact that automatic detectors are trained to produce literal captions. To evaluate the suitability of automatic captions for gist detection, we pair manual or automatic captions with the manual image tags and provide them as input for our pipeline (Figure 4.2). The results, shown in Table 5.7 (lines 1 and 2), indicate that similar to the case of automatic image labeling, our approach suffers from a mild yet clear decrease in performance (MAP: 0.74 vs. 0.48). Finally, we test the performance of the system when using *both* automatic object labels and captions (Table 5.7, line 4): in this case, the dramatic performance decrease (MAP: 0.14) indicates that our method is robust whenever it is provided with at least one signal source (i.e., visual or textual) that is manually produced and cannot cope with purely automatically generated input.

Summary of finding. The overlap between the entities found within automatic and manual captions is low (3-10%). The automatic captions are often short, and the focus of the captions does not always match the focus of the manual caption (e.g., the example in Figure 4.1a receives the caption "There is a bench", without considering the orangutan, although it was detected as a monkey by the automatic image tagging). The results on gist detection, however, are similar to those obtained using automatic image tagging, but drastically drop when providing the system with a purely automatically generated input (i.e., automatically generated image labels and captions).

5.10 RQ10: Manual vs. automatic input – Does an automatic approach capture more literal or non-literal aspects?

In the previous two RQs we benchmarked the performance degradation of a manual versus automatically generated input, i.e., image labels (RQ8) and captions (RQ9). We now turn to the complementary question of which kind of image-caption pairs are better captured by an automatic approach. That is, we investigate the question: is the output of a state-of-art image tagger and caption generator better suited to identify the gist of literal or non-literal image-caption pairs? Again, we rely on the Computer Vision API of Microsoft for such purpose. As shown in Table 5.7 (line 4) using both automatic image tags and captions leads to a moderate ranking for the literal pairs (MAP: 0.20), whereas performance for the non-literal pairs is much lower (MAP: 0.09). This effect is likely due to the fact that the image understanding system we use is trained on a much different kind of data and with a purpose other than detecting (possibly, abstract) image gists. Microsoft Cognitive Service uses, in fact, a network pre-trained on ImageNet and includes a CNN, which assigns labels to image regions trained on Microsoft COCO data [Lin et al., 2014]. Similarly, the language generation uses a language model built using 400,000 (literal) image descriptions.

Compared to the approach that uses the manual assigned image object labels, the 'realistic' approach (i.e., the one using human captions and automatic image labels) has a consistent performance decrease of around 40% for both image-caption pair types (5.7, line 3 and 1 – MAP: 0.43 vs. 0.74). Substituting only the manual captions with the automatic ones (5.7, line 2), instead, results in a lower performance drop than when using automatic image tags for the literal pairs (MAP: 0.61 and 0.46, respectively), but a lower overall performance for the non-literal pairs (MAP: 0.36 and 0.40, respectively). Again, this is likely to be due to the caption generator being able to leverage background knowledge for literal, i.e., descriptive caption generation on the basis of the underlying language model. Such an approach, however, cannot, and is not meant to generate non-literal, topically abstract captions. This finding is even more underlined considering the performance decrease: while the literal pairs have a higher performance drop, when the automatic object detection is combined with the manual captions, compared to the manual object detection and the automatic caption (-45% vs. -27%), the non-literal have a lower decrease when the automatic object detection is combined with the manual captions, compared to the manual object detection and the automatic captions (-38% vs. -44%)

Summary of findings. The evaluation results across all input signal combinations confirm that gists of non-literal pairs are generally more difficult to detect. Automatic approaches can account for descriptive pairs by detecting important objects in the image and describe those in the caption. However, the automatic approaches are not able to produce high-level, abstract image descriptions that are salient to detect the gist of non-literal pairs. That is, to detect the gist of non-literal image-caption pairs, to date, we need to rely on manually produced captions, a requirement that can be dropped to detect the message of literal pairs only.

5.11 RQ11: Visual vs. textual information – Does the image or the caption convey the gist?

Motivated by the significant benefit over single modality approaches, observed from the multimedia community when using different types of modalities fused in one approach, we also look at the role of different kinds of signals within our approach. To this end, we test the performance on gist detection when using only visual (5.7, lines 5–6) or textual (lines 7–8) information separately. For each modality, i.e., visual or textual, we additionally benchmark performance as obtained when using automatically versus manually created image labels or captions. That is, we additionally cast RQs 8 and 9 in a single modality setting.

Given the manual image tags as input signal only (line 5), gist detection on literal pairs suffers from a lower performance drop as when compared to non-literal pairs (MAP: 0.68 and 0.28, respectively). Using automatic object labels (line 6) only additionally lowers performance, with a massive drop for non-literal gists (MAP: 0.06 and 0.20). These very same trends are shown also when using either manually (line 7) or automatically (line 8) generated captions only: using textual information only leads to a high performance decrease for both image-caption pair types (MAP: 0.45 and 0.31), which is even higher in the case of automatically generated captions (MAP: 0.09 and 0.05). Nevertheless, all configurations are able to outperform a baseline obtained by using the Vision API's confidence values for each image directly to establish the ranking (line 9). When investigating different signal sources separately, we are able to corroborate our previous findings that the gists of literal pairs are easier to detect than the gist of non-literal ones. Besides, given that performance substantially decreases when using only the image tags or captions, we show that image and caption are complementary sources of information to effectively detect the message of image-caption pairs. This is in line with many previous contributions from the field of multi-modal modeling that have demonstrated improvements by combining textual and visual signals.

Summary of findings. The evaluation results across different modalities indicate the complementary nature of visual and textual information for detecting the gist of both, literal and non-literal image-caption pairs. That is, by showing that performance on gist detection is reduced when the image tags or only the caption are provided, we show that both image and caption are required in order to capture the message of images.

5.12 RQ12: Visual Linking – Benefit of measuring the descriptiveness of image and caption?

One characteristic of literal pairs is a caption that describes the image, whereas typically a strong focus is put on the most interesting, salient object(s), interaction(s) of object(s), and their attributes etc. In consequence, things that are mentioned in the caption, can typically be seen in the image. Additionally, the predominant amount of depictable gists within literal pairs and the high amount of depictable concepts in the seed nodes (cf. statistical evaluation of RQ 6), motivates this RQ.

We call the connection between a component of a caption, e.g., part(s) of speech, that refers to a component in the image, e.g., an object, a **visual link**. There are different strategies to find and create the visual links, such as the one proposed by [Weegar et al., 2014], who benefit from lexical hierarchies. In contrast to our previous work [Weiland et al., 2015], where we build upon similarity measures and object detector outputs to create a visual link between image objects and caption nouns, we again take a step back and similar to the manual object detection in the images, we use a manual created visual linking (gold standard). This procedure allows us to investigate, whether the gist ranking can be improved with this feature.

As we want to investigate the impact of the visual linking feature to the overall understanding of the gist, we compare with all features, which is the best performing feature combination. We again evaluate the results according to literal, non-literal, and both types of pairs.

Contrasting to the assumption that there is a clear separation between literal and non-literal pairs, the gold standard (cf. Section 3.3) for the visual linking reveals that non-literal pairs tend to encapsulate literal elements. Non-literal pairs can have visual links between objects in the image and parts of the text. Nevertheless, as the amount of non-literal visual links is per pair and on average over all pairs lower than for the literal pairs, we study the respective feature.

Comparing between all features and the combination of all and the visual linking features, shows a significant loss in the performance for both types of pairs when including the visual linking (MAP: 0.74 vs. MAP: 0.70, cf. Table 5.8). The study of the different types of pairs, reveals that the performance degradation is more drastically for the non-literal pairs. The non-literal pairs lose nearly 10% with respect to MAP, when the visual linking feature is considered (MAP: 0.64 vs. MAP: 0.58). Even the literal pairs have a lower MAP of 0.82 than without the visual linking (MAP: 0.84).

Table 5.8 Evaluation of the feature set enhanced with Visual Linking. Entity ranking results (grade 4 or 5) of supervised learning-to-rank. Significance is indicated by * (paired t-test, p-value ≤ 0.05).

	Both				Non-Literal				Literal			
	MAP	Δ%	NDCG @10	P @10	MAP	Δ%	NDCG @10	P @10	MAP	Δ%	NDCG @10	P @10
All Features	0.74	0.0	0.78	0.74	0.64	0.0	0.69	0.62	0.84	0.0	0.87	0.86
All Features + Visual Linking	0.70	-5.4*	0.73	0.69	0.58	-9.6*	0.60	0.56	0.82	-2.1	0.86	0.83

On the first sight it might seem surprisingly that the literal pairs have a decrease in performance. However, there are two important aspects to note. First, there are commonly used phrases in the media, which are not represented by an article or a re-direct, e.g., solar park which often occurs in media does not exist as an article in Wikipedia, hence the visually recognizable object 'solar panel' cannot be linked to the textual phrase 'solar park'. There are concepts that contain the noun phrase 'solar park', e.g., GUJARAT SOLAR PARK, however, as we are using a string-match based concept linking approach the solar park is not matched to one of these instances (cf. RQ1 for the details on concept linking). Consequently, no or misleading visual links can be found. Second, on the vision side there are texture-like image properties conveying important aspects of the gist - such as smog - which can neither be assigned with an object label, nor be delimited via object contours from other objects, as smog just covers the whole scenery.

In turn, the non-literal pairs contain visual links, which also exist in their literal pendants. The usage of correspondences between image and caption components (visual links) follow an explanatory objective: the association which is required to understand the image-caption pair and its gist, is sometimes far to abstract, difficult, or unknown. Supplying the association between the pair and its gist by pointing and explaining the visually recognizable objects, helps to convey the meaning. Despite the fact that encoding the visual-links helps a human to understand the gist of non-literal pairs, it confuses the machine and distracts from learning the gist of literal versus non-literal pairs.

Summary of findings. Even if the visual linking feature does not improve the performance of the gist detection, it helps to understand the nature of image-caption pairs, which convey a meaning. One must note that non-literal pairs contain literal elements and vice versa - which makes the distinction between those two types of pairs even more difficult. However, being aware of these characteristics helps to understand the gist of image-caption pairs.

5.13 Conclusion

Our experiments show that the candidate selection and ranking of gist concepts is a more difficult problem for non-literal image-caption pairs than for literal image-caption pairs. Nevertheless, we demonstrated that using features and concepts from both modalities (image and caption) improves the performance for all types of pairs – a finding which is in line with research on multimodal approaches for other related tasks. Additionally, a feature ablation study shows the complementary nature and usefulness of different types of features, which are collected from different kinds of semantic graphs of increasing richness. We compare manually to automatically gathered information created by automatic detectors. The evaluation is conducted on the complete test collection of 328 image-caption pairs, with respect to the different input signals, signal combination, and single signal analysis. Finally, we experimented with a state-of-the-art image object detector and caption generator to evaluate the performance of an end-to-end solution for our task.

The results indicate that using state-of-the-art open-domain image understanding provides us with an input that is good enough to detect gists of image-caption pairs, with nearly half of the predicted gists being relevant. However, it also demonstrates that improved object detectors could avoid a drop of 38% mean-average precision. Additionally, the caption contains useful hints especially for non-literal pairs. However, without considering the information of the image the performance is significantly degraded.

Chapter 6

Using Gist Detection for Multimedia Indexing - A Use Case

Since gist detection is a novel research problem, devising the existing research directions further, we take a step back and evaluate the performance of the gist detection pipeline in an established research problem, such as multimedia indexing for image and text. Multimedia indexing is about representing multimedia data according to syntactic and semantic features, often accompanied by a classification according to the features to conduct and allow for a retrieval of the respective data. The most important difference between multimedia indexing and gist detection is the missing distinction of literal and non-literal pairs. Furthermore, the topics for semantically classifying the data are often based on the salient objects, which can be seen in the image. This fact does apply for non-literal image-caption pairs, instead, most often the gist of non-literal pairs is something which cannot be visually recognized. In this use case study, we want to demonstrate the benefits and review the disadvantages within a detailed evaluation of the gist detection pipeline applied to an established task such as multimedia indexing. The goal is to point to future work in bridging the gap between established and novel research directions.

6.1 Introduction

Conveying meaning by the use of different modalities is probably as old as human mankind. Nowadays, where everyone takes pictures constantly, storage is cheap, and several platforms allow for the distribution, sharing, and modification of multimodal data, the need of adequately representing the content of such multimodal data is more important than ever. Research on multimedia data, independent of the goal - i.e., whether it is modality generation,

modality retrieval, or representation of the modalities - have shown to perform better with a joint representation of the different modalities. Thus, multimedia researchers motivate the joint modeling by observing ("people often caption an image to say things that may not be obvious from the image itself, such as the name of the person, place, or a particular object in the picture." [Srivastava and Salakhutdinov, 2012, p. 2950]). This is what communication scientists and linguists also report as research results: modalities such as images and text form a communicative unit, with often complementary information [Horn, 1999]. Multimedia indexing, which is in the literature also referred to as classification or representation, requires to adequately represent the respective multimedia data. There are different options on how to represent multimedia data, such as semantic or syntactic classes, by features and similarity measures, or by meta-information.

Representing the multimedia data is relevant to allow for a retrieval which is accurate and precise, while at the same time allowing for a diverse result, in terms of a query, e.g., a user's information need. Sometimes the multimedia representation is a pre-processing of a cross-media retrieval task, where given an image as query, retrieve or generate a caption that best describes the image.

In this work, we focus on multimedia representation. Thus, we make use of the MIR Flickr dataset [Huiskes and Lew, 2008], which provides - besides the images and their affiliated tags - a gold standard multi-label annotation, where the labels represent semantic class topics, such as 'sky' or 'people'. A lot of research has been conducted providing a detailed analysis on the MIR Flickr dataset, these works can be mainly grouped into deep learning, auto-encoder, hashing-based and other approaches. To the best of our knowledge, none of the previous works have made benefit from knowledge bases, such as Wikipedia.

Task: Given an image and its tags, represent this pair by concepts from a knowledge base. Furthermore, create a mapping between the concepts, which represent a pair, to the target class topics.

To represent images, we adapt, modify, and expand the work of Weiland et al. [2016]. Their task is to understand the *gist* (equivalent to meaning or message) of an image-caption pair. Their idea is to represent an image-caption pair by concepts. Based on these initial concepts, additional concepts, which are semantically close to the initial ones are collected from the knowledge base. For each concept a feature vector is generated. This feature vector consists of a diverse set of 15 different features, such as text-similarity or graph-connectivity measures. In a learning-to-rank setting these concepts are ranked. The final ranking represents the gist of an image-caption pair.

There are two modifications in the original understanding the gist pipeline presented in Chapter 4. First, we need to create a mapping between the ranked list of concepts and the

target class topic(s) to conduct the multimedia indexing. Second, we encode only one of the two methods for collecting additional concepts. Namely, the one which searches for commonalities between the initial concepts.

This modified gist pipeline is an end-to-end setting, in that the method relies on out-of-the-box object detectors and a model that has been pre-trained on a dataset with the purpose of gist understanding. We follow two assumptions. First, according to Weiland et al. [2017] the performance for understanding the gist is better when considering information from the image - even though this might include false positive detections from a less precise automatic object detection - than discarding the image information. The second assumption is that training a model for gist detection in one domain, is able to produce a valuable ranking of concepts representing the gist of a query image-text pair of another domain. In deep learning domain adaption as a part of transfer learning has gained more popularity. Thus, our approach benefits from domain adaption in that we make use of a model that has been trained on the annotated gist dataset of Weiland et al. [2016]. Consequently, we do not manually annotate nor rank the concepts representing each of the 25,000 image-text instances of the benchmarking dataset.

In an experimental evaluation, we study the per class and the overall performance of the modified gist pipeline. We compare these results with shallow, deep, hashing, and auto-encoder approaches. We demonstrate the usefulness of the concept representation in that it outperforms shallow and auto-encoder approaches and competes, especially for infrequent represented classes, with deep and hashing approaches. Furthermore, we evaluate the modifications of the pipeline and study the effect of using different sets of concepts, e.g., concepts generated from the Flickr tags only.

We find that not only representing the pair as concepts, but the expansion steps to collect semantically related concepts, has a positive impact to the overall performance. Overall, we confirm that the combination of tags and a deep learning based automatic object detection - even though the latter encodes noise - achieves better performance than single modality representations.

6.2 Methodology

Since we base our method on the pipeline introduced by Weiland et al. [2016] (cf. Fig. 4.2), we focus on our novel changes (cf. Fig. 6.1). However, we start with an overview of the complete pipeline and indicate where required the modifications compared to the original pipeline.

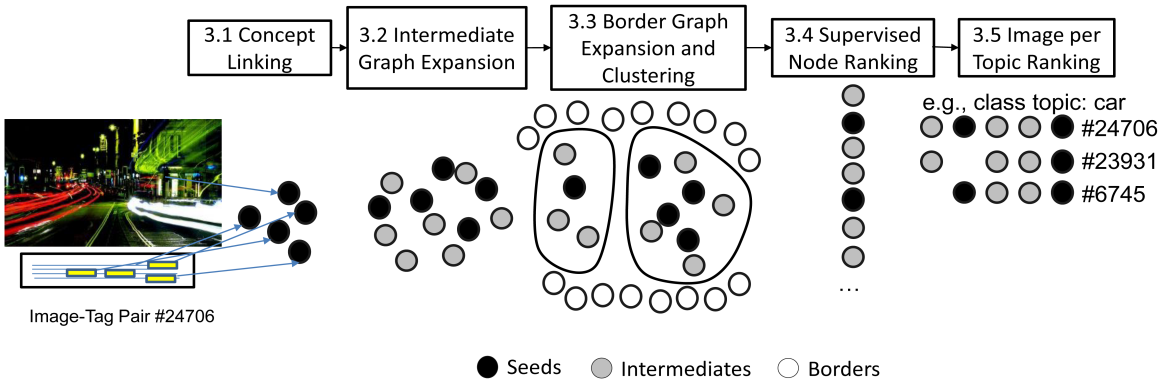


Figure 6.1 The twice-learning-to-rank pipeline for the multimedia indexing task. Image-text pairs are represented as concepts from a knowledge base (Image #24706: Flickr/clare savory, CC BY 2.0).

The main idea of the approach is to represent the images and their affiliated text by concepts from a knowledge base, such as Wikipedia.

In the knowledge base representation of Wikipedia we are using, article pages and categories are the concepts (which is the reason, why we are referring to concept instead of entity linking). Besides the representation of Wikipedia as a knowledge base, we benefit from a representation as a knowledge graph, where categories and articles are nodes in the graph. The article redirects and the category links are the edges in the graph.

Using different strategies, based on semantic relatedness and graph traversal, additional potentially relevant concepts are collected from the knowledge base and added to the set of concepts. In a learning-to-rank (l2r) setting these concepts are ranked. The final ranking represents the gist of an image-text pair. The top-k concepts of such a ranking are transferred to a feature vector. This feature vector serves as input for a second learning-to-rank setting, with which the multi-label classification is conducted and a mapping between gist concepts and class topic(s) is achieved.

6.2.1 Seed Node Linking

There are a few preliminaries for the respective data to fulfill, before the actual concept linking (following the idea of entity linking) can be conducted. The objects in the image, their interactions, and the complete scenery need to be detected. To achieve this, we benefit from a state-of-the-art image processing API, such as Microsoft Cognitive Services [1]. The

[1]Note: By End of 2017 the API is integrated into Microsoft Azure. The images had been annotated before this integration.

result of this API are textual object labels, categories, and captions. The objects name what can be seen in an image, e.g., car. The categories, categorize the scenery of the image, e.g., outdoor. The captions describe how objects in the image in the specified scenery relate to each other, e.g., there is a yellow car on the street. We use the textual object labels as candidates for concept mentions (cf. entity mentions) for the concept linking step.

If the image is accompanied by a text, which is not already tokenized, a NLP preprocessing pipeline consisting of tokenization, lemmatization, noun, and noun phrase detection is applied, whereas the nouns and noun phrases serve as concepts mentions for the concept linking step. Otherwise, if the text is already tokenized, the tokens serve as candidates for concept mentions for the concept linking step. In the MIRflickr25k benchmarking dataset, the Flickr tags directly serve as candidates for the concept linking.

Via a string-match based concept linking, the concept mentions from both, the image and the text, are linked to concepts of a knowledge base. The resulting set of concepts is what is called the seed nodes, consisting of nodes with origin in the image or the textual Flickr tags, in the following we refer to these concepts as S_i and S_t, or to their combination as $S_{t,i}$, respectively.

6.2.2 Intermediate Graph Creation

Under the assumption that the image and the text convey some specific - sometimes complementary - aspect of a common meaning, the next step is to search for semantic connections between the image and its text. Given the seed nodes from the previous step, we collect all category nodes on shortest paths (up to length 4) between all pairs of seed nodes. Therefore edges representing category links are followed. The result is a sub-graph, which consists of seed nodes, the collected category nodes (called intermediate nodes) and the edges connecting the seed and intermediate nodes. In the following we refer to this procedure as the intermediate graph creation and to concepts collected in this step to as I. The gist understanding pipeline consists of two different strategies for collecting additional candidates. The intermediate graph creation is the first strategy. It searches for common concepts on paths, thus semantic commonalities, between the seeds of an instance. In fact, as several instances, which are similar and sometimes differ only slightly, are assigned to a class topic, the intermediate graph creation achieves some kind of normalization between instances of the same class topic. **Example:** One instance has the seed nodes volvo and car, where a node on the intermediate path is Category:Motor vehicles. Another instance has the seed nodes motorcycle and traffic, which are also connected to Category:Motor vehicles. Consequently, the instances which initially did not share a concept, now share a concept.

6.2.3 Border Graph Expansion and Clustering

Given the intermediate graph, we are now collecting all top-k (k = 3) shortest paths up to length 4, between all seed and intermediate nodes in the knowledge graph. The nodes on these paths are called the border nodes and the resulting graph a border graph.

The border graph creation is the second strategy to collect additional concepts. Different to the intermediate graph, which helps to find common concepts and semantic commonality between seeds and instances of the same class, the border graph targets at the fine-grained differences between similar instances. Thus, in contrast to Weiland et al. [2016], we do not include these in the set of candidates for the concept ranking. Instead, we benefit from the borders to achieve a clustering and a valuable relatedness measure computation for each concept. Even though we do not use the border nodes in the set of concepts, preliminary experiments have shown that the border nodes help to produce a better quality in clustering and the relatedness measure for the seed and intermediate nodes. By applying Louvain clustering [Blondel et al., 2008] based on the path based semantic relatedness measure of Hulpuş et al. [2015], we address the fact that an instance can be assigned to several class topics, consequently, the seed concepts are not about one semantic topic.

6.2.4 Ranking the Nodes

For the seed and intermediate nodes, 16 different features encoding graph connectivity and content based measures are calculated: 5 Boolean features (e.g., is seed node?), Page Rank and Betweenness Centrality on intermediate graph, Jenson-Shannon Divergence between the texts of the article pages of a concept, in-degree of the node, clustering coefficient, 3 features with the relatedness measure (max, avg, sum), 2 numerical features about the clusters, and the query likelihood on the text of the article pages of a concept (for further details, please refer to [Weiland et al., 2016]).

As there is no gold standard about the relevance of the so far collected concepts given (recall: a concept in the knowledge base has an equivalent node in the knowledge graph, but as we are not doing a graph ranking, we switch back the notation of concept), we rely on a pre-trained model using the data and gold standard of Weiland et al. [2016] to rank the concepts.

Learning-to-rank model (l2r). Our generated feature vectors serve as input for a list-wise learning-to-rank model [Li, 2011]. In a learning-to-rank setting, the image-caption pairs are the set of documents D, with d_i as the i-th document (following the notation of [Li, 2011]). For each instance the candidate concepts are the set of queries, denoted by Q, where q_i is the i-th query. The j-th image-text pair for a query q_i is represented

as a feature vector $x_{i,j} = \phi(q_i, d_{i,j})$ of feature functions ϕ, such as Betweeness Centrality. Finally, $S' = (x_i, y_i)$ represents the training data for q_i, with y denoting the set of labels $\{1,2,...,5\}$. The objective is to find a parameter setting $\widetilde{\phi}$ that maximizes the scoring function: $\widetilde{\pi} = \text{argmax}_{\pi_i \in \Pi_i} S(x_i, \pi_i)$. Specifically, we use RankLib[2], trained with respect to the target metric Mean-average precision (MAP). For optimization we use coordinate ascent with a linear kernel [Metzler and Bruce Croft, 2007]. The result is a ranked list of concepts for each image-text pair.

6.2.5 Image per Topic Ranking

As the objective is not a ranking of concepts, that best represent the gist of an image-text pair, but a multi-label classification, the ranked list of concepts representing an image needs to be mapped to the class topics, allowing one feature vector to be assigned to multiple classes. There are several ways of creating such a mapping, such as using lexical and conceptual hierarchies (e.g., WordNet), textual similarities and/or pre-trained models of textual similarities etc. The issue with these methods is that each single gist concept needs to be matched to the MIR topics, but tags and images in their single instances are not completely semantically coherent with the topic of a class, e.g., a building will not be matched to the class topic people. However, the combination of the gist concepts make the meaning, e.g., crowd, man, street, building as gist concepts of an image showing a crowded city street that can be assigned to the class topic people.

Hence, we benefit from an additional learning-to-rank approach, with a feature vector representing the ranked gist concepts of an input image-tag pair. Each entry of the feature vector is one concept. If the concept is in the list of the ranked gist concepts for the respective pair, the value of the ranking score of the first l2r setting is its feature value. If it is not in the ranked gist of the respective pair, its value is set to zero. Across all ranked concepts we build a lexicon, where the size of the lexicon is the number of dimensions for the feature vectors, which will be created. Each concept in the vector has a unique id. The feature ids of the ranked concepts, which represent an image, are assigned the relevance values of the respective concept. Given the class as a query, the task is to rank images that fit the class highest.

[2]https://sourceforge.net/p/lemur/wiki/RankLib

6.3 Experimental Evaluation

In the following, we study the performance of framing the multimedia indexing as a modified gist understanding task. We investigate each pipeline step to allow for a detailed analysis of flaws and benefits, where we focus on both, the gist understanding and the dataset particularities. Given an image-tags pair the task is to assign at least one label of the 29 (38 assuming the distinction between potential and relevant topic labels as being different classes) class topics to such a query pair. We provide a comparison to state-of-the-art approaches of the multimedia indexing task conducted on the same MIR Flickr benchmarking dataset. However, only a few evaluate using the complete dataset. Others use only a subset of the data, where either the selection criteria remains unclear and/or they do not report the results according to all evaluation measures, or use a lower rank (e.g., just 20 instead of the first 50 results for precision). Consequently, we compare on a per class topic basis to the results reported by Mark J. Huiskes and Lew [2010]. Additionally, we compare to Chen et al. [2016] and Srivastava and Salakhutdinov [2012] on an overall basis (calculating the evaluation measures across all classes). Finally, we compare the performance of the extended gist pipeline, when considering three different sets of concepts in the gist ranking task: concepts from the seed nodes of the tags (Gist(S_t)), concepts from the seed nodes of the tags and the images (Gist($S_{t,i}$)), and concepts from the seed and the intermediate nodes (Gist($S_{t,i}, I$).

Dataset and Gold Standard. We evaluate the applicability of the modified gist detection on the MIR Flickr dataset [Huiskes and Lew, 2008]. The MIR Flickr dataset contains around 25,000 images from Flickr with textual tags. These image-tag pairs are assigned to 10 general- and 19 subtopics, e.g., sky, clouds (cf. first column in Table 6.1). Whereas an image can be assigned to several of these class topics. Furthermore, there are two levels of classes: the topic is relevant to the image or it is partially/potentially relevant, resulting in 38 class topics. Some of the topics have less image instances than the other, resulting in an instance per topic range between 116 to 10,373 for the topics baby (relevant) and people (potential), respectively (cf. second column indicated with '#' in Table 6.1). In the MIR flickr 25,000 dataset some of the proposed subtopics have not been annotated -these are not considered for the evaluation (i.e., architecture city/urban, building, house, bridge, road/street).

There are no image object annotations provided with the dataset. Thus, even though we extend the dataset with image object annotations generated with the Computer Vision API of Microsoft Cognitive Services [3], we cannot afford a gold standard for these.

At the time of conducting the experiments we do not know, whether the predicted image objects from Microsoft Cognitive Services are correct, but we follow the working hypothesis

[3] https://www.microsoft.com/cognitive-services/en-us/computer-vision-api

Table 6.1 Numbers (#) of images per General Topic and per Subtopic according to type of annotation: potential or relevant (the latter is marked with (r)). Classification performance of the gist detection pipeline per topic according to **MAP** and **P@50**, compared to the four methods from [Mark J. Huiskes and Lew, 2010] (the two low-level settings are abbreviated with LL. Original table from [Mark J. Huiskes and Lew, 2010]).

		MAP					**P@50**		
General Topic	#	Gist($S_{t,i}, I$)	LL LDA	LDA	LL SVM	SVM	Gist($S_{t,i}, I$)	LDA	SVM
sky	7,912	63.33	74.9	80.0	77.5	**82.3**	92.0	87.6	100.0
water	3,331	**58.47**	35.7	57.5	44.8	52.7	92.0	89.6	89.6
people	10,373	61.91	62.8	73.1	63.1	**74.8**	84.0	79.2	99.6
people(r)	7,849	62.32	54.4	**66.4**	55.8	56.5	94.0	79.6	72.4
night	2,711	58.32	51.5	**61.5**	55.4	58.8	92.0	83.2	89.6
night(r)	669	**63.90**	25.2	42.0	39.0	45.0	92.0	62.8	68.8
plant life	8,763	61.45	64.2	**70.3**	68.7	69.1	68.0	83.6	96.4
animals	3,216	**60.30**	26.8	53.7	27.8	53.1	98.0	86.0	96.8
structures	9,992	60.72	61.5	**70.9**	62.6	69.5	80.0	81.2	94.0
sunset	2,135	56.63	42	52.8	58.8	**61.3**	94.0	84.8	98.0
indoor	8,313	60.53	58.2	66.3	60.5	**68.3**	100.0	60.4	89.2
transport	2,895	**63.79**	25.2	41.1	29.8	36.9	86.0	80.0	76.0
food	990	**48.62**	19.6	43.9	29.3	30.8	84.0	86.0	66.8
		MAP					**P@50**		
Subtopic		Gist($S_{t,i}, I$)	LL LDA	LDA	LL SVM	SVM	Gist($S_{t,i}, I$)	LDA	SVM
cloud	3,700	60.65	57.7	65.1	65.1	**69.5**	96.0	87.6	99.2
cloud(r)	1,350	**63.60**	44.5	52.8	51.1	43.4	98.0	77.6	69.6
sea	1,322	**65.95**	25.5	47.7	36.6	52.9	90.0	82.0	91.2
sea(r)	214	**63.61**	9.1	19.7	12.6	20.1	76.0	32.8	28.8
river	894	**67.90**	3.0	31.7	17.9	15.8	90.0	74.8	26.0
river(r)	149	**72.32**	6.9	13.4	10.2	10.9	76.0	20.0	12.4
lake	791	**66.15**	13.9	25.8	18.8	20.7	92.0	59.2	40.0
portrait	3,931	**65.50**	43.2	54.3	49.3	48	98.0	78.0	76.4
portrait(r)	3,829	**59.88**	42.9	54.1	49.3	55.8	98.0	78.0	89.6
male	6,081	**61.13**	35.6	43.4	40.7	41.3	94.0	60.0	58.8
male(r)	3,647	**63.33**	26.7	35.4	29.4	33.5	92.0	58.8	61.2
female	6,148	**58.19**	40.4	49.4	46.1	46.5	96.0	70.4	68.0
female(r)	3,982	**60.62**	34.2	45.4	38.9	45.1	98.0	73.6	75.6
baby	259	**66.31**	6.9	28.5	8.4	20.0	76.0	60.8	39.6
baby(r)	116	**75.65**	6.6	30.8	8.8	16.5	74.0	42.0	20.4
tree	4,683	**61.35**	43.4	51.5	51.4	55.9	90.0	78.8	94.4
tree(r)	668	**62.72**	14.4	34.2	20.5	32.1	98.0	57.2	64.4
flower	1,823	55.74	30.1	**56.0**	46.9	48.0	96.0	92.4	96.0
flower(r)	1,077	41.24	31.8	62.3	51.9	**71.7**	64.0	86.0	98.4
dog	684	**66.97**	10.8	62.1	15.5	60.7	92.0	95.6	96.8
dog(r)	590	62.01	11.2	**66.3**	15.6	64.1	86.0	94.4	92.8
bird	742	**62.54**	9.7	42.6	12.8	44.3	90.0	92.4	94.8
bird(r)	484	**54.70**	9.6	50.0	12.9	52.0	84.0	89.2	88.8
car	1,177	**63.74**	14.2	29.7	17.9	33.9	94.0	73.2	86.8
car(r)	380	**53.24**	12.2	38.9	22.7	43.4	64.0	72.4	75.2

Table 6.2 Number of candidates and seed nodes after the concept linking according to Flickr Tags (text) or image as their origin.

	Total		Unique		Empty Instances
	Candidates	Seed Nodes	Candidates	Seed Nodes	
Tags	223,537	157,473	74,427	34,127	2,128
Images	514,819	487,724	976	964	166

that the method is robust enough to handle eventual noise and overall benefit from the image object labels. This assumption is in line with the findings of the study of Weiland et al. [2017], who have reported an improvement of using automatic object detection combined with the manual given text for an image-caption pair, over just using the text only and discard the visual information.

Evaluation Measures. As there is no explicit gold-standard gist annotation and ranking provided, the output of the candidate gist ranking of the learning-to-rank approach cannot be evaluated directly. However, since the task of multimedia indexing is accomplished as a classification, we benefit from learning and evaluate the mapping between a ranked list of gist candidates to a class. We report the results according to the evaluation measures Mean Average Precision (MAP) and Precision of the first 50 positions of our ranking output (P@50).

6.3.1 Seed Node Linking

There are no gold standard entity links provided in the dataset and it is not feasible to annotate the entity links for all of the 25,000 pairs. Consequently, we provide a statistical overview on how many of the concept candidate mentions are actually linked to a concept in the knowledge base. In Table 6.2 the total number of candidates and the finally total number of linked entities are given, which we refer to as seed nodes. We provide the numbers separately according to the texts and the images. Furthermore, we study how many unique candidates and seed nodes are found across the dataset.

The sum of all tags for all image instances (25,000) of the dataset, is over 223,000 tags (cf. Table 6.2), which results in an average of 9 tags per image. However, around 2,000 images do not have a tag at all (cf. Table 6.2, Empty Instances). For the images around 500,000 image objects are detected and annotated by the Computer Vision API, which is an average of 20 visually recognizable objects per image. On 166 images no visually recognizable object is

detected, whereas 2 images could not be processed by the API due to an extreme format (e.g., 500px width, 49px height).

For both types of media - image and text - around half of the candidate concept mentions can be linked to actual concepts in the knowledge base. This results in 157,473 and 487,724 seed nodes for the texts and the images, respectively. The statistics about uniqueness across the dataset reveals that the tags are more diverse than the visually recognizable objects: Even though the total number of visually recognizable objects are twice the number of tags, the image objects are from 976 semantic concepts, whereas the tags are from 74,427 semantic concepts.

6.3.2 Multimedia Classification (Multimedia Indexing)

In the gist detection task, [Weiland et al., 2016] have shown that selecting a top-k set of the highest ranked gist concepts, improves the performance of the gist understanding. One image instance is per average assigned with 9 Flickr tags. One of our methods we compare is provided with the Flickr tags only, the other methods are provided with more candidates, e.g., by using the object labels and the tags. Consequently, a normalization to make the methods comparable is required. Therefore, we use the top-k (k@10) ranked gist concepts. These top-10 concepts are used to build the feature vector for the final ranking of the image-tags pairs for a class.

After performing the entity linking for the candidates of the query pair (image and tags), the intermediate graphs are created. 175 image-tag pairs are not represented by an intermediate graph, because none of their candidate mentions are linked to a seed node (empty query). Furthermore, 157 intermediate graphs contain only one seed. These empty and one-seed pairs are left out of the multimedia indexing evaluation.

In the following we study the per class and overall performance. Furthermore, we evaluate variations of the modified gist detection pipeline.

Per Topic Evaluation. The modified gist detection pipeline performs best for five of thirteen topical classes of the general topics, e.g., water with MAP: 58.47 (cf. Table 6.1). The two comparison classification approaches using the low-level image features only, never outperform. Both classifiers LDA and SVM based on the combination of low-level image features and the textual description of visual objects perform best in four classes each.

For the sub-topic classification the extended gist detection pipeline outperforms the four other approaches in 21 of 25 cases (cf. Table 6.1). Even for the four remaining classes the gist detection performance is pretty close to the performance of the best approach, e.g., flower, Gist Detection MAP: 55.74 vs. LDA MAP: 56.0 (cf. Table 6.1). Again, the classifiers, which

Table 6.3 Classification performance of the different gist detection pipeline settings across all topics according to **MAP** and **P@50**, compared to the two best methods from [Mark J. Huiskes and Lew, 2010] (the two low-level settings are discarded here. Original numbers from [Mark J. Huiskes and Lew, 2010]), the random baseline, the Deep Boltzmann Machine (DBM) (numbers and approach of [Srivastava and Salakhutdinov, 2012]), and the 32-bit robust multi-label hashing (RMLH) of [Chen et al., 2016].

Method	Gist($S_{t,i}, I$)	Gist($S_{t,i}$)	Gist(S_t)	LDA	SVM	Random	RMLH, 32-bit	DBM
MAP	61.42	45.56	35.87	49.2	47.5	12.4	65.6	52.6
P@50	88.3	53.71	49.18	74.5	75.8	12.4	-	79.1

are using the combination of features perform best for the four remaining classes, whereas the low-level approaches never perform best.

Overall Performance Evaluation. We compare the overall classification results of our gist detection to five other methods from related works, relying on their results without re-implementation. We provide a per class evaluation, compared to only one work [Mark J. Huiskes and Lew, 2010], as the others do not provide a per class evaluation.

The extended gist detection pipeline outperforms all of the comparison approaches in terms of precision (88.3, cf. Table 6.3). Additionally, all but one approach are outperformed according to Mean Average Precision (MAP: 61.42) by the extended gist detection pipeline (cf. Table. Only, the multi-label hashing method performs better (MAP: 65.6). However, there is no result reported for this approach with respect to precision.

Modified Gist Pipeline - Variation Comparison. We study three different variations of the modified gist pipeline. The modified gist pipeline (Gist($S_{t,i}, I$)), applies the gist concept ranking to the nodes collected from both types of seed nodes and the intermediate nodes (cf. Section 6.2, S_t and S_i, and I). This is our main approach, which we have used in the experiments from the previous subsections.

A variation is to leave the intermediate graph creation out and to apply the clustering and ranking border graph produced by the set of seed nodes only (Gist($S_{t,i}$)). This variation is studied to evaluate whether the modified gist pipeline with all of its required steps does increase the performance of the multimedia indexing task.

The third variation is to leave out the intermediate graph creation and the seed nodes from the image. This variation - called Gist(S_t) - allows to study the question, whether the multimedia indexing can be achieved in the simplified pipeline without the intermediate graph creation step and with the seed nodes linked from the Flickr tags only (and discard any visual information).

The lexicon of Gist($S_{t,i}, I$) consists of 1,538 different top-10 concepts. This is an average of 40.5 concepts of representing one class. The lexicon of Gist($S_{t,i}$) and Gist(S_t) contain 5,809 and 11,755, respectively. The higher number of concepts for Gist(S_t) underlines the diversity across the concepts using the tags only (cf. Seed Node Linking) - which might not be the best characteristic to find semantic similarity.

That the main of our methods (Gist($S_{t,i}, I$) has the smallest number of different concepts, confirms the idea which is described in Section 6.2.2 of finding common concepts with the modified gist understanding pipeline. Studying the concepts which most often appear as representation for an instance in the main method, reveals true semantic connections between the concepts and the class topics, e.g., for the class car the concepts are Category:Automobiles, Category:Roads in Lahore, Category:Streets in Los Angeles County, California. However, it also reveals weaknesses, as the concept Category:Rainbow appears in nearly every class. The reason is that the object detector is highly confident with assigning color labels, such as white, blue, green.

The comparison of all three variations shows that the extended gist pipeline performs best in terms of MAP and P@50 (MAP: 61.42 and P@50: 88.3, cf. Table 6.3). Comparing the MAP results (MAP: 45.56 vs. 35.87) of the two simplified pipelines Gist($S_{t,i}$) and Gist(S_t), indicate the benefit of including information from the image even though it might contain false positive detections. This observation is also confirmed by the result of P@50: 53.71 vs. 49.18).

This study demonstrates that all pipeline steps and the automatic object detection, which allows for an end-to-end approach, increase the performance in the multimedia indexing task. Finally, it confirms the benefit of understanding the gist of multimedia pairs, such as images and texts, in established research domains.

Summary of Findings. The evaluation of the modified gist detection pipeline has shown to perform well on established tasks, such as multimedia classification. Especially the detailed per class evaluations have demonstrated the robustness of our approach with respect to classes with fewer or very few instances. In nearly 40% of the general topic and 84% of the sub-topic cases, the extended gist detection pipeline outperforms the comparison approaches. The classifier approaches LDA and SVM, which use the combination of low-level and visual concepts as features, are the best performing comparison approaches. Nevertheless, these encode the textual visual concepts as features, which highly overlap with the target classes, e.g., sky is both a visual concept and a target class topic. Comparing the overall performance, the gist detection pipeline outperforms the Deep Boltzmann Machine approach and is comparable to the robust multi-label hashing approach (cf. Table 6.3, DBM and RMLH, respectively), where for the latter one no result is given for precision. Finally, it can

be shown that the modified gist pipeline helps to find common concepts for instances of the same class topic and that these common concepts are semantically close to the target class topics. As future work, optimizations towards initial concept filtering should be conducted to lower the influence of concepts that occur in several classes with a very high frequency, such as the mentioned color problem (cf. Modified Gist Pipeline - Variation Comparison) e.g., by tf-idf filtering over the complete collection. Furthermore, an in-depth study on how many concepts are required to represent an image-text pair and its gist would provide useful insights not only to multimedia indexing, but also to other domains, e.g., image understanding or question-answering.

6.4 Conclusion

To lower the barrier of fostering novel research tasks, such as the one of gist detection and understanding far beyond the literal meaning of images, we have shown that established research questions with multimedia data benefit from understanding the gist. Understanding the gist is framed as: Given an image-text pair represented by concepts from a knowledge base, collect semantically-related additional concepts, and finally rank these concepts, according to which best represent the gist. To achieve this goal typically two strategies of collecting additional concepts are combined. The first searches for semantic commonalities between the concepts representing the pair and the second adds concepts, which are highly related to these concepts and which allows for a fine-grained distinction between very similar pairs. We have modified this gist pipeline to benefit from the strategy of finding semantic commonalities, while conducting a clustering based on relatedness. Both aspects address the characteristic of the multimedia indexing task: instances can be assigned to several classes, as they convey different semantic meanings, however, at the same time, they share semantics with instances assigned to the same class. As we make use of a benchmarking dataset, we are able to compare the achieved performance of the modified gist detection pipeline to other methodologies, ranging from low-level features with SVM classification to deep-learning based approaches. Overall, the modified gist pipeline reaches a comparable performance or even outperforms the other approaches, e.g., DBM, SVM, LDA, while being less affected of classes with small numbers of training data.

Chapter 7

Conclusion

In this thesis we formulate the novel task of understanding the gist of image-caption pairs using a reference external source, i.e., a knowledge base. The task defines the gist of an image represented by a ranked list of concepts with the image-caption pair - also represented as concepts - as query. As reference knowledge base the proposed approach utilizes Wikipedia. Besides concepts and their structure in the knowledge base, the proposed approach benefits from graph connectivity and relatedness measures, and different content-based measures, such as Jenson-Shannon divergence between Wikipedia article texts, generated with the texts affiliated to a concept in the knowledge base. Within a detailed analysis and evaluation, we have shown that the understanding of gist, thus, the initial candidate concept selection and the concept ranking for non-literal pairs, is an even harder problem. However, within an experimental study of applying the approach to the problem of indexing image-text pairs, we have shown the usefulness of gist detection.

In Chapter 3, we have revised image and multimodal datasets and proofed the observation that previously there has not been any dataset combining literal and non-literal pairs, and providing gist annotations. In turn, and based on research we have conducted before, we built a novel dataset, addressing the task of understanding the gist of image-caption pairs. Therefore we collect from news media sources non-literal image-caption pairs related to the topic of global warming. For each of the images of the collected pairs, annotators provide an alternative descriptive caption. As these pairs convey gists based on common knowledge, it is a realistic, but challenging dataset. In order to allow for an extensive study of the problem formulation of understanding the gist, the dataset is provided with a gold standard addressing several aspects. The objects in the images are assigned with bounding boxes and textual labels, allowing to conduct evaluations of methodologies to understand the gist, without the influence of noise from automatic image object detectors. Representing the gist as concepts from a knowledge base requires at some point the mapping from the initial query

to concepts in the knowledge base, known as entity linking. To benefit the most from the knowledge base and its structure as a graph, the entity linking step is directly in the beginning of the application. Consequently, the image-caption pair is represented by concepts in the knowledge base. The gold standard provides human evaluated entity linking annotations for each image-caption pair. The most important part of the gold standard, is provided with the human annotated ranking of concepts representing the gist for an image-caption pair. Even though the dataset is compared with others rather small (328 pairs), the annotators have annotated more than 8,000 gist concepts. Finally, the gold standard provides human labeled annotations whether the concepts are in principle visually recognizable (depictable vs. non-depictable). This annotation is the complementary information to descriptive texts, with the goal of describing what can be seen in an image.

Based on the task definition, we have proposed in Chapter 4 a pipeline which benefits from external knowledge, several graph and content measures. From each pipeline step different features are collected and utilized in a learning-to-rank framework. The evaluation is conducted with a 5-fold cross validation. The results indicate that the gist of non-literal pairs are generally more difficult to detect. Overall, with the proposed approach a reasonable gist ranking for image-caption query pairs can be produced.

As the proposed approach considers very different methodological aspects, we study the impact of the pipeline steps separately. The simple string match-based entity linking approach with graph-based heuristic to resolve ambiguous links, is compared with two state of the art entity linkers, TagMe! and Wikipedia-based retrieval index of texts. The evaluation result indicates whenever interpretation and association in linking entities is required, the two state of the art systems are strong systems. However, using the simple approach performs better in our use case, because the concepts that need to be linked are common nouns, which can easily be found in a knowledge base such as Wikipedia.

In each pipeline step several features are generated. Furthermore, there are content- and graph connectivity-based features that are collected globally. To evaluate the contribution of different types of features, an ablation study, which helps the feature selection is conducted. Overall, the results indicate the relevance of the selected diverse set of different features. The best results can be achieved, when content, graph, locally (on sub-graphs) and globally (on the knowledge base) generated features are combined.

The concepts that represent the image-caption query are used to traverse the knowledge base. Different strategies to generate a sub-graph containing the query (seed) concepts and the gist candidates are realized. To investigate which of these strategies have an impact to the candidate selection, we evaluate on single strategy and strategy subsets. Overall, the

combination of all graph expansion steps performs best, indicating the relevance of close neighbors and related concepts in the proximity of the seed concepts.

Considering the results of previous studies, we investigated on the complete feature and candidate set the question, whether a top-k candidate selection does help to identify the gists. Better gists can be identified, when ranking only the top-k candidates.

Furthermore, we have tested the influence of the clustering to the gist candidate selection. Again, we have compared on the complete feature and node set. The difference for the candidate selection is negligible, nevertheless, clustering lowers the computational cost (cf. all shortest paths calculation on the complete sub-graph vs. on sub-graphs of the clusters).

Encoding the correspondences between image objects and the caption text nouns as feature, which is called visual linking, reveals that across the non-literal pairs literal elements, which follow an explanatory objective, can be found. Vice versa, literal pairs can contain descriptions that are not necessarily visually recognizable or image objects that are not part of the caption. Even if the visual linking feature does not improve the performance of the gist detection, it helps to understand the nature of image-caption pairs and their gists. Finally, it underlines the difficulty of the task of understanding the gist and the gradient-like transition from literal to non-literal, or vice versa.

Even though multimedia studies have proven the usefulness of combining different media types, e.g., images and text, it is also an interesting research direction for gist detection. The evaluation results across different modalities indicate the complementary nature of visual and textual information for detecting the gist of both, literal and non-literal image-caption pairs. Consequently, both image and caption are required to capture the gist.

Instead of using the manual image object labels, in an end-to-end approach, we benefit from the image object detection results of an automatic image object detector (Microsoft Cognitive Services API) instead. The objects in the images are not always detected correctly, nevertheless in combination with the human captions, a reasonable gist ranking can be achieved. These results confirm the strength and robustness of our approach. Substituting the manual captions by results of the caption generation of the same API and combining those with the manual image object labels, reveal the ability of automatic caption generation approaches to detect important objects in an image. Yet, such an approach cannot produce high-level abstract image captions. Both experiments involving state-of-the-art object detection and caption generation, confirm that understanding the gist of non-literal pairs are an even harder problem.

The task of understanding the gist of images, also requires to understand the characteristic of a gist. Thus, different statistics using the gold standard and our pipeline are calculated. Statistics reveal that around half of the relevant gist nodes can already be found among the

seed nodes, for the other half one must look further away from the seeds. This effect is significantly stronger for the core gists of the non-literal pairs - nearly 35% of the core gists are found across the border nodes (nodes that are collected via the relatedness measure). The need for semantic methods when approaching the gist is also confirmed by results to the question whether gists are depictable - thus, visually recognizable - concepts. It can be shown that a predominant amount of gists for non-literal pairs is non-depictable. Investigating the descriptiveness of automatic approaches, we can conclude that automatic approaches account for descriptive aspects of an image-caption pair, but automatic approaches fail to produce high level abstract captions and image aspects, that are necessary to detect the gist of non-literal pairs with its predominant non-depictable (abstract) nature.

These statistics about distribution of gist across all types of nodes, the amount of non-depictable gists, and the amount of literal or non-literal aspects of automatic approaches, are in line with our motivation: it is not sufficient to detect what objects are on an image, but what message the image conveys in combination with the caption - thus, detecting the non-depictable gist. Finally, in an use case scenario, we have shown the benefit of gist detection in established research problems. Within a comparative study, we have shown that gist detection is able to compete with or even outperforms shallow and deep approaches for multimedia indexing. This result is an important finding, as it underlines the need of methods that are capable of representing abstract aspects and additional knowledge, such as associations to broader, complex, or visually not-recognizable domains. Finally, it indicates that gist detection is useful for downstream tasks, some of which are image search and retrieval, or cross-domain search and retrieval, i.e., recommending images for texts.

7.1 Future Tasks and Limitations

Independent from the combinations of features or strategies, the non-literal gist detection remained the most challenging. Even though we report the results on a realistic and challenging dataset, it has a focus on the theme global warming. Thus, it would be interesting to see, whether the reported conclusions also hold true for other themes, especially those with non-literal gists. Based on the use case scenario one can already get a notion for the results of understanding the gist applied to bigger datasets.

Generally, the question exists, whether there is one approach that can handle the gist detection for both types of pairs and satisfy the requirements given by literal and non-literal pairs? This general question is accompanied by the concrete idea that the type of pair first should be identified to select different strategies, which better account for the specifications,

e.g., their distribution across the node types or their depictability, of the gist for literal versus non-literal pairs.

In Chapter 2 we are talking about the linguistic and communication science perspective of the message conveyed by an image-caption pair. There, we find that the message is strongly influenced by social, cultural, and also by time periods. These findings motivated the decision to use as a core component a knowledge base like Wikipedia, which exists in multiple languages and also versions from multiple time periods. Interesting questions are whether the gist changes during time (recall the initial example of smoking trains and smokestacks during the industrial revolution, in these days they might have been less effected by a negative connotation). In turn, the question for cultural and language differences might reveal completely new research directions: whilst some images are without ambiguous meaning, some others might not (e.g., its not always a wedding when people are dressed in white) and maybe our detectors are 'taught' too much in western language and culture.

Gist image identification is a small, yet arguably crucial part of the much bigger problem of interpreting images beyond their denotation. As such, we see this study as a starting point for research on gist-oriented image search and classification, detection of themes in images, and recommending images from the web when writing new articles for news, blogs, or Wikipedia. But even in the simple form of casting image understanding as a concept ranking problem, we see many potential benefits for a wide range of applications: with our method, for instance, large image collections, such as Wikimedia Commons (more than 30 million images) could potentially be explored in a new way by annotating the contained images with (possibly abstract) concepts from Wikipedia. We leave the exploration of such high-end task that could profit from gist detection for future work.

Bibliography

Aletras, N. and Stevenson, M. (2012). Computing similarity between cultural heritage items using multimodal features. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@EACL 2012, 24 April 2012, Avignon, France*, pages 85–93.

Altadmri, A. and Ahmed, A. (2009). Video databases annotation enhancing using commonsense knowledgebases for indexing and retrieval. In *IASTED International Conference on Artificial Intelligence and Soft Computing, ASC 2009, Palma de Mallorca, Spain, September 7–9, 2009*.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379.

Balog, K., Serdyukov, P., and de Vries, A. P. (2010). Overview of the TREC 2010 entity track. In *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010*.

Barbu, A., Bridge, A., Burchill, Z., Coroian, D., Dickinson, S. J., Fidler, S., Michaux, A., Mussman, S., Narayanaswamy, S., Salvi, D., Schmidt, L., Shangguan, J., Siskind, J. M., Waggoner, J. W., Wang, S., Wei, J., Yin, Y., and Zhang, Z. (2012). Video in sentences out. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, pages 102–112.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artifical Intelligence Research*, 55(1):409–442.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia – A crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008:1–12.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Bruni, E., Uijlings, J. R. R., Baroni, M., and Sebe, N. (2012). Distributional semantics with eyes: using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012*, pages 1219–1228.

Carmel, D., Roitman, H., and Zwerdling, N. (2009). Enhancing cluster labeling using wikipedia. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 139–146.

Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., and Trani, S. (2013). Learning relatedness measures for entity linking. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 139–148.

Chandler, D. (1994). Semiotics for beginners. WWW document.

Chen, H., Zhao, Y., Zhu, L., Chen, G., and Sun, K. (2016). Learning robust multi-label hashing for efficient image retrieval. In *Advances in Multimedia Information Processing - PCM 2016 - 17th Pacific-Rim Conference on Multimedia, Xi'an, China, September 15-16, 2016, Proceedings, Part II*, pages 285–295.

Chen, X. and Zitnick, C. L. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2422–2431. IEEE Computer Society.

Croft, B., Metzler, D., and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition.

Dalton, J., Dietz, L., and Allan, J. (2014). Entity query feature expansion using knowledge base links. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2014, Gold Coast , QLD, Australia, July 06 - 11, 2014*, pages 365–374.

Das, P., Srihari, R. K., and Corso, J. J. (2013a). Translating related words to videos and back through latent topics. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 485–494.

Das, P., Xu, C., Doell, R. F., and Corso, J. J. (2013b). A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, Portland, OR, USA, June 23-28, 2013*, pages 2634–2641.

Demartini, G., Firan, C. S., Iofciu, T., and Nejdl, W. (2008). Semantically enhanced entity ranking. In *Web Information Systems Engineering, WISE 2008, 9th International Conference, Auckland, New Zealand, September 1-3, 2008. Proceedings*, pages 176–188.

Demartini, G., Iofciu, T., and de Vries, A. P. (2009). Overview of the INEX 2009 entity ranking track. In *Focused Retrieval and Evaluation, 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009, Brisbane, Australia, December 7-9, 2009, Revised and Selected Papers*, pages 254–264.

Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., and Mitchell, M. (2015). Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 100–105.

Dodge, J., Goyal, A., Han, X., Mensch, A., Mitchell, M., Stratos, K., Yamaguchi, K., Choi, Y., III, H. D., Berg, A. C., and Berg, T. L. (2012). Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2012, Montréal, Canada, June 3-8, 2012*, pages 762–772.

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., and Darrell, T. (2017). Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691.

Drechsel, B. (2010). The berlin wall from a visual perspective: comments on the construction of a political media icon. *Visual Communication*, 9(1):3–24.

Elliott, D. and de Vries, A. (2015). Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 42–52.

Elliott, D. and Keller, F. (2013). Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1292–1302.

Elliott, D., Lavrenko, V., and Keller, F. (2014). Query-by-example image retrieval using visual dependency representations. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 109–120.

Escalante, H. J., Hernández, C. A., González, J. A., López-López, A., y Gómez, M. M., Morales, E. F., Sucar, L. E., Pineda, L. V., and Grubinger, M. (2010). The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428.

Everingham, M., Gool, L. J. V., Williams, C. K. I., Winn, J. M., and Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Fang, H., Gupta, S., Iandola, F. N., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, L., and Zweig, G. (2015). From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1473–1482.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 15–29, Berlin, Heidelberg. Springer-Verlag.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Feng, Y. and Lapata, M. (2008). Automatic image annotation using auxiliary text information. In *Proceedings of ACL-08: HLT*, pages 272–280, Columbus, Ohio.

Feng, Y. and Lapata, M. (2010a). How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, Uppsala, Sweden, July 11-16, 2010*, pages 1239–1249.

Feng, Y. and Lapata, M. (2010b). Topic models for image annotation and text illustration. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2012, Los Angeles, California, USA, June 2-4, 2010*, pages 831–839.

Ferragina, P. and Scaiella, U. (2010). TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 1625–1628.

Fleiss, J. et al. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, page 215.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI 2007, Hyderabad, India, January 6-12, 2007*, pages 1606–1611.

Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., and Lazebnik, S. (2014). Improving image-sentence embeddings using large weakly annotated photo collections. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 529–545, Cham. Springer International Publishing.

Griffiths, T. L., Tenenbaum, J. B., and Steyvers, M. (2007). Topics in semantic representation. *Psychological Review*, 114:211–244.

Grubinger, M., Clough, P., Müller, H., and Deselaers, T. (2006). The IAPR TC-12 benchmark – a new evaluation resource for visual information systems.

Gunter R. Kress, T. v. L. (1996). *Reading Images: The Grammar of Visual Design*. Psychology Press.

Gupta, A., Verma, Y., and Jawahar, C. V. (2012). Choosing linguistics over vision to describe images. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, Ontario, Canada, July 22-26, 2012*, pages 606–612.

Hare, J. S. and Lewis, P. H. (2010). Automatically annotating the MIR flickr dataset: experimental protocols, openly available data and semantic spaces. In *Proceedings of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2010, Philadelphia, Pennsylvania, USA, March 29-31, 2010*, pages 547–556.

Harrison, C. (2003). Visual social semiotics: Understanding how still images make meaning. *Technical Communication*, 50(1).

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Hoffart, J., Seufert, S., Nguyen, D. B., Theobald, M., and Weikum, G. (2012). KORE: keyphrase overlap relatedness for entity disambiguation. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 545–554.

Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61.

Horn, R. E. (1999). Information design: Emergence of a new profession. In Jacobson, R., editor, *Information design*, pages 15–33, Cambridge, MA, USA. MIT Press.

Hovy, E., Navigli, R., and Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Huiskes, M. J. and Lew, M. S. (2008). The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR 2008, pages 39–43, New York, NY, USA. ACM.

Hulpuş, I., Hayes, C., Karnstedt, M., and Greene, D. (2013). Unsupervised graph-based topic labelling using dbpedia. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 465–474.

Hulpuş, I., Prangnawarat, N., and Hayes, C. (2015). Path-based semantic relatedness on linked data and its use to word and entity disambiguation. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I*, pages 442–457.

Irene Hammerich, C. H. (2002). *Developing Online Content: The Principles of Writing and Editing for the Web*. Wiley.

Jeon, J., Lavrenko, V., and Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR 2003, pages 119–126, New York, NY, USA. ACM.

Jewitt, C. and Oyama, R. (2001). Visual meaning: a social semiotic approach. In van Leeuwen, T. and Jewitt, C., editors, *Handbook of visual analysis*, chapter 7, pages 134–156. Sage, London.

Jia, X., Gavves, E., Fernando, B., and Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In *Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015*, pages 2407–2415.

Jiang, L., Kalantidis, Y., Cao, L., Farfade, S., Tang, J., and Hauptmann, A. G. (2017). Delving deep into personal photo and video search. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM 2017, pages 801–810, New York, NY, USA. ACM.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Li, F.-F. (2015). Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 3668–3678.

Karpathy, A., Joulin, A., and Li, F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. *CoRR*, abs/1406.5679.

Karpathy, A. and Li, F. (2015). Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. (2015). Unifying visual-semantic embeddings with multimodal neural language models. *Advances in Neural Information Processing Systems Deep Learning Workshop*.

Kofler, C., Larson, M., and Hanjalic, A. (2016). User intent in multimedia search: A survey of the state of the art and future challenges. *ACM Comput. Surv.*, 49(2):36:1–36:37.

Krishnamoorthy, N., Malkarnenkar, G., Mooney, R. J., Saenko, K., and Guadarrama, S. (2013). Generating natural-language video descriptions using text-mined knowledge. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, Washington, USA, July 14-18, 2013*.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1601–1608.

Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., and Choi, Y. (2012). *Collective generation of natural image descriptions*, volume 1, pages 359–368.

Kuznetsova, P., Ordonez, V., Berg, T. L., and Choi, Y. (2014). Treetalk: Composition and compression of trees for image descriptions. *TACL*, 2:351–362.

Lavrenko, V., Manmatha, R., and Jeon, J. (2004). A model for learning the semantics of pictures. In Thrun, S., Saul, L. K., and Schölkopf, P. B., editors, *Advances in Neural Information Processing Systems, NIPS 2016*, pages 553–560. MIT Press.

Lebret, R., Pinheiro, P. H. O., and Collobert, R. (2015). Phrase-based image captioning. In *International Conference on Machine Learning (ICML)*, volume 37, page 2085–2094. JMLR.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2015). DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195.

Lemke, J. (1990). *Talking science: Language, learning and values.* Ablex Publishing Corporation.

Li, H. (2011). A short introduction to learning to rank. *IEICE Transactions*, 95-D(10):1854–1862.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., and Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 220–228, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lin, D., Fidler, S., Kong, C., and Urtasun, R. (2015). Generating multi-sentence natural language descriptions of indoor scenes. In *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, pages 93.1–93.13.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A. (2015). Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *IEEE International Conference on Computer Vision, ICCV 2015*, pages 2533–2541.

Mark J. Huiskes, B. T. and Lew, M. S. (2010). New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In *MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval*, pages 527–536, New York, NY, USA. ACM.

Mason, R. and Charniak, E. (2014). Nonparametric method for data-driven image captioning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 592–598.

Mei, Q., Shen, X., and Zhai, C. (2007). Automatic labeling of multinomial topic models. In *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining, ACM SIGKDD 2007, San Jose, California, USA, August 12-15, 2007*, pages 490–499.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*, pages 1–8.

Metzler, D. and Bruce Croft, W. (2007). Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274.

Milne, D. N. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, pages 509–518.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A., Berg, T., and Daume III, H. (2012). Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2012, Avignon, France, April 23-27, 2012*, pages 747–756.

Nastase, V. and Strube, M. (2012). Transforming Wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85.

Navigli, R. and Lapata, M. (2010). An experimental study on graph connectivity for unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692.

Navigli, R. and Ponzetto, S. P. (2012a). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif.ícial Intelligence*, 193:217–250.

Navigli, R. and Ponzetto, S. P. (2012b). BabelRelate! A joint multilingual approach to computing semantic relatedness. In *AAAI*, pages 108–114.

O'Neill, S. and Nicholson-Cole, S. (2009). Fear won't do it: promoting positive engagement with climate change through imagery and icons. *Science Communication*, 30(3):355–379.

O'Neill, S. and Smith, N. (2014). Climate change and visual imagery. *Wiley Interdisciplinary Reviews: Climate Change*, 5(1):73–87.

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 1143–1151.

Ortiz, L. G. M., Wolff, C., and Lapata, M. (2015). Learning to interpret and describe abstract scenes. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2015, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1505–1515.

Patterson, G., Xu, C., Su, H., and Hays, J. (2014). The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1):59–81.

Perlmutter, D. D. (1997). A picture's worth 8,500,000 people: News images as symbols of china. *Visual Communications Quaterly*, 4(2):4–7.

Perlmutter, D. D. (1998). *Photojournalism and Foreign Policy*. Praeger.

Perlmutter, D. D. and Wagner, G. L. (2004). The anatomy of a photojournalistic icon: Marginalization of dissent in the selection and framing of 'a death in Genoa'. *Visual Communication*, 3(1):91–108.

Ponzetto, S. P. and Navigli, R. (2010). Knowledge-rich Word Sense Disambiguation rivaling supervised system. In *ACL*, pages 1522–1531.

Ponzetto, S. P. and Strube, M. (2011). Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175:1737–1756.

Ponzetto, S. P., Wessler, H., Weiland, L., Kopf, S., Effelsberg, W., and Stuckenschmidt, H. (2015). Automatic classification of iconic images based on a multimodal model : an interdisciplinary project. In Wildfeuer, J., editor, *Sprache - Medien - Innovationen-Building bridges for multimodal research : international perspectives on theories and practices of multimodal analysis*, 7, pages 193–210, Frankfurt am Main ; Bern; Wien. Peter Lang Edition.

Raiber, F. and Kurland, O. (2013). Ranking document clusters using markov random fields. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland, July 28 - August 01, 2013*, pages 333–342.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G., Levy, R., and Vasconcelos, N. (2010). A New Approach to Cross-Modal Multimedia Retrieval. In *ACM International Conference on Multimedia*, pages 251–260.

Raviv, H., Kurland, O., and Carmel, D. (2016). Document retrieval using entity-based language models. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 65–74.

Ren, M., Kiros, R., and Zemel, R. S. (2015). Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2953–2961.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Sadeghi, M. A. and Farhadi, A. (2011). Recognition using visual phrases. *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1745–1752.

Schuhmacher, M. and Ponzetto, S. P. (2014). Knowledge-based graph document modeling. In *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*, pages 543–552.

Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., and Manning, C. D. (2015). Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language, VL@EMLP 2015, Lisbon, Portugal, September 18, 2015*, pages 70–80, Lisbon, Portugal. Association for Computational Linguistics.

Shutova, E., Kiela, D., and Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016, San Diego California, USA, June 12-17, 2016*, pages 160–170.

Shutova, E., Teufel, S., and Korhonen, A. (2013). Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Socher, R. and Fei-Fei, L. (2010). Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 966–973.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2:207–218.

Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 2222–2230.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW 2007, pages 697–706.

Talukdar, P. P., Reisinger, J., Pasca, M., Ravichandran, D., Bhagat, R., and Pereira, F. C. N. (2008). Weakly-supervised acquisition of labeled class instances using graph random walks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 582–590.

Thomee, B. and Popescu, A. (2012). Overview of the imageclef 2012 flickr photo annotation and retrieval task. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*.

Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., and Sienkiewicz, C. (2016). Rich image captioning in the wild. *arXiv preprint arXiv:1603.09016*.

Ushiku, Y., Yamaguchi, M., Mukuta, Y., and Harada, T. (2015). Common subspace for model and similarity: Phrase learning for caption generation from images. In *IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2668–2676.

Verma, Y. and Jawahar, C. V. (2014). Im2text and text2im: Associating images and texts for cross-modal retrieval. In *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014*.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164.

Wang, C., Yang, H., and Meinel, C. (2016). A deep semantic framework for multimodal representation learning. *Multimedia Tools and Applications*, 75(15):9255–9276.

Weegar, R., Hammarlund, L., Tegen, A., Oskarsson, M., Åström, K., and Nugues, P. (2014). Visual entity linking: A preliminary study. In *AAAI-14 Workshop on Computing for Augmented Human Intelligence*.

Weiland, L., Dietz, L., and Ponzetto, S. P. (2015). Image with a message: Towards detecting non-literal image usages by visual linking. In *Proceedings of the 2015 EMNLP Workshop on Vision and Language, (VL'15)*, pages 40–47.

Weiland, L., Effelsberg, W., and Ponzetto, S. P. (2014). Weakly supervised construction of a repository of iconic images. In *In Proceedings of the Workshop on Vision and Language 2014, (VL'14) at the 25th International Conference on Computational Linguistics (COLING '14)*.

Weiland, L., Hulpuş, I., and Ponzetto, Simone Paolo Dietz, L. (2017). Using object detection, nlp, and knowledge base to understand the message of images. In *MultiMedia Modeling - 23nd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017*, pages 405–418. Springer International Publishing.

Weiland, L., Hulpuş, I., Ponzetto, S. P., and Dietz, L. (2016). Understanding the message of images with knowledge base traversals. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, September 12- 6, 2016*, pages 199–208.

Weiland, L., Hulpuş, I., Ponzetto, S. P., Effelsberg, W., and Dietz, L. (2018a). Knowledge-rich image gist understanding beyond literal meaning. *Data & Knowledge Engineering*.

Weiland, L., Ponzetto, S. P., Effelsberg, W., and Dietz, L. (2018b). Understanding the gist - ranking of concepts for multimedia indexing. In *arXiv preprint, arXiv:1809.08593*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. cite arxiv:1502.03044.

Yagcioglu, S., Erdem, E., Erdem, A., and Çakıcı Ruket (2015). A distributed representation based query expansion approach for image captioning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 106–111. ACL.

Yang, Y., Teo, C. L., III, H. D., and Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 444–454.

Yatskar, M., Galley, M., Vanderwende, L., and Zettlemoyer, L. (2014). See no evil, say no evil: Description generation from densely labeled images. *Lexical and Computational Semantics (* SEM 2014)*, page 110.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214.

Zhang, Z., Gentile, A. L., and Ciravegna, F. (2013). Recent advances in methods of lexical semantic relatedness - a survey. *Natural Language Engineering*, 19(4):411–479.

Zhiltsov, N., Kotov, A., and Nikolaev, F. (2015). Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 253–262.

Zitnick, C. L. and Parikh, D. (2013). Bringing semantics into focus using visual abstraction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, Portland, OR, USA, June 23-28, 2013*, pages 3009–3016.