Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics (Extended Abstract) *

Micah Hodosh, Peter Young, Julia Hockenmaier
Department of Computer Science
University of Illinois at Urbana-Champaign
{mhodosh2, pyoung2, juliahmr}@illinois.edu

Abstract

In [Hodosh *et al.*, 2013], we establish a ranking-based framework for sentence-based image description and retrieval. We introduce a new dataset of images paired with multiple descriptive captions that was specifically designed for these tasks. We also present strong KCCA-based baseline systems for description and search, and perform an in-depth study of evaluation metrics for these two tasks. Our results indicate that automatic evaluation metrics for our ranking-based tasks are more accurate and robust than those proposed for generation-based image description.

1 Introduction

Over the last few years, the challenge of developing systems that associate images with natural language sentences that describe what is depicted in them has received a significant amount of attention. Initially, this has been posed as a natural language generation task, in which systems have to automatically produce novel sentences (e.g. [Farhadi *et al.*, 2010; Ordonez *et al.*, 2011; Gupta *et al.*, 2012; Kuznetsova *et al.*, 2012; Mitchell *et al.*, 2012]). In contrast to these approaches, we propose to frame sentence-based image description as the task of ranking a pool of (previously unseen) captions for each test image. This establishes a natural parallel between sentence-based image description and sentence-based image search, and allows models to focus on the semantic question of whether a sentence provides a correct description of an image. But the main advantage of using a ranking-based framework is that it allows for large-scale quantitative evaluations that enable a direct comparison of different systems. Our in-depth comparison of evaluation metrics shows that metrics previously used to evaluate generation-based approaches such as BLEU [Papineni *et al.*, 2002] or ROUGE [Lin, 2004] show much weaker correlation with human judgments than those that can be used for ranking-based tasks. We also provide crowdsourcing-based methods to produce benchmark datasets to train and evaluate image description systems, and release an initial version of such a dataset.

*This paper is an extended abstract of the Journal of Artificial Intelligence Research publication [Hodosh *et al.*, 2013]

2 Contributions of this Work

Our paper makes the following contributions:

- We release a dataset of images that capture a wide range of everyday actions and events, each paired with five sentences that describe what is depicted in them.
- We introduce a ranking-based framework to evaluate systems on a sentence-based image description and a sentence-based image retrieval task.
- We develop a number of strong baseline systems. In contrast to prior work, we show that explicit object and scene detectors may not be required to perform well on this task. We focus on linguistic features, and show that models that capture lexical similarities and word order outperform simple bag of words approaches.
- We compare several automatic evaluation metrics for this task, and show that ranking-based metrics correlate better with human judgements than other metrics such as BLEU or ROUGE.

3 A Dataset for Image Description

We argue that the captions that are used to train and evaluate image description systems need to provide explicit descriptions of what is depicted in the image. However, Gricean maxims [Grice, 1975] imply that the captions that people normally provide for images do not simply restate what can be already observed from the image, but instead provide additional information or context. We therefore argue that datasets that consist of images and captions from newswire [Feng and Lapata, 2010] or Flickr [Ordonez *et al.*, 2011] are not directly suitable for the task that we have in mind.

In this paper, we present two datasets with captions specifically created for sentence-based image description: the PASCAL VOC-2008 dataset and the Flickr 8K dataset. Each image in these datasets is associated with five different captions that describe the entities and events depicted in the image that were collected via crowdsourcing (Amazon Mechanical Turk). By associating each image with multiple, independently produced sentences, our dataset captures some of the linguistic variety that can be used to describe the same image. An example from our Flickr8K dataset is shown in Figure 1. As illustrated by this example, different captions of the same image may focus on different aspects of the scene, or use

A man is doing tricks on a bicycle on ramps in front of a crowd.
A man on a bike executes a jump as part of a competition while the crowd watch
A man rides a yellow bike over a ramp while others watch.
Bike rider jumping obstacles.
Bmx biker jumps off of ramp.

Figure 1: An example of an image from the Flickr 8K dataset. Each of the captions literally describe what is being depicted in the photograph while also mentioning different entities and exhibiting linguistic variation

different linguistic constructions. The PASCAL VOC-2008 dataset consists of 1,000 images randomly selected from the training and validation set of the PASCAL 2008 object recognition challenge [Everingham *et al.*, 2008]. The larger Flickr 8K dataset features $8,092$ "action" images of scenes featuring people and animals.

4 Image Description as a Ranking Task

In order to advance the state of the art of sentence-based image description, we sought to design a task to quantitatively and accurately evaluate image description models. We argue that image description is, at its core, a semantic problem: all systems that perform this task need to capture the quality of a sentence as a description of an image (e.g. through an affinity function or probability distribution).

Much of the related work has focused on models that produce novel captions for images, e.g. [Kulkarni *et al.*, 2011; Gupta *et al.*, 2012; Kuznetsova *et al.*, 2012; Mitchell *et al.*, 2012]. But since these systems have to generate sentences that are not just accurate, but also grammatically correct and appropriate for the image, we argue that framing image description as a natural language generation task introduces syntactic and pragmatic difficulties that distract from the underlying semantic question. Moreover, evaluating these machine-made captions relies on the repeated collection of human judgments, which are difficult to compare across experiments, or automatic scores such as BLEU or ROUGE, that we show to be unreliable metrics for this task.

In order to measure how well a model understands the relationship between an image and the space of appropriate captions that can describe it, we propose to evaluate systems directly on man-made captions. This naturally casts the problem as a ranking or retrieval task. We hold out a pool of images and their corresponding captions, and evaluate for each test image how well the system ranks the caption of that images over the captions of all other test images. This also allows us to evaluate the problem of sentence-based image description in the same framework as sentence-based image search. In both cases, systems are expected to return a query-dependent ranking over the pool of possible responses (captions for image description, images for image search), and in both cases, the pool contains one item that was originally associated with the query. This allows us to define a number of metrics that can be computed automatically. Since it is important to measure the average rank of the correct response, as well as how often the correct response appears among the top few results, we measure recall at a number of fixed ranks k ($R@k$, for $k \in \{1, 5, 10\}$), which corresponds to the percentage of test queries for which the correct response was among the top k results, as well as the median rank r of the correct response among all test queries.

5 Our Image Description Models

In order to capture the association of images and text, we sought to project images and text into a shared latent 'semantic' space, using a technique known as Kernel Canonical Correlation Analysis (KCCA) [Bach and Jordan, 2002; Hardoon *et al.*, 2004]. Given a set of images and their corresponding captions' (kernelized) feature representations, KCCA learns projections such that the representation of images and the captions that describe them are maximally correlated in a new common space. Unlike most of the related work e.g. [Farhadi *et al.*, 2010; Kulkarni *et al.*, 2011; Li *et al.*, 2011; Yang *et al.*, 2011; Ordonez *et al.*, 2011; Mitchell *et al.*, 2012], we therefore, do not need to define an explicit semantic representation that consists only of a fixed number of scenes and objects that are each predicted by pre-trained detectors. Furthermore, [Kuznetsova *et al.*, 2012] specifically evaluate on a subset of images on which their detectors work well. It is also unclear how well these approaches generalize beyond the PASCAL VOC-2008 dataset, since it consists of in-domain images on which the detectors may have been trained. In our experiments, we focus on developing and analyzing the effects of more advanced text representations, and fix the image representation to a baseline spatial pyramid kernel [Lazebnik *et al.*, 2009] over basic color, texture, and SIFT [Lowe, 2004; Vedaldi and Fulkerson, 2008] features.

Since bag-of-words representations ignore the word order within a sentence, they may lose important contextual information, despite their strong word overlap: *'A small child with red hair playing with a large brown dog on white carpet'* looks quite different from *'A small white dog playing with a large red ball on brown grass'*. To capture word order information, we use a subsequence string kernel [Shawe-Taylor and Cristianini, 2004]. Due to the brevity of our image captions, this kernel was truncated to subsequences of up to length three.

Different situations, events, and entities of photographs can be described in a myriad of ways; however, a basic text kernel captures only exact word matches. We extended our text kernel to use a lexical-based similarity kernel to allow for "partial matches" to better capture when two words or phrases describe the same concept. Our final model incorporates two

Image Retrieval: Rank of the original item				
	R@1	**R@5**	**R@10**	**Median r**
NN	$2.5^{\diamond\diamond\diamond}$	$4.7^{\diamond\diamond\diamond}$	$7.2^{\diamond\diamond\diamond}$	$272.0^{\diamond\diamond\diamond}$
BOW1	$4.5^{\diamond\diamond\diamond}$	$14.3^{\diamond\diamond\diamond}$	$20.8^{\diamond\diamond\diamond}$	$67.0^{\diamond\diamond\diamond}$
BOW5	$5.8^{\diamond\diamond}$	$16.7^{\diamond\diamond\diamond}$	$23.6^{\diamond\diamond\diamond}$	$60.0^{\diamond\diamond\diamond}$
TAGRANK	$5.4^{\diamond\diamond\diamond}$	$17.4^{\diamond\diamond\diamond}$	$24.3^{\diamond\diamond\diamond}$	$52.5^{\diamond\diamond\diamond}$
TRI5	$6.0^{\diamond\diamond\diamond}$	$17.8^{\diamond\diamond\diamond}$	$26.2^{\diamond\diamond\diamond}$	$55.0^{\diamond\diamond\diamond}$
TRI5SEM	**7.6**	**20.7**	**30.1**	**38.0**

Table 1: Model performance as measured by the rank of the original image or caption. R@k: percentage of queries for which the correct response was among the first X results. Median r: Median position of the response in the ranked list of results. Superscripts indicate statistically significant difference to TRI5SEM ($^{\diamond\diamond}$: $p \leq 0.05$, $^{\diamond\diamond\diamond}$: $p \leq 0.01$). BoW1 is the Bag of Words baseline using 1 caption per training image, BoW5 is with all 5 captions, TagRank [Hwang and Grauman, 2012], TRI5 is the baseline using the subsequence string kernel [Shawe-Taylor and Cristianini, 2004], and TRI5SEM is our final model incorporating lexical semantics. For the image annotation results, see [Hodosh *et al.*, 2013]

kinds of similarity. We learn a novel alignment-based similarity on our corpus through the machine translation IBM Models 1-2 [Brown *et al.*, 1993] to capture directly when nouns and verbs can refer to the same object. In addition, we use distributional similarity to capture co-occurrence information to push sentences on the same topic closer together. During training, we pool the responses of all five captions, in order to get a more robust and accurate picture of what is important about an image which transfers even when testing on only one response.

In Table 1, we show that training on five captions of an image, as well as using lexical similarity and subsequence kernels significantly increased performance over a baseline nearest-neighbor method, a bag of words kernel trained on one or five captions per image, and the text kernel of [Hwang and Grauman, 2012], which has had success on related tasks.

6 An Analysis of Evaluation Metrics

Generation-based image description systems typically return a single caption for each image, whereas ranking-based systems return a ranked list of captions. This difference has implications for evaluation. In our analysis of evaluation metrics, we were guided by two questions: 1. How well do automatic evaluation metrics for single captions correlate with human judgments? and 2. How well do automatic evaluation metrics for our ranking task correlate with human judgments?

To answer the first question, we had "experts" (students at Illinois who we trained for the task) grade the quality of the first retrieved annotation for each test image on a scale from 1 to 4. Since BLEU [Papineni *et al.*, 2002] and ROUGE [Lin, 2004] scores are commonly used to evaluate generation-based systems [Ordonez *et al.*, 2011; Yang *et al.*, 2011; Kuznetsova *et al.*, 2012], we compared them against these human judgments. Our results indicate that BLEU and ROUGE are not useful metrics for this task. BLEU in particular failed to reveal statistically significant differences between systems that the "experts" identified. Across multiple thresholds for both BLEU and ROUGE, and the expert scores, we found that the correlation with experts maximized at Cohen's $\kappa = 0.72$ and 0.54, respectively. However, this required a threshold that typically occurred only when the returned caption was the one that was originally written for the image. To more realistically measure the quality of 'novel' captions, we remove the query from the reference captions. This reduces the correlation to $\kappa = 0.52$ and 0.51. When only a single reference caption per image is available (as in [Ordonez *et al.*, 2011; Grubinger *et al.*, 2006]), the correlation further decreases to $\kappa = 0.36$ and 0.42.

To evaluate the quality of the ranked lists returned by our system, we had to collect human judgments on a much larger scale. But by directly annotating image caption pairs from our test pool, these judgements can be reused, facilitating comparison between future models. For the top 10 results returned by each of our systems, we collected simple binary judgements through Crowdflower.com at the cost of 0.9¢ per image caption pair. We found that these binary scores correlate strongly with the "expert" scores ($\kappa = 0.79$) while being much more efficient to collect. Since each image may now have multiple relevant captions, we used R-precision [Manning *et al.*, 2008] and "success @k" metrics. While these metrics suggest that the R@k scores underestimate performance, we also show that the scores that rely only on the position of a gold response may be a suitable proxy when human judgements are unavailable, since they result in system rankings that correlate very strongly (with Spearman's ρ of up to 0.97) with those that take human judgments into account. Our analysis suggests that the evaluation of ranking-based image description can be automated.

7 The Current State of Related Work

Since publication, there has been a significant amount of progress utilizing our work.

We have released a larger dataset (Flickr30K) containing over 30,000 images [Young *et al.*, 2014]. In [Gong *et al.*, 2014], we present the first image description evaluation on Flickr30K, and show that a CCA-based approach can benefit significantly from the incorporation of Flickr meta-data.

Following our crowdsourcing approach to collect multiple captions per image, Lin et al. [2014] have recently released the first version of the *Microsoft Common Objects in Context* (COCO) dataset, which currently contains over 82,000 training images and 40,000 validation images harvested from Flickr that are each associated with five captions. Such larger datasets are crucial for this task, since their size should allow for a more meaningful exploration of the subtleties and rarer phenomena of language.

[Elliott and Keller, 2014] also evaluate automatic metrics for image description. Their findings concurred with ours and they suggested using either Meteor [Denkowski and Lavie, 2014], a modified smoothed BLEU score [Clark *et al.*, 2011], or the ROUGE-SU4 variant (skip bigram with a maximum gap of 4 tokens) for novel text.

The Flickr30K and COCO datasets have made it possible for recent advances in deep learning for this task. For example, state-of-the-art neural network vision features such as VGGNet [Simonyan and Zisserman, 2014] have been shown to be expressive enough to significantly increase performance independent of text and multimodal modeling [Mao *et al.*, 2014; Kiros *et al.*, 2014]. Regarding the question of how to use neural networks on the language side, there are currently two schools of thought, both utilizing Recurrent Neural Networks (RNNs) [Elman, 1990] or Long Short-Term Memory Networks (LSTMs) [Hochreiter and Schmidhuber, 1997]. One set of models seek to maximize the probability of generating a sentence token by token, conditioned on the image e.g. [Mao *et al.*, 2014; Kiros *et al.*, 2014; Vinyals *et al.*, 2014; Chen and Zitnick, 2014]. The other school of thought is more similar to our KCCA-based approach in that they use deep-learning to induce a common semantic space for complete sentences (or pieces of text) and images e.g. [Karpathy and Fei-Fei, 2014; Kiros *et al.*, 2014]. The first has the advantage of being able to produce novel captions directly (rather than via a second decoder model), while the latter can more directly perform image search as well as annotation and avoids the need to directly model the distribution of the language.

8 Conclusion

In this work, we have introduced a novel dataset for sentence-based image description and have proposed to evaluate sentence-based image description systems on a ranking task. We have shown that this task lends itself to automatically computable evaluation metrics that correlate highly with human judgments. Ranking-based evaluations are now commonly used by image descriptions papers and we continue to question the usefulness of using BLEU or ROUGE scores, as these metrics fail to correlate strongly with human judgments. Although our models predate recent significant advances and jumps in performance, we have shown that this task can be done without explicit detectors on the vision side, and that significant increases in performance can be obtained on the language side by moving beyond simple bag-of-words models.

9 Acknowledgements

We gratefully acknowledge support for this project from the National Science Foundation through IIS Medium grant 0803603, CAREER award 1053856, CNS-1205627 CI-P and CNS 1405883.

References

[Bach and Jordan, 2002] Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[Brown *et al.*, 1993] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, June 1993.

[Chen and Zitnick, 2014] Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. *CoRR*, abs/1411.5654, 2014.

[Clark *et al.*, 2011] Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, pages 176–181, 2011.

[Denkowski and Lavie, 2014] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

[Elliott and Keller, 2014] Desmond Elliott and Frank Keller. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland, June 2014.

[Elman, 1990] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[Everingham *et al.*, 2008] Mark Everingham, Luc Van Gool, C.K.I. Williams, J. Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/, 2008.

[Farhadi *et al.*, 2010] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of the European Conference on Computer Vision (ECCV), Part IV*, pages 15–29, Heraklion, Greece, September 2010.

[Feng and Lapata, 2010] Yansong Feng and Mirella Lapata. 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)*, pages 1239–1249, July 2010.

[Gong *et al.*, 2014] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision (ECCV)*, 2014.

[Grice, 1975] H. Paul Grice. Logic and conversation. In Donald Davidson and Gilbert H. Harman, editors, *The Logic of Grammar*, pages 64–75. Dickenson Publishing Co., 1975.

[Grubinger *et al.*, 2006] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The IAPR benchmark: A new evaluation resource for visual information systems. In *OntoImage 2006, Workshop on Language Resources for Content-based Image Retrieval during LREC 2006*, pages 13–23, Genoa, Italy, May 2006.

[Gupta *et al.*, 2012] Ankush Gupta, Yashaswi Verma, and C. Jawahar. Choosing linguistics over vision to describe images. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, July 2012.

[Hardoon *et al.*, 2004] David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664, December 2004.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[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 Artifical Intellegence*, 47:853–899, 2013.

[Hwang and Grauman, 2012] SungJu Hwang and Kristen Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International Journal of Computer Vision*, 100(2):134–153, 2012.

[Karpathy and Fei-Fei, 2014] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

[Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[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 *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1608, 2011.

[Kuznetsova *et al.*, 2012] Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 359–368, Jeju Island, Korea, July 2012.

[Lazebnik *et al.*, 2009] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Spatial pyramid matching. In B. Schiele S. Dickinson, A. Leonardis and M. Tarr, editors, *Object Categorization: Computer and Human Vision Perspectives*, chapter 21, pages 401–415. Cambridge University Press, 2009.

[Li *et al.*, 2011] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 220–228, Portland, OR, USA, June 2011.

[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 (ECCV)*, 2014.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, July 2004.

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

[Manning *et al.*, 2008] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[Mao *et al.*, 2014] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632, 2014.

[Mitchell *et al.*, 2012] Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daume III. 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)*, pages 747–756, 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 24*, pages 1143–1151, 2011.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, July 2002.

[Shawe-Taylor and Cristianini, 2004] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[Simonyan and Zisserman, 2014] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[Vedaldi and Fulkerson, 2008] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[Vinyals *et al.*, 2014] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.

[Yang *et al.*, 2011] Yezhou Yang, Ching Teo, Hal Daume III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 444–454, 2011.

[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.