Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures (Extended Abstract)*

Raffaella Bernardi[1], Ruket Cakici[2], Desmond Elliott[3], Aykut Erdem[4], Erkut Erdem[4], Nazli Ikizler-Cinbis[4], Frank Keller[5], Adrian Muscat[6], Barbara Plank[7]

[1]University of Trento
[2]Middle East Technical University
[3]University of Amsterdam
[4]Hacettepe University
[5]University of Edinburgh
[6]University of Malta
[7]University of Groningen

raffaella.bernardi@unitn.it, ruken@ceng.metu.edu.tr, {aykut, erkut, nazli}@cs.hacettepe.edu.tr, adrian.muscat@um.edu.mt, d.elliott@uva.nl, keller@inf.ed.ac.uk, b.plank@rug.nl

Abstract

Automatic image description generation is a challenging problem that has recently received a large amount of interest from the computer vision and natural language processing communities. In this survey, we classify the known approaches based on how they conceptualise this problem and provide a review of existing models, highlighting their advantages and disadvantages. Moreover, we give an overview of the benchmark image-text datasets and the evaluation measures that have been developed to assess the quality of machine-generated descriptions. Finally we explore future directions in the area of automatic image description.

1 Introduction

The task of *automatic image description* involves taking an image, analyzing its visual content, and generating a textual description (typically a sentence) that verbalizes the most salient aspects of the image. This requires the joint use of both Computer Vision (CV) and Natural Language Processing (NLP) techniques.

From a CV point of view, the description could in principle cover any visual aspect of the image: it can talk about objects and their attributes, features of the scene (e.g., indoor/outdoor), or verbalize the interaction of the people and objects in the scene. More challenging, it can reference objects that are not depicted (e.g., people waiting for a train, even when the train is not visible because it has not arrived yet) and provide background knowledge that cannot be derived directly from the image (e.g., the person depicted is the Mona Lisa). In other words, image understanding (which essentially produces an unstructured list of object, scene and interaction labels) is necessary, but clearly not sufficient for producing a good description. A good description should be comprehensive but concise (talk about all and only the important things in the image), while being formally correct, i.e., consist of grammatically well-formed sentences. In this review, we follow Hodosh *et al.* [2013] and assume that the descriptions that are of interest for this survey are the ones that verbalize visual and conceptual information in an image.

The NLP task of natural language generation (NLG) takes a non-linguistic representation, in our case an image representation (e.g., a list of objects and their spatial relationships) and turns it into human-readable text, e.g., a sentence in a natural language. Generating text involves a series of steps: we need to decide which aspects of the input to talk about (content selection), then we need to organize the content (text planning) and verbalize it (surface realization). Surface realization in turn requires choosing the right words (lexicalization), using pronouns whenever appropriate (referential expression generation), and grouping of related information (aggregation).

In summary, automatic image description requires both image understanding and natural language generation, thus bridging the CV and the NLP communities. An extensive relevant literature has appeared over the last five years. The aim of this survey is to give a comprehensive overview of this literature, covering state-of-the-art models, datasets, and evaluation metrics that have been developed or adopted in this area of research.

In this extended abstract, we first group automatic image description models into three categories: those that generate the description from scratch, those that search a visual space and those that search a multimodal space. In Section 3, we examine the available multimodal image datasets used for the training and testing of the models and in Section 4 we review evaluation measures that have been used to gauge the quality of generated descriptions. We then discuss future research directions in Section 5.

*This paper is an extended abstract of an article in the Journal of Artificial Intelligence Research [Bernardi *et al.*, 2016].

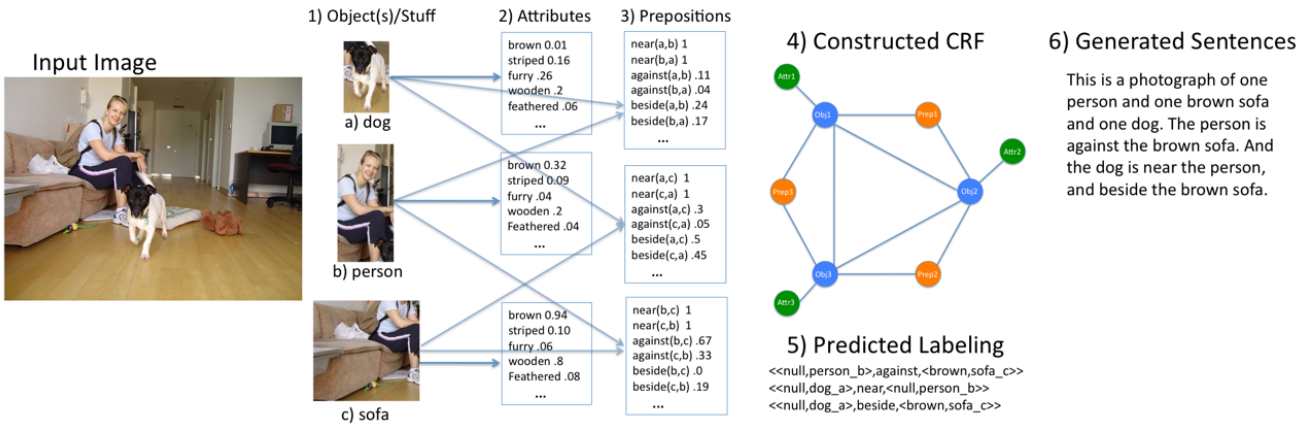

Figure 1: The automatic image description generation system proposed in [Kulkarni *et al.*, 2011].

2 Image Description Models

In this review, we organise the models into three categories. The first group follows the classical NLG pipeline, i.e., the image description is generated starting from a detailed list of image content. The fact that it is difficult to evaluate machine-generated sentences has led to the second group of models that cast the problem as a retrieval problem. The sophisticated NLG requirement is relaxed by transferring human-authored descriptions (directly or synthesised) from a set of similar images, which are retrieved from a database of images–description pairs. The second group is further subdivided according to whether the retrieval is carried out in visual space or multimodal space.

2.1 Description as Generation from Visual Input

Models in this category follow a standard pipeline. CV techniques are first applied to recognize the scene, objects, spatial relationships, and actions present in the image. The words or phrases obtained in the first step are then combined to produce a natural language description, using techniques from NLG (e.g., templates, n-grams, grammar rules). An illustration of an early example of this type of model is given in Figure 1. These models differ along two main dimensions: (a) which image representations they derive descriptions from, and (b) how they address the sentence generation problem.

In terms of the image representations used we differentiate between those systems that rely on a list of image attributes, such as objects and their relationships, often expressed as tuples or triples and those that make explicit use of structure in an image. The early work of Kulkarni *et al.* [2011] and Mitchell *et al.* [2012] are examples of the former. Relationships are also augmented using models that make use of a linguistic corpus in addition to CV techniques. For example, the work of Fang *et al.* [2015] trains detectors from images and their associated descriptions using a weakly supervised approach. The idea of explicitly representing image structure was first explored by Elliott and Keller [2013], who developed a Visual Dependency Representations (VDR) framework to capture the spatial relations between the objects in an image in the form of a dependency graph. We review further work that makes use of scene graphs in the next paragraph.

In terms of addressing the sentence generation problem we differentiate between systems that are based on n-gram language models, sentence templates and more sophisticated NLG pipeline methods. Early examples based on n-gram based language models, which capture the probability of generating a word given the words that precede it, include the work by Kulkarni *et al.* [2011]. More recent work, for example Kiros *et al.* [2015], use recurrent neural networks (RNNs), which can be seen as advanced language models. The RNN is trained to generate the next word from both the previous words and image features, and is therefore a novel joint language-visual model. On the other hand sentence templates are pre-defined sentence frames (often manually generated) in which the missing words are filled with objects labels, relations, or attributes. For instance, the model [Elliott and Keller, 2013] traverses a VDR to fill in the sentence templates, and selects content by learning associations between VDRs and syntactic dependency trees. More linguistically sophisticated approaches have been applied to the generation of sentences. For example, Mitchell *et al.* [2012] uses a tree-substitution grammar to recombine syntactically well-formed generated sentence fragments, and Ortiz *et al.* [2015] model image description as machine translation over VDR–sentence pairs and perform explicit content selection and surface realization.

While the general pipeline architecture described above does not require a large data set for training, it does constrain the generated descriptions to a predefined set of semantic classes of scenes, objects, attributes, and actions and assumes that the detectors for each semantic class are accurate, an assumption that is not always met in practice.

2.2 Description as a Retrieval in Visual Space

The studies in this group cast the problem of automatically generating the description of an image by retrieving images similar to a query (i.e., the new image to be described); this is illustrated in Figure 2. In other words, these systems exploit similarity in the visual space to transfer descriptions to the query images.

Models based on visual retrieval typically follow a two-step pipeline. Based on the chosen visual feature space and

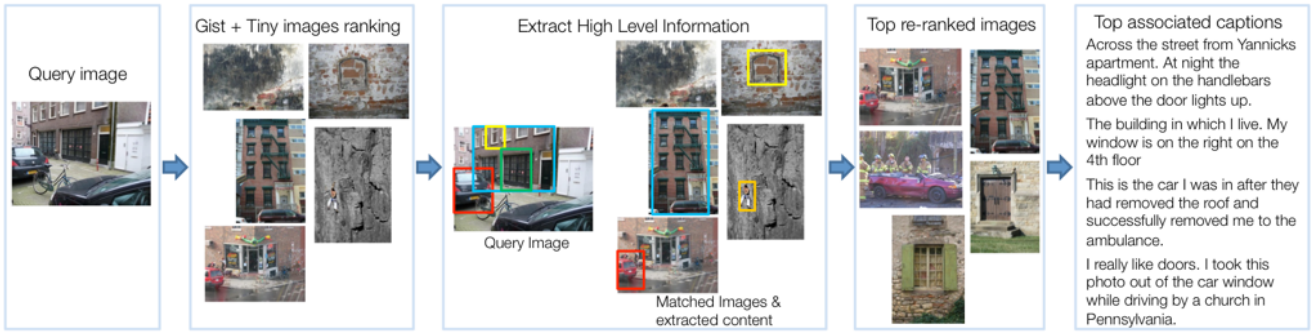

Figure 2: The description model based on retrieval from visual space proposed by [Ordonez *et al.*, 2011].

the similarity function, a set of candidate images are retrieved from the training set. The descriptions of the candidate images are re-ranked by further making use of visual and/or textual information contained in the retrieval set, or alternatively by combining fragments of the candidate descriptions according to certain rules. In most studies, the result of the first step is considered as a baseline on which the re-ranking step improves. Retrieval-based systems therefore differ mainly in how low-level and high-level visual features are used in the retrieval and re-ranking steps and in how to select (or compose) the final description.

The IM2TEXT model [Ordonez *et al.*, 2011] (see Figure 2), makes use of Gist and Tiny Image descriptors in the first retrieval step, while using a range of detectors (e.g., objects, stuff, pedestrians) and scene classifiers specific to the entities mentioned in the candidate descriptions. The re-ranking is carried out via a classifier trained over these semantic features. On the other hand, the re-ranking step of Mason and Charniak [2014] considers only textual information and the final output description is determined by using extractive summarization techniques. Later studies make use of convolutional neural networks (CNNs) to compute the image features [Devlin *et al.*, 2015]. Phrase-based approaches to synthesise the output were first reported in Kuznetsova *et al.* [2012]. Similar detectors and classifiers used in the re-ranking step of the IM2TEXT model are applied on a query image to extract and represent its semantic content. Then a separate image retrieval step for each visual entity in the query image is carried out to collect related phrases from the retrieved descriptions. For instance, if a dog is detected in the given image, then the retrieval process returns the phrases referring to visually similar dogs in the training set. Finally, a description is composed from a selection of the retrieved phrases, considering factors such as word order or redundancy.

Compared to models that generate descriptions directly (see Section 2.1), retrieval models typically require a large amount of training data in order to provide relevant descriptions. However human authored sentences or phrases are used in the output, increasing its quality.

2.3 Description as a Retrieval in Multimodal Space

The third group of studies casts image description generation as a retrieval problem from a multimodal space [Hodosh *et al.*, 2013; Socher *et al.*, 2014; Karpathy *et al.*, 2014]. The overall approach is to first learn a common multimodal space for the visual and textual data using a training set of image–description pairs and then given a query, use this joint representation to perform cross-modal (image–sentence) retrieval. The advantage of this approach is that it allows bi-directional models, i.e., the common space can also be used for retrieving the most appropriate image for a query sentence.

Approaches in multimodal retrieval-based systems differ mainly in the way the common multimodal space is learnt. The seminal paper of Hodosh *et al.* [2013] makes use of KCCA, a kernelized version of Canonical Correlation Analysis (CCA). A drawback of KCCA is that it is only applicable to smaller datasets, as it requires two kernel matrices to be kept in memory during training. This becomes prohibitive for very large datasets. Neural network models are more efficient in constructing a multimodal space, and are now the method of choice. For example, Socher *et al.* [2014] use Dependency Tree Recursive Neural Network (DT-RNN) for building sentence representations and a nine layer neural network for building image vector representations that are then mapped into a common embedding space.

Socher's multimodal embedding model was extended by Karpathy *et al.* [2014]. Rather than directly mapping entire images and sentences into a common embedding space, the model embeds more fine-grained units, i.e., fragments of images (objects) and sentences (dependency tree fragments), into a common space. The final model outperforms the DT-RNN approach. Other variants of deep neural networks have been used, for example, Long–Short Term Memory (LSTM) recurrent neural networks [Kiros *et al.*, 2015], or convolutional network (CNN) to compute the multimodal space.

Models based on retrieval and ranking are limited by the availability of very large datasets with descriptions. A good number of multimodal models have therefore been developed to not only to rank sentences, but also to generate them, for example [Kiros *et al.*, 2015; Vinyals *et al.*, 2015; Karpathy and Fei-Fei, 2015].

3 Datasets

A wide range of datasets for automatic image description research is available. Images in these datasets are associated with textual descriptions and differ from each other in terms of size, the format of the descriptions and in how the descriptions were collected. A selection of datasets are listed in Ta-

	Images	Texts	Objects
VLT2K [Elliott and Keller, 2013]	2,424	3	Partial
Flickr30K [Young *et al.*, 2014]	31,783	5	No
MS COCO [Lin *et al.*, 2014]	164,062	5	Partial
BBC News [Feng and Lapata, 2008]	3,361	1	No
SBU1M Captions [Ordonez *et al.*, 2011]	1,000,000	1	No

Table 1: A selection of datasets, image *description* datasets (top) and *caption* datasets (bottom).

ble 1, and an image-description pair example from VLT2K [Elliott and Keller, 2013] is given in Figure 3.

Image descriptions verbalize what can be seen in the image, i.e., they refer to the objects, attributes, actions, scene type, etc. Captions, on the other hand, are texts associated with images that verbalize information that cannot be seen in the image. The texts in image–*description* datasets are usually crowd-sourced from Amazon Mechanical Turk or Crowdflower; whereas the texts in image–*caption* datasets are harvested from photo-sharing sites, such as Flickr, or from news providers. Furthermore, captions are usually collected without financial incentive because they are written by the people sharing their own images, or by journalists.

1. There are several people in chairs and a small child watching one of them play a trumpet
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

Figure 3: An example of an image and descriptions from the VLT2K [Elliott and Keller, 2013] benchmark image dataset.

4 Evaluation Measures

We review two types of evaluations, human judgments and automatic measures. Human evaluations mostly make use of the techniques that have been originally developed for evaluating NLG systems. Typically, judges are provided with the image as well as with the description during evaluation tasks; the evaluation is often performed on Mechanical Turk and includes questions on grammar, content and fluency.

The other approach for evaluating descriptions is to use automatic measures, such as BLEU [Papineni *et al.*, 2002], Meteor [Denkowski and Lavie, 2014], and CIDEr [Vedantam *et al.*, 2015]. These measures were originally developed to evaluate the output of machine translation engines or text summarisation systems, with the exception of CIDEr, which was developed specifically for image description evaluation. All these measures compute a score that indicates the similarity between the system output and one or more human-written reference texts. This approach to evaluation has been subject to much study, discussion, and criticism [Kulkarni *et al.*, 2011; Hodosh *et al.*, 2013; Elliott and Keller, 2014]. More recently [Anderson *et al.*, 2016] proposed a new measure, SPICE, which is based on semantic scene graphs and showed that it correlates much better with human judgements, in comparison to the above standard measures.

5 Future Directions

In spite of the recent advances in the quality of automatically generated descriptions, a series of challenges remain. In this section, we point to future directions that image description generation is likely to benefit from.

Vinyals *et al.* [2015] demonstrate that learning a model from MS COCO and applying it to datasets collected in different settings such as SBU1M Captions or Pascal1K, leads to a degradation in BLEU performance. This is attributed mainly to the differences in vocabulary and in the quality of descriptions. Supervised learning is therefore likely to benefit from larger and diversified datasets that share a common, unified, comprehensive vocabulary. On the other hand taking advantage of larger unsupervised data or weakly supervised methods is another challenge to explore.

Experiments carried out by Fang *et al.* [2015] reveal that human judgments are best used for evaluation. However, since conducting human judgment experiments is costly, there is a need for automatic measures that are more highly correlated with human judgments [Elliott and Keller, 2014].

Current systems are limited in the diversity of output they generate. For example, Devlin *et al.* [2015] show that their best model generates only 47.0% unique descriptions. Systems that generate diverse and original descriptions should not just repeat what they have already seen, but also infer the underlying semantics, allowing it to generate novel descriptions.

Future work should also investigate whether multilinguality results in improved descriptions compared to monolingual baselines. The Multimodal Translation shared task [Specia *et al.*, 2016] is leading efforts in this direction with the Multi30K benchmark dataset [Elliott *et al.*, 2016].

6 Conclusions

In this survey, we discussed recent advances in automatic image description generation. We reviewed a large body of the existing work, highlighting common characteristics and differences between existing research. In addition, we briefly reviewed the existing corpora and automatic evaluation measures, and discussed some future directions for vision and language research.

References

[Anderson *et al.*, 2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016: Lecture Notes in Computer Science*, Springer, Cham, 2016.

[Bernardi *et al.*, 2016] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016.

[Denkowski and Lavie, 2014] Michael Denkowski and Alon Lavie. Meteor Universal: Language Specific Translation

Evaluation for Any Target Language. In *Conference of the European Chapter of the Association for Computational Linguistics Workshop on Statistical Machine Translation*, 2014.

[Devlin *et al.*, 2015] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language Models for Image Captioning: The Quirks and What Works. In *Annual Meeting of the Association for Computational Linguistics*, 2015.

[Elliott and Keller, 2013] Desmond Elliott and Frank Keller. Image Description using Visual Dependency Representations. In *Conference on Empirical Methods in Natural Language Processing*, 2013.

[Elliott and Keller, 2014] Desmond Elliott and Frank Keller. Comparing Automatic Evaluation Measures for Image Description. In *Annual Meeting of the Association for Computational Linguistics*, 2014.

[Elliott *et al.*, 2016] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German Image Descriptions. In *Sixth Workshop on Vision and Language*, 2016.

[Fang *et al.*, 2015] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[Feng and Lapata, 2008] Yansong Feng and Mirella Lapata. Automatic Image Annotation Using Auxiliary Text Information. In *Annual Meeting of the Association for Computational Linguistics*, 2008.

[Hodosh *et al.*, 2013] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[Karpathy *et al.*, 2014] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Advances in Neural Information Processing Systems*, 2014.

[Kiros *et al.*, 2015] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *Advances in Neural Information Processing Systems Deep Learning Workshop*, 2015.

[Kulkarni *et al.*, 2011] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Baby talk: Understanding and generating simple image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[Kuznetsova *et al.*, 2012] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective Generation of Natural Image Descriptions. In *Annual Meeting of the Association for Computational Linguistics*, 2012.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

[Mason and Charniak, 2014] Rebecca Mason and Eugene Charniak. Nonparametric Method for Data-driven Image Captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2014.

[Mitchell *et al.*, 2012] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, A C Berg, Kota Yamaguchi, T L Berg, Karl Stratos, Hal Daume, III, and III. Midge: generating image descriptions from computer vision detections. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2012.

[Ordonez *et al.*, 2011] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, 2011.

[Ortiz *et al.*, 2015] Luis M. G. Ortiz, Clemens Wolff, and Mirella Lapata. Learning to Interpret and Describe Abstract Scenes. In *Conference of the North American Chapter of the Association of Computational Linguistics*, 2015.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002.

[Socher *et al.*, 2014] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Ng. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[Specia *et al.*, 2016] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *Conference on Machine Translation*, 2016.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[Young *et al.*, 2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.