Image with a Message:
Towards detecting non-literal image usages by visual linking

Lydia Weiland, Laura Dietz and **Simone Paolo Ponzetto**
Data and Web Science Group
University of Mannheim
68131 Mannheim, Germany
{lydia, dietz, simone}@informatik.uni-mannheim.de

Abstract

A key task to understand an image and its corresponding caption is not only to find out what is shown on the picture and described in the text, but also what is the exact relationship between these two elements. The long-term objective of our work is to be able to distinguish different types of relationship, including literal vs. non-literal usages, as well as fine-grained non-literal usages (i.e., symbolic vs. iconic). Here, we approach this challenging problem by answering the question: 'How can we quantify the degrees of similarity between the literal meanings expressed within images and their captions?'. We formulate this problem as a ranking task, where links between entities and potential regions are created and ranked for relevance. Using a Ranking SVM allows us to leverage from the preference ordering of the links, which help us in the similarity calculation for the cases of visual or textual ambiguity, as well as misclassified data. Our experiments show that aggregating different features using a supervised ranker achieves better results than a baseline knowledge-base method. However, much work still lies ahead, and we accordingly conclude the paper with a detailed discussion of a short- and long-term outlook on how to push our work on relationship classification one step further.

1 Introduction

Despite recent major advances in vision and language understanding, the classification of usage relationships between images and textual captions is still an open challenge, which is still to be addressed from a computational point of view. Relationships between images and texts can be classified from a general perspective into three different types, namely literal, non-literal and no-relationship. Literal relations cover captions and/or longer corresponding texts that have a descriptive character with respect to the associated image. Non-literal refers instead to images and captions having a relationship that arouses broad associations to other topics, e.g., abstract topics.

The class of non-literal relationships itself can be further divided: *Symbolic photographs* are a common example of non-literal relations. Pictures of this kind can be used without any further explanation on the basis of common socially-mediated understanding, e.g., a heart as a symbol of love, an apple and the snake as an symbol of original sin, or the peace symbol. *Social media* typically use another type of language and sometimes can only be understood by insiders or people, who attended to the situation the photo has been taken, e.g., "Kellogs in a pizza box", with a photo showing a cat sleeping in a pizza box. Without the image, it would have been only clear to those who know Kellogs that a cat is meant by this caption. To the ordinary reader, this would rather suggest a typo and thus, cereals in the pizza box. Those types of relationships can often be found on Flickr, e.g., in the SBU 1M dataset (Ordonez et al., 2011).

A third category is the one of *Media icons* (Perlmutter and Wagner, 2004; Drechsel, 2010), which is typically focused on hot, sensitive, and abstract topics, which are hard to depict directly. Pictures of this kind are often used by news agencies, politicians, and organizations, e.g., a polar bear on an ice floe for global warming. This type of non-literal relationship uses a combination of descriptive parts and language beyond a literal meaning, which assumes fine-grained domain and background knowledge, e.g., the ice floe melting as a result of global warming. When knowledge of this kind is not readily available to the reader, it can be

Proceedings of the 2015 Workshop on Vision and Language (VL'15), pages 40–47,
Lisbon, Portugal, 18 September 2015. ©2015 Association for Computational Linguistics.

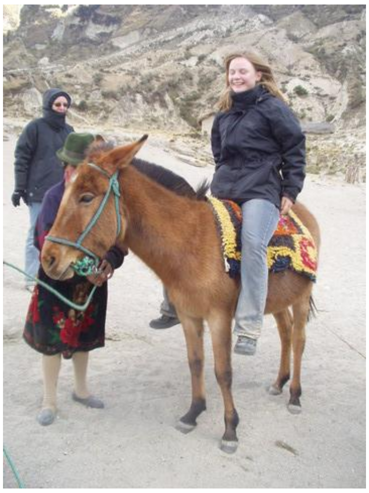
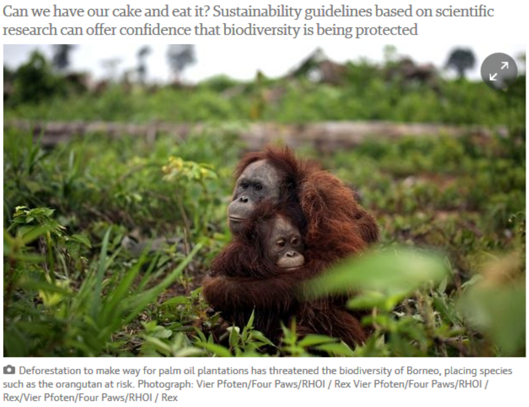


Figure 1: Caption: ”A girl with a black jacket and a blue jeans is sitting on a brown donkey; another person is standing behind it; a brown, bald slope in the background.”



Can we have our cake and eat it? Sustainability guidelines based on scientific research can offer confidence that biodiversity is being protected

Deforestation to make way for palm oil plantations has threatened the biodiversity of Borneo, placing species such as the orangutan at risk. Photograph: Vier Pfoten/Four Paws/RHOI / Rex Vier Pfoten/Four Paws/RHOI / Rex/Vier Pfoten/Four Paws/RHOI / Rex

Figure 2: Non-literal caption: “Deforestation to make way for palm oil plantations has threatened the biodiversity of Borneo, placing species such as the orangutan at risk.”. Literal caption: “Two orangutans hugging each other on a field with green leaves. A wooden trunk lays in the background.”. Photograph: BOSF I VIER PFOTEN

still acquired by reading associated articles or, in general, by getting to know further facts about a topic. This way the readers are able to create the association of the topic to the image-caption pair.

In our work, we aim at developing methods for automatic understanding of relations between natural language text and pictures *beyond literal meanings and usages*. In particular, we ultimately aim to automatically understand the cultural semantics of iconic pictures in textual contexts (i.e., captions, associated texts, etc.). Far from being an abstract research topic, our work has the potential to impact real-world applications like mixed image-text search (Panchenko et al., 2013), especially in cases of ambiguous or abstract topics in textual queries. Even if current state-of-the-art search engines perform very well, not every search query is answered with what a user expects, e.g., in cases of ambiguity or image and text pairs with non-literal meaning. Being able to assess if a caption and an image are in literal, non-literal, or no relationship can have positive effects to search results. Another, more specific use case is the training of image detectors with the use of captions, which are available in large amounts on the World Wide Web. Training image detectors requires image-caption pairs of the literal class, so being able to reliably identify such instances will arguably produce better, more reliable, and precise object or scene detection models. This is particularly of interest in the news and social media domain, where customizing image detectors for trending entities is of high interest.

Most of the datasets used for training and testing methods from natural language processing, computer vision, or both, are focusing on images with literal textual description. When humans are asked to annotate images with a description, they tend to use a literal caption (cf., e.g., Figure 1). However, captions in real world news articles are devised to enhance the message and build bridges to a more abstract topic, thus have a non-literal or iconic meaning – cf., e.g., the caption of Figure 2 on deforestation in combination with an image showing the orangutan mother with her baby in an open field without trees. Note that image-captions of this kind are typically designed to arouse an emotive response in the reader: in this case, the non-literal usage aims at leading the reader to focus on an abstract topic such as the negative impacts of palm oil plantations. In contrast, the literal caption for this image would rather be “Two orangutans hugging each other on a field with green leaves. A wooden trunk lays in the background.” The literal image-caption pair, without further background knowledge, does not trigger this association.

Existing methods from Natural Language Processing (NLP), Computer Vision (CV) do not, and are not meant to find a difference between the same images being used in another context or the same textual contexts depicted with other viewpoints of an abstract topic. In the case of image detection there is no difference between the image with the literal or non-literal caption – it is still the same image, classified as e.g., orangutans. Only when the caption is incorporated into the prediction process, we are able to identify the image-caption pair into the appropriate usage classes, either in a coarse-grained (i.e., 'literal' versus 'non-literal') or fine-grained (e.g., 'media icons' versus 'symbolic photographs').

Spinning our example further, if we would replace the image of Figure 2 with a picture showing a supermarket shelf with famous palm-oil-rich products, it should still be classified as non-literal. However, when regarding the caption as arbitrary text without the context of a picture, this does not have any iconic meaning. Likewise, image processing without considering text cannot predict the relationship to this abstract topic. Therefore, the classification into 'literal' or 'non-literal' (respectively 'media iconic') needs to integrate NLP and CV together. Our working assumption is that the iconic meaning reveals itself through the mismatches between objects mentioned in the caption and objects present in the image.

In this paper we set to find methods and measures to being able to classify these different image-text usage relationships. Consequently, we aim at answering the following research questions:

- What constitutes a literal class of image-caption pair?
- Which method or measure is required to classify a pair as being literal?
- Are we able to derive methods and measures to approach the detection of non-literal pairs?
- How to differentiate literal, non-literal, and not-related classes from each other?

As a first step towards answering these questions, we focus here on detecting literal text-image usages. Therefore, we focus on a dataset of images and captions with literal usages. Our hunch is that *the more links between entities from the caption and regions in the image we can create, the more literal the relationship becomes*. In order to verify this hypothesis, we need to create links between entities from the text and regions with an object in the image, a problem we next turn to.

2 Methods

We provide a first study of the problem of visual entity linking on the basis of a machine learning approach. To the best of our knowledge, Weegar et al. (2014) is the only previous work to address the problem of automatically creating links between image segments and entities from the corresponding caption text. For their work, they use the segmented and annotated extension of the IAPR-TC12 dataset (Grubinger et al., 2006), which consists of segmented and textual annotated images and corresponding captions – we refer to this dataset as SAIAPR-TC12 (Escalante et al., 2010) in the following. In contrast to their work we aim at exploring the benefits of a supervised learning approach for the task at hand: this is because, in line with many other tasks in NLP and CV, we expect a learning framework such as the one provided by a Ranking SVM to effectively leverage labeled data, while coping with ambiguity within the images and associated text captions.

2.1 Ranking SVM

Given a tuple (Q, S, M), with Q as a query, S the ranked segments of an image, and M defined based on the different methods to generate and extract features. Then the score $H_\theta(Q, S)$ between a query Q and a segment S, can be obtained by maximizing over M (Lan et al., 2012; Joachims, 2002): $H_\theta(Q, S) = \arg\max_M F_\theta(Q, S, M)$, where θ is the feature vector consisting of at least one feature or a combination of features. We now proceed to describe such features in details.

2.2 Ranking SVM with Textual Features

GloVe-based cosine similarity: We use the distributional vectors from Pennington et al. (2014) to provide us with a semantic representation of the captions. For each noun of the caption, the GloVe vector calculated on a pre-trained model (Wikipedia 2014, 300d) is used to calculate semantic similarity as:

$$\sum_{q_i \in q \setminus q_{color} \cap l} \alpha(f(q), f(l))$$

where $q \setminus q_{color}$ refers to queries without color entities. l is defined with $l \in I_j$, where l denotes the label of the segment of the current image (I_j). $f(q)$ and $f(l)$ is defined as the feature vector from GloVe and α is defined as the cosine similarity function between those vectors.

GloVe-based cosine similarity with color entities: Besides nouns, GloVe is also able to implicitly associate colors to words, allowing us to determine that, e.g., the color name 'green' and the noun 'meadow' have a high similarity. The SAIAPR-TC12 dataset has more descriptive captions, where a lot of color names are used to describe how the objects and scenes look like. Besides, the text-based label catalog uses color names to further specify a subcategory of a diverse hypernym, e.g., 'sky' can be 'blue', 'light', 'night' and, 'red-sunset-desk'. We accordingly extend the GloVe feature as:

$$\sum_{q_i \in q \cap l} \alpha(f(q), f(l))$$

where q consists of all possible queries, including the color entities.

In the text-only setting the ranking SVM uses only the textual description of the labels and no visual features. The ranking SVM features thus consist of cosine similarities between segment labels and a query consisting of entities and color names. The result thus consists of a ranking of potential segment labels.

2.3 Ranking SVM with Visual Features

HOG: Since images usually do not come with manual segmented and textual annotated regions, we include visual features to systematically substitute textual and manually set information in the images. Thus, we make use of image features as an alternative to the text-based label catalog.

Histogram of Oriented Gradients (HOG): In this stage we still leverage from the image segments, but instead of using the textual label, we apply a classification to every segment. Based on the label statistics from our dataset, models are trained using a binary SVM. For each label, we collect data from ImageNet (Deng et al., 2009), where bounding box information for some objects are provided. With the images from ImageNet, SVM classifiers based on Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005) are trained[1]. After training, bounding boxes around every segment are defined. From the normalized version of bounding boxes, HOG features are extracted. These features are then used to classify the test data within every of the trained models. The resulting predictions are stored and serve as features for the Ranking SVM. Thus, our HOG-based features are defined as:

$$\sum_{q_i \in q \setminus q_{color} \cap s} \beta_i^T f(S)$$

Where β_i is the prediction of a linear SVM of detecting object i and $f(S)$ denotes the HOG feature vector of segment S.

HOG and Color Names: Based on Ionescu et al. (2014), we use eleven different color names, which are extracted from the captions of the texts from our dataset. For every normalized bounding box of the segments from the training dataset, color histograms are calculated. The bins of the color histograms serves as a feature vector for the color Ranking SVM. The colors of the bounding boxes are ranked with respect to the context of the color in the caption:

$$\sum_{q_i \in q \setminus q_{entities} \cap s} \gamma_i^T f(S)$$

The queries are now color names without object entities, $f(S)$ defines the distribution of a color defined in γ. We assume entities, which are further described with a color name in the caption, as multi-word queries. The predictions from both rankings are summed to build the final ranking.

3 Evaluation

3.1 Dataset

We conduct experiments on the SAIAPR-TC12 dataset (Escalante et al., 2010). Whilst the Flickr30k dataset (Plummer et al., 2015) is 1.5 the size of the SAIAPR-TC12, it lacks accurate segmentations, which might be relevant for image processing. The IAPR-TC12 consists of 20,000 images with a caption each. The images are covering topics of interesting places, landscapes, animals, sports, and similar topics, which can typically be found in image collections taken from tourists on their holidays. A caption consists of one to four sentences (23.06 words per caption on average (Grubinger et al., 2006)). In addition, the extension delivers segmentation masks of each image, where an image can have multiple segmentation (4.97 segments per image on average (Es-

[1] Note that for our purposes we cannot use existing models, like Pascal VOC (Everingham et al., 2010), for instance, because it has only a small overlap in the set of objects in our data.

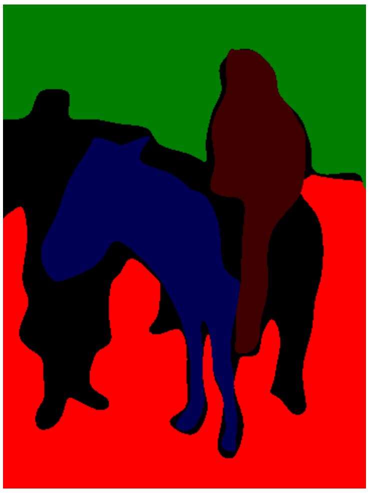


Figure 3: Figure 1 with segmentation masks. The segments are labeled with: mammal-other, mountain, woman, sand-desert

calante et al., 2010)). Each segmentation has exactly one label from a predefined catalog of 276 annotations created by a concept hierarchy. Furthermore, spatial relations (adjacent, disjoint, beside, X-aligned, above, below and Y-aligned) of the segmentation masks are defined and image features are given, with respect to the segments (area, boundary/area, width and height of the region, average and standard deviation in x and y, convexity, average, standard deviation and skewness in both color spaces RGB and CIE-Lab).

An example image of the SAIPR-TC12 with segmentation masks and the affiliated caption is given in Figure 1 and 3. The example also shows that due to the limited amount of labels objects are not inevitably represented by the same word in images and captions. Links between entities of the captions and corresponding image segments are not given by default. Due to the topics, covered by the dataset, which are similar to other datasets, the SAIAPR-TC12 can be used as training data. Whereas other, non segmented datasets can be used as testing data, e.g., MediaEval Benchmarking (Ionescu et al., 2014).

3.2 Baseline

We build upon previous work from Weegar et al. (2014) and develop a text-based baseline for our task. To this end, we selected 39 images with 240 segments (from 69 different objects) and corresponding captions with 283 entities (133 different entities), with an average of 6.15 and 7.26, respectively. An overview of object representations in the amount of labels and entities, and their distribution within the test data is given by Table 1.

Entity	Amount	Label	Amount
Sky	12	Leaf	16
Mountain	9	Rock	14
Rock	8	Sky (Blue)	13
Tree	7	Plant	13
House	7	Man	11
Wall	6	Woman	11
People	6	Mountain	10
Building	5	Ground	9
Woman	4	Grass	8
Water	4	Vegetation	8

Table 1: Most common 10 labels and entities of test data selection.

From each of the 39 images we use the textual image segment labels (in the latter referred to as label) and the captions. With Stanford CoreNLP (Manning et al., 2014) we extract the nouns (NN/NNS) from the captions (in the latter referred to as 'entity'). If a noun is mentioned in plural form (NNS), we use the lemma instead (e.g., horses is stored as horse). The extracted entities and labels are stored and further processed image-wise, so that only links between an image segment and an entity from the corresponding caption can be created.

With WordNet and the similarity measure according to WUP (Wu and Palmer, 1994), we calculated the similarity between every label and every entity. A link is stored between the most similar label and entity. Whereas we allow to link multiple segments to one entity. This is done to be able to link multiple instances of one object in an image to the lemmatized entity. To simplify the method with respect to any ambiguity, we used the most frequent sense in WordNet. Overall, the method results in precision of 0.538 and F1 measure of 0.493, thus providing us with a baseline approach with results comparable to the original ones from Weegar et al..

3.3 Experimental Settings and Results

We manually created links between the 240 segments and 231 entities of the originally 281 extracted ones. Since some entities are abstract words, describing images, e.g. 'background',

Different Ranking SVM	Precision	Recall	F1-Measure
Baseline	0.5375	0.45583	0.4933
Cosine Similarity of GloVe	0.7853	0.9392	0.7473
Cosine Similarity of GloVe (Color Entities included)	0.6848	0.9003	0.6551
HOG	0.5459	0.5322	0.3512
HOG and CN	0.6379	0.5796	0.4059

Table 2: Results of the baseline and the different ranking SVM with the two metrics for relevance (Precision), diversity (Recall), and mean of relevance and diversity (F1-Measure).

those entities are filtered in advance (already in the baseline). Overall, 98 color names, that are further describing entities, can be extracted. All links are rated with respect to the query. Within a leave-one-out approach we cross validated every method. As color features are low level features, and rather supposed to enrich the HOG model, it is not separately evaluated. All Ranking SVM results are evaluated for Precision (P), Recall (R) and F1-Measure (F1).

The text-based Baseline achieves precision and F1 with around 50% (cf Table 2). The also text-based Cosine Similarity of GloVe achieves around one and a half better results than the baseline, but these results are reduced for around 10% after integrating the cosine similarities of color names and labels. Vice versa, the two visual feature approaches show better results when integrating both feature types – HOG and color (P: 63.79% vs. 54.59%, F1:40.59% vs. 35.12%).

The results indicate, that visual feature selection and extraction needs further improvement, but they also show, that a post-processing, e.g., reranking with aggregation can have positive impacts.

4 Related Work

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. 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 (Kulkarni et al., 2011; Yang et al., 2011; Gupta et al., 2012), and videos (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 (2010) and Das et al. (2013a) being notable exceptions. However, even these latter contributions address the easier task of generating visual descriptions for standard, news text. But while processing newswire text is of great importance, this completely disregards other commonly used, yet extremely challenging, dimensions of natural language like metaphorical and figurative language usages in general, which are the kinds of contexts we are primarily interested in. The ubiquity of metaphors and iconic images, in particular, did not inspire much work in Computer Science yet: researchers in NLP, in fact, only recently started to look at the problem of automatically detecting metaphors (Shutova et al., 2013), whereas research in computer vision and multimedia processing did not tackle the problem of iconic images at all.

To the best of our knowledge there is only one related work about the link creation between image segments and entities from the corresponding caption text, namely the study from Weegar et al. (2014), who use the segmented and annotated extension (Escalante et al., 2010) of the IAPR-TC12 dataset (Grubinger et al., 2006), which consists of segmented and textual annotated images and corresponding captions. Due to the textual annotated images, Weegar et al. are able to follow a text-only approach for the linking problem. They propose a method which is based on word similarity using WordNet, between extracted nouns (entities) from the caption and the textual annotation labels of the image segments. For evaluation purposes, they manually created links in 40 images from the dataset with 301 segments and 389 entities. The method results in a precision of 55.48% and serves as an inspiration for the baseline used to compare our own method.

In Plummer et al. (2015) annotators were asked to annotate only objects with bounding boxes that were mentioned in the caption. Not every object in images is asked for a bounding box and an annotation, but those which are mentioned in the captions. Within experiments (bidirectional image-sentence retrieval and text-to-image co reference), they showed the usefulness of links between images and captions, but they also pointed out the issue we are addressing here: Leveraging the links is dependent on a high accuracy between the regions of an image and the textual phrases.

Hodosh at al. (2015) formulates the image description task as ranking problem. Within their method five different captions for one image are ranked. Their results show that metrics using ranked lists, and not only one query result, are more robust.

Dodge et al. (2012) developed methods to classify noun phrases into visual or non-visual text. Visual means things that can be seen on an image. Their results indicate, that using visual features improves the classification. Overall, the classification of visual and non-visual text is especially interesting for the classification of literal and non-literal pairings.

5 Conclusions and Future work

In this work we developed a supervised ranking approach to visual linking. Ranking links between entities and segments is inspired by several aspects of creating the links between caption entities and segments. First, there might be several segments which perfectly fit to one mention in the caption. Second, as object detection approaches are far from being robust and perfect, it might be helpful to limit ourselves not to one decision (binary) but rather to use a ranking, where correct object class might be on lower rank but still to considered. Third, if an object is not covered within a pre-trained model, these objects either will not be considered in the detection and evaluation or wrongly classified.

Visual linking provides us with a first attempt in the direction of solving the question of whether caption is the literal description of the image it is associated with. That is, our goal is not to find an object detector with the highest precision (e.g., answering the question “Is there an orangutan or a chimpanzee on the image?”), but rather if and how much related the images and the captions are to each other. If the caption is talking about palm-oil harvesting and the image shows an orangutan to depict the endangered species, we are interested in receiving detector results with a high probability for an animal as such, and being able to create the non-literal link between these two topics.

In the short term, a necessary step is to develop a model that does not rely on manually defined enrichments of the dataset (e.g., textual labels or segmentation masks). We will accordingly look at ways to perform predictions about regions of interest from the linear SVM and work without the bounding boxes from the dataset. To this end, our dataset needs to be extended, so that we can apply our improved methods also on non-literal image-caption pairings.

In the long term, we need to directly investigate the hypothesis of whether the more links between entities from the caption and regions in the image can be created, the more literally the relationship becomes. That is, a hypothesis for non-literal relationships needs to be computationally formulated and also investigated. Besides this, it would be interesting to discover interesting discriminative characteristics between literal and non-literal images. Finally, future work will concentrate on the differentiation of cultural influences in the interpretation of non-literal image-caption pairs, for instance by taking the background of coders into account (e.g., on the basis of a crowdsourced-generated dataset).

Acknowledgments

The authors would like to thank Vier Pfoten (`www.vier-pfoten.de`) for the permission to use the orangutan photograph in this publication. This work is funded by the Research Seed Capital (RiSC) programme of the Ministry of Science, Research and the Arts Baden-Württemberg, and used computational resources offered from the bwUniCluster within the framework program bwHPC.

References

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893.

Pradipto Das, Rohini K. Srihari, and Jason J. Corso. 2013a. Translating related words to videos and back through latent topics. In *Proc. of WSDM-13*, pages 485–494.

Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. 2013b. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proc. of CVPR-13*, pages 2634–2641.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR09*.

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé, III, Alexander C. Berg, and Tamara L. Berg. 2012. Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 762–772.

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

Hugo Jair Escalante, Carlos A. Hernández, Jess A. González, Aurelio López-López, Manuel Montes y Gómez, Eduardo F. Morales, Luis Enrique Sucar, Luis Villaseñor Pineda, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428.

Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vision*, 88(2):303–338.

Yansong Feng and Mirella Lapata. 2010. Topic models for image annotation and text illustration. In *Proc. of NAACL-HLT-10*, pages 831–839.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 benchmark – a new evaluation resource for visual information systems.

Ankush Gupta, Yashaswi Verma, and C. V. Jawahar. 2012. Choosing linguistics over vision to describe images. In *Proc. of AAAI-12*, pages 606–612.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2015. Framing image description as a ranking task: Data, models and evaluation metrics (extended abstract). In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 4188–4192.

Bogdan Ionescu, Anca-Livia Radu, María Menéndez, Henning Müller, Adrian Popescu, and Babak Loni. 2014. Div400: A social image retrieval result diversification dataset. In *MMSys '14*, pages 29–34.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *KDD '02*, pages 133–142.

Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond Mooney, Kate Saenko, and Sergio Guadarrama. 2013. Generating natural-language video descriptions using text-mined knowledge. In *Proc. of AAAI-13*.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating image descriptions. In *In Proc. of CVPR-11*, pages 1601–1608.

Tian Lan, Weilong Yang, Yang Wang, and Greg Mori. 2012. Image retrieval with structured object queries using latent ranking svm. In *Proc. of ECCV'12*, pages 129–142.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. of ACL-14*, pages 55–60.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.

Alexander Panchenko, Pavel Romanov, Olga Morozova, Hubert Naets, Andrey Philippovich, Alexey Romanov, and Cédrick Fairon. 2013. Serelex: Search and visualization of semantically related words. In *Proc. of ECIR-13*, pages 837–840.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP-14*, pages 1532–1543.

David D. Perlmutter and Gretchen L. Wagner. 2004. The anatomy of a photojournalistic icon: Marginalization of dissent in the selection and framing of 'a death in genoa'. *Visual Communication*, 3(1), February.

Bryan Plummer, Liwei Wang, Chris Cervantes, Juan Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Rebecka Weegar, Linus Hammarlund, Agnes Tegen, Magnus Oskarsson, Kalle Åström, and Pierre Nugues. 2014. Visual entity linking: A preliminary study. In *Proc. of the AAAI-14 Workshop on Computing for Augmented Human Intelligence*.

Zhibiao Wu and Martha Stone Palmer. 1994. Verb semantics and lexical selection. In *Proc. of ACL-94*, pages 133–138.

Yezhou Yang, Ching Lik Teo, Hal Daumé, III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proc. of EMNLP-11*, pages 444–454.