

# Exploring Nearest Neighbor Approaches for Image Captioning

Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, C. Lawrence Zitnick

**Abstract**—We explore a variety of nearest neighbor baseline approaches for image captioning. These approaches find a set of nearest neighbor images in the training set from which a caption may be borrowed for the query image. We select a caption for the query image by finding the caption that best represents the “consensus” of the set of candidate captions gathered from the nearest neighbor images. When measured by automatic evaluation metrics on the MS COCO caption evaluation server, these approaches perform as well as many recent approaches that generate novel captions. However, human studies show that a method that generates novel captions is still preferred over the nearest neighbor approach.



## 1 INTRODUCTION

The automatic generation of captions for images has recently received significant attention [18], [26], [35], [16], [19], [6], [8], [2], [23], [22], [21], [36]. This surge in research is due in part to the creation of large caption datasets [11], [29], [13], [37], [1], [24], and new learning techniques [20], [12]. Recently proposed methods for caption generation share many similarities, including the use of deep learned image features [20], [14], [32], and language models using maximum entropy [8], recurrent neural networks [2], [16], and LSTMs [35], [6]. An integral feature of all these methods is their ability to generate novel captions.

We seek to better understand how important the generation of *novel* captions is for the task of automatic image captioning when using the benchmark MS COCO dataset [24]. Previously, several papers proposed producing image captions by first finding similar images, and then copying their captions [9], [29], [13]. Given larger caption datasets such as the MS COCO [24] dataset, which contains 100,000s of captions, the chances of finding an appropriate caption may increase, making such approaches more useful. Vinyals et al. [35] found that up to 80% of the captions generated by their approach were identical to captions in the MS COCO training dataset, while still achieving near state-of-the-art results. This provides evidence that copying captions may indeed achieve good results. However, if the images in the MS COCO dataset contain too much diversity, or capture many rare occurrences, approaches that copy captions directly may not perform as well as those that can additionally generate novel captions.

In this paper we expand on [5] by providing a detailed exploration into nearest neighbor (NN) approaches for image captioning. Nearest neighbor approaches have a rich history in work on predicting words given images, used in face recognition [31] as well as recent work in retrieval-based caption generation [9], [27]. We focus on nearest neighbor approaches to gain further insight into the limitations of the captioning task, and to explore the properties of the largest captioning dataset to date, the MS COCO dataset. We hope to

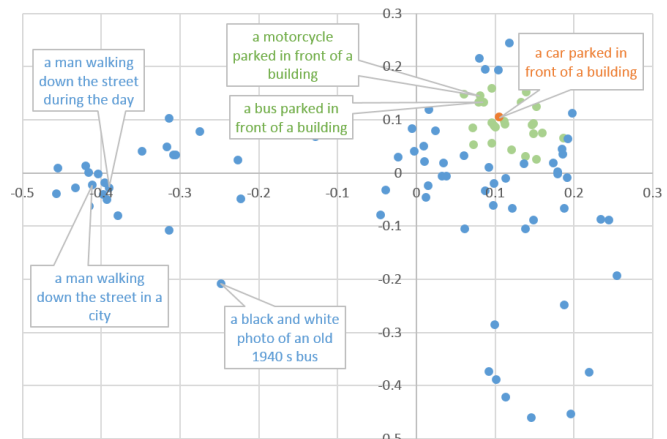


Fig. 1: Example of the set of candidate captions for an image, the highest scoring  $m$  captions (green) and the consensus caption (orange). This is a real example visualized in two dimensions.

provide context for the recent advances in this area [18], [26], [35], [16], [19], [6], [8], [2], [23], [22], [21], [36].

Our nearest neighbor approach finds a set of  $k$  nearest images. Which images are “nearest” can be defined in several ways, and we examine using GIST [28], pre-trained deep features [32], and deep features fine-tuned for the task of caption generation [8]. Once a set of  $k$  NN images are found, the captions describing these images are combined into a set of candidate captions from which the final caption is selected, Figure 1. We select the best candidate caption by finding the one that scores highest with respect to the other candidate captions. We refer to this as the “consensus” caption. The scores between pairs of captions are computed using either the CIDEr [34] or BLEU [30] metric.

Surprisingly, we find that this simple NN approach outperforms many novel caption generation approaches as measured by BLEU [30], METEOR [4] and CIDEr [34] on the MS COCO testing [24] dataset. We find that using simple features for finding nearest neighbors such as GIST [28] do not perform well. However, deep features, especially those fine-tuned specifically for caption generation [8] are very effective at finding images from which high-scoring captions may be borrowed. While the NN approaches perform well

- J. Devlin, R. Girshick, M. Mitchell, and C. L. Zitnick are with Microsoft Research, Redmond.
- Saurabh Gupta is with University of California, Berkeley.

when evaluated using automatic metrics, a crowdsourced study shows that humans still prefer a system that generates novel captions [8] by a significant margin. Further human studies still need to be performed to see how the NN approaches compare to other generation-based approaches.

## 2 RELATED WORK

Several early papers proposed producing image captions by copying captions from other images [9], [29], [13], [15]. [9] use nearest neighbors to define image and caption features, capturing information about objects, actions, and scenes, where [29] use a combination of object, stuff, people and scene information. [27] use GIST nearest neighbors to the query image. [13] use Kernel Canonical Correlation Analysis to map images and captions to a common space where the nearest caption can be found. While not using explicit captions, [15] explores the task of captioning images using surrounding text on webpages.

Hodosh et al. [13] popularized the task of image and caption ranking. That is, given an image, rank a set of captions based on which are most relevant. They argued that this task was more correlated with human judgment than the task of novel caption generation measured using automatic metrics. Numerous papers have explored the task of caption ranking [26], [33], [10], [17], [2], [25]. These approaches could also be used to rank the set of training captions, and used to select the one that is most relevant. As far as we are aware, how well such an approach would perform on the MS COCO caption dataset for generation is still an open question. In this paper, we only explore a simple nearest neighbor baseline approach.

## 3 APPROACH

In this section, we describe our set of approaches for image captioning. We assume a dataset of training images with a set of corresponding captions. We use the MS COCO [24] training dataset containing 82,783 images with 5 captions each, for a total of 413,915 captions. In our approach [5] we first find a set of  $k$  NN images in the training dataset. The consensus caption returned by our approach is selected from the set of candidate captions describing the set of  $k$  NN training images.

### 3.1 Nearest Neighbor Images

Our first task is to find a set of  $k$  nearest training images for each query image based on visual similarity.<sup>1</sup> We find the  $k$  NNs using cosine similarity with the following feature spaces:

- **GIST**: We use the popular approach of [28] to compute a set of global image features based on the summation of low-level image features, such as contours or textures. GIST is computed on images resized to  $32 \times 32$  pixels.
- **fc7**: Our first set of deep features are computed using the fc7 layer of the VGG16 Net [32]. The network was trained using the 1,000 ImageNet classification task [3]. The features are computed using a single window with resolution  $224 \times 224$ . Images are rescaled to make the longer side 224 pixels. Empty image regions were replaced by the mean image.

- **fc7-fine**: These features are computed in the same manner as fc7. However, the weights of the VGG16 network are fine-tuned for the image captioning task. Specifically, its weights are initialized using the ImageNet task, and the weights are fine-tuned on the task of classifying the 1,000 most commonly occurring words in image captions [8].

The image features are computed for every image in the training dataset. The neighbor images are found by exhaustively computing the cosine similarity between the query image and the training images. In Figure 2, we show several examples of NN matches using different feature spaces. Notice how the NNs found using deep features are more semantically similar.

### 3.2 Consensus Caption

Given  $k$  nearest training images for a given test image, we take the union of their captions to create a set  $C$  of  $n$  candidate captions. Our task is to select the “best” or consensus caption from the set  $C$ , as seen in Figure 1. There are five captions per image in the MS COCO dataset, so  $n = 5k$ . We define the consensus caption  $c^*$  as the one that has the highest average lexical similarity to the other captions in  $C$ . This scoring function is:

$$c^* = \operatorname{argmax}_{c \in C} \sum_{c' \in C} \operatorname{Sim}(c, c'), \quad (1)$$

where  $\operatorname{Sim}(c, c')$  is the similarity score between two captions  $c$  and  $c'$ . We explore two similarity functions: BLEU [30], which measures 1-to-4-gram overlap, and CIDEr [34], which measures tf-idf weighted 1-to-4-gram overlap. Intuitively, this tf-idf weighting means that CIDEr pays more attention to rarer, more descriptive phrases. We use the CIDEr-D variant of CIDEr.

Some of the candidate captions in  $C$  might be outliers and add noise to the computation of Equation (1). A solution to this problem is to compute Equation (1) only over a subset  $M$  of  $C$ , where the number  $m$  of captions in  $M$  is less than  $n$ . This can be thought of as finding the centroid of a large cluster of captions, as demonstrated in Figure 1. Using  $M$ , our final consensus caption is:

$$c^* = \operatorname{argmax}_{c \in C} \max_{M \subset C} \sum_{c' \in M} \operatorname{Sim}(c, c'). \quad (2)$$

The inner maximization is over all size- $m$  subsets  $M$  of  $C$ .

Intuitively, the consensus caption is a *single* caption from the training data that can be used to describe *many* images that are visually similar to the test image. Ideally, then, this caption is likely to also be an adequate description of the test image.

If the NN images are diverse, one would expect the chosen caption to be more generic, while if the NN images are quite similar, the selected caption may be specific. The reason is that the descriptiveness of captions is a basic risk vs. reward trade-off: a red car is a better description than a car if the car is red, but a significantly worse description when the car is actually blue. As a result, the caption’s detail is directly dependent on the diversity of the dataset used to gather candidate captions.

In Figure, 6 we show several examples of consensus captions using both CIDEr and BLEU. Subjectively, we can see

1. The value of  $k$  is chosen optimally for each feature set, and typically ranges from 50-200.



Fig. 2: Example nearest neighbor matches using different feature spaces.

that the CIDEr-tuned captions tend to be more descriptive than the BLEU-tuned captions, which is likely due to CIDEr’s preference for rarer n-grams.

## 4 RESULTS

In this section, we provide results for several variants of the NN approach. For all our experiments, we use 82,783 MS COCO training images for training, and split 40,504 validation set into two halves: a “tuning” set for hyperparameter optimization, and a “testval” set for reporting results.

We begin by exploring the effect of  $k$  and  $m$  on the final accuracy. Next, we explore the different feature spaces that may be used to find NN images. Finally, we perform human studies to evaluate how well NN approaches perform as judged by humans, and report results on the MS COCO testing set.

### 4.1 The Number of Nearest Neighbors

How important is the selection of  $k$ , the number of nearest neighbor images used to create the candidate caption set  $C$ ? In Figure 3, we show BLEU scores [30] as we vary  $k$ . Notice that for  $k < 20$ , significantly worse results are achieved. For reference, if only one image is selected and a caption is randomly chosen, the BLEU score is 11.2 [6]. For  $k > 60$  the BLEU scores are roughly similar.

In Figure 4, we show results when we vary  $m$ , the number of candidate captions used to select the consensus caption (Equation 2). High scores are achieved for a variety of values for  $m$  ranging between 50 and 200. If  $m = n$ , worse results are achieved, supporting our hypothesis that outlier captions should be removed.

As shown in Table 2, finding the best caption in Equation (2) is slightly better using CIDEr (fc7-fine (CIDEr)) than BLEU (fc7-fine (BLEU)). CIDEr performs better as measured by both CIDEr and METEOR. Not surprisingly, optimizing using BLEU preforms better when measured by BLEU.

### 4.2 Different Feature Spaces

We now explore the effect of using different feature spaces. For these experiments, we use the values of  $k$  and  $m$  that produced the highest BLEU/CIDEr scores on the tuning half of the validation set. Results on the testval half of the validation set

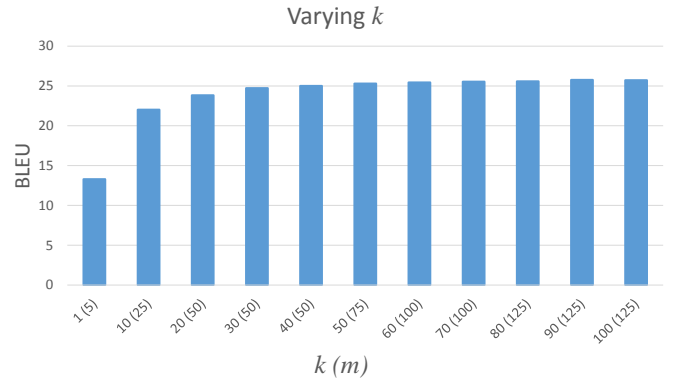


Fig. 3: Resulting BLEU scores when varying number of NN images,  $k$ . The optimal  $m$  for each  $k$  is shown in parentheses.

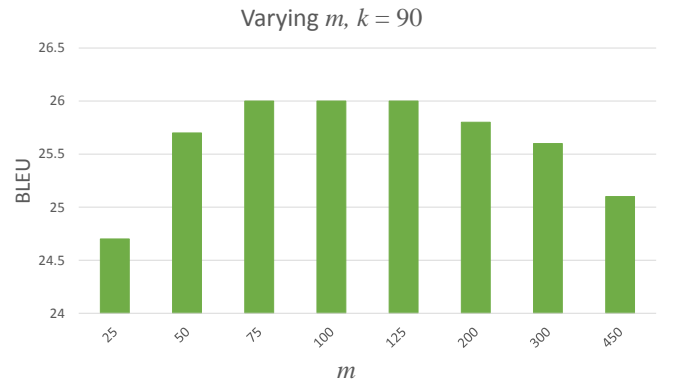


Fig. 4: Resulting BLEU scores when varying the number of captions  $m$  used to compute the consensus score.  $k$  is held constant at 90.

are shown in Table 2.<sup>2</sup> GIST performs poorly since it doesn’t capture the high-level semantics of the scenes as shown in Figure 2. The deeply learned features fc7 and fc7-fine do significantly better. For comparison, we show the results of [8] (ME+DMSM) using a Maximum Entropy (ME) language model and a Deep Multimodal Similarity Model (DMSM). Interestingly, the fc7-fine (BLEU) performs comparably to the ME+DMSM approach [8] as measured by BLEU.

How do these methods perform on query images that are *not* visually similar to the training data, versus images that are visually similar? To gain insight into this question, we

2. Results were computed on 4 references rather than 5 for consistency with [8].



Method	c5			c40		
	BLEU 4	CIDEr	METEOR	BLEU 4	CIDEr	METEOR
ME + DMSM [8]	<b>29.1</b>	<b>0.912</b>	24.7	<b>56.7</b>	<b>0.925</b>	33.1
LRCN [6]	27.7	0.869	24.2	53.4	0.891	32.2
Vinyals et al. [35]	27.2	0.834	23.6	53.8	0.842	32.7
Xu et al. [36]	26.8	0.850	24.3	52.3	0.878	32.3
m-RNN [25]	27.9	0.819	22.9	54.3	0.828	31.2
MLBL [18], [19]	26.0	0.740	21.9	51.7	0.752	29.4
NeuralTalk [16]	22.4	0.674	21.0	44.6	0.692	28.0
fc7-fine (CIDEr)	27.9 (2)	0.886 (2)	23.7 (3)	54.2 (2)	0.916 (2)	31.8 (5)
Human	21.7	0.854	<b>25.2</b>	47.1	0.910	<b>33.5</b>

TABLE 1: Results on the MS COCO test set for c5 (left) and c40 (right). Best scores are shown in bold. Results on *fc7-fine* (CIDEr) are shown, with its relative ranking compared to the automatic approaches shown in parentheses. For comparison, results using captions written by humans are also shown.

Features	k	m	BLEU	CIDEr	METEOR
GIST	80	100	9.0	0.23	12.2
fc7	130	150	22.3	0.72	20.3
fc7-fine (BLEU)	90	125	26.0	0.85	22.5
fc7-fine (CIDEr)	80	200	25.1	0.90	22.8
ME + DMSM [8]			25.7	0.92	23.6

TABLE 2: BLEU [30], METEOR [4] and CIDEr [34] scores on testval for NN approaches using different feature spaces. See text for descriptions of the feature spaces.

Approach	Human Judgements			BLEU
	Better	Equal	Better or Equal	
<i>k</i> -NN fc7-fine (BLEU)	5.5%	22.1%	27.6%	26.0
<i>k</i> -NN fc7-fine (CIDEr)	6.3%	20.2%	26.5%	25.1
ME + DMSM [8]	7.8%	26.2%	34.0%	25.7

TABLE 3: Results when comparing produced captions to those written by humans, as judged by humans. The percentage that are better than, equal to, and better than or equal to the captions written by humans are shown.

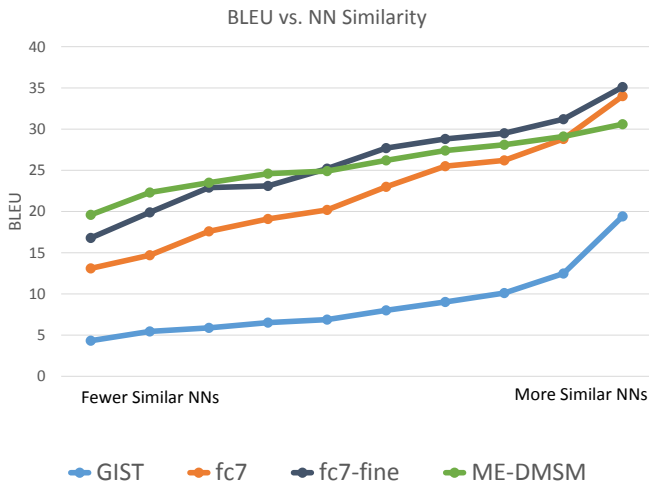


Fig. 5: BLEU scores for various approaches when the testval images are split into 10 equally sized bins based on visual similarity to the training data. The bins are arranged from those with fewer close NNs (left), to those with more NNs images (right).

compute the mean distance of the testval images to their 50 nearest training images based on *fc7-fine* cosine distance. We then sort the mean distances and place the query testval images into ten bins, ranging from those with the closest NN images to those with the furthest.

The BLEU scores for each of these bins are shown in Figure 5. Unsurprisingly, the images that are most visually similar to the training achieve the highest BLEU scores across all approaches. However, compared to the generation-based approach of [8], the nearest neighbor approaches perform better for highly similar images, but worse for highly dissimilar images. This suggests that the generation-based approach generalizes better to less common images, but doesn’t do as well as borrowing captions for common images.

### 4.3 Human Evaluation

An interesting question is how the captions selected by the NN approaches would perform when judged by humans. To explore this, we use the same experimental setup as [8], in which human subjects are asked to judge whether a caption generated by a system is better than, worse than, or equal to a caption written by a human for that image. Each caption is evaluated 5 times and the majority is recorded. If a tie occurs (2-2-1), each of the top choices are given half a vote. The results are shown in Table 3. The generation-based approach of [8] significantly outperforms the nearest neighbor approaches, despite the similar BLEU scores. As a point of reference, a baseline system from [8] that uses non-fine-tuned *fc7* features achieves 21.1 BLEU and 23.3% “Better or Equal to Human.” Therefore, we believe that the 27.6% achieved by the *k*-NN models is still relatively competitive with respect to the state-of-the-art.

However, given the strong BLEU/CIDEr performance of the nearest neighbor systems, this provides additional evidence that automatic metrics may only be a rough estimate of human judgments, as also noted in [7], [13], [34].

### 4.4 MS COCO Caption Test

In Table 1, we show results on *fc7-fine* (CIDEr) on the MS COCO caption test set. Surprisingly, *fc7-fine* is ranked second or third by most metrics. The METEOR metric computed using 40 captions per image (c40) ranks *fc7-fine* fifth. As stated before, further human studies are still needed to gain a better understanding into how the captions produced by NN approaches are perceived by humans relative to generative approaches.

## 5 DISCUSSION

The success of nearest neighbor approaches to image captioning draws attention to the need for better evaluation and testing datasets. Ideally, we desire approaches that can generalize to



















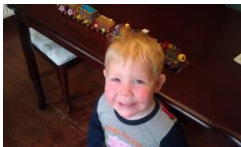

Image	Selected Caption (BLEU)	Selected Caption (CIDEr)	Image	Selected Caption (BLEU)	Selected Caption (CIDEr)
	A bedroom with a bed and a couch.	A hotel room with two beds and a table.		A man riding a wave on a surfboard.	A man riding a wave on a surfboard in the ocean.
	A train is stopped at a train station.	A red and white train parked in a train station.		A person flying a kite in the sky.	A person flying a kite in the sky.
	A group of people sitting around in a living room.	A group of people sitting on a couch in a living room.		A cat sitting in a bathroom sink.	A black and white cat sitting in a bathroom sink.
	A group of people washing elephants in the water.	An elephant is swimming in the water near the rocks.		A wooden bench in front of a building.	A wooden bench in front of a building.
	Two zebras and a giraffe in a field.	Two zebras and a giraffe in a field.		A baby elephant is standing in a field.	An elephant is walking through a grassy field.
	A car parked in front of a building.	A motorcycle parked in front of a brick building.		A cup of coffee on a plate with a spoon.	A plate of food and a cup of coffee.
	A laptop computer sitting on top of a desk.	A laptop computer sitting on top of a desk.		A group of people sitting at a table with laptop computers.	A group of people sitting around a table with laptops.
	A clock sitting on top of a table.	A white airplane hanging from a ceiling in a museum.		A wooden bench in front of a building.	A window display on the front of a building.
	A baseball player holding a bat on a field.	A baseball player holding a bat on a field.		A building with a clock on the top.	A clock tower on the top of a building.
	A little boy sitting at a table eating food.	A little boy sitting at the table with food.		The side of a passenger train at a train station.	A bus that is on the side of a road.

Fig. 6: Several examples of randomly selected images and their selected consensus captions. The consensus caption is shown using the BLEU metric and CIDEr metric for scoring. Notice the chosen captions using CIDEr are more detailed.

images beyond those found in the training set. How can we build a testing set that measures this ability? One obvious approach is to measure the similarity of each testing image with those in the training set, similar to Figure 5. We could then examine how well approaches do on unusual or more diverse images. Another option would be to collect a new testing dataset using a different set of queries than those used for the MS COCO dataset. This would ensure the distribution of images in testing and training is different, and help us measure how well our approaches generalize.

The success of recent approaches such as [18], [26], [35], [16], [19], [6], [8], [2], [23], [22], [21] demonstrates another problem with the task of image captioning. For each of these approaches, we know that human generated captions are typically preferred over the automatically generated captions. However, for many automatic evaluation metrics, the human captions have lower scores than the automatically generated captions. This suggests that the advancement towards human-like captions may not be properly benchmarked using automatic approaches. Further research into automatic evaluation metrics that are highly correlated with human judgment is essential [7], [13], [34].

A further difficulty when performing human evaluations on which caption is “best” is that we miss the nuances in the similarities/differences between the systems when we ask for humans’ overall preferences. For example, we know that all of the NN captions are pretty fluent, while that may not be true for novel generated captions. However, we’re not directly measuring fluency, so it is possible that generation-focused approaches are correctly capturing content, but are sometimes not fluent. On the other hand, NN approaches may produce very generic captions, causing them to not be preferred even when technically correct. Hopefully future experiments will shed light on these questions.

In this paper, we only explored very simple NN approaches to provide a baseline for the image captioning community. More sophisticated approaches that have been proposed for the task of caption ranking [13], [26], [33], [10], [17], [2] may generate even better results. It may also be interesting to explore hybrid approaches that use NN approaches for query images with many similar training images and generation-based approaches for other images.

## REFERENCES

- [1] J. Chen, P. Kuznetsova, D. Warren, and Y. Choi. Déjà image-captions: A corpus of expressive image descriptions in repetition. In *NAACL*, 2015. 1
- [2] X. Chen and C. L. Zitnick. Mind’s eye: A recurrent visual representation for image caption generation. *CVPR*, 2015. 1, 2, 6
- [3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 2
- [4] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *EACL Workshop on Statistical Machine Translation*, 2014. 1, 4
- [5] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*, 2015. 1, 2
- [6] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CVPR*, 2015. 1, 3, 4, 6
- [7] D. Elliott and F. Keller. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 452–457, 2014. 4, 6
- [8] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. *CVPR*, 2015. 1, 2, 3, 4, 6
- [9] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, 2010. 1, 2
- [10] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013. 2, 6
- [11] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *LREC Workshop on Language Resources for Content-based Image Retrieval*, 2006. 1
- [12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1
- [13] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 47:853–899, 2013. 1, 2, 4, 6
- [14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 1
- [15] A. Kannan, S. Baker, K. Ramnath, J. Fiss, D. Lin, L. Vanderwende, R. Ansary, A. Kapoor, Q. Ke, M. Uyttendaele, X.-J. Wang, and L. Zhang. Mining text snippets for images on the web. In *Proc. of the 20th ACM SIGKDD*, KDD ’14, pages 1534–1543, 2014. 2
- [16] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *CVPR*, 2015. 1, 4, 6
- [17] A. Karpathy, A. Joulin, and F.-F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 2, 6
- [18] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *ICML*, 2014. 1, 4, 6
- [19] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 1, 4, 6
- [20] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 1
- [21] A. Lazaridou, N. T. Pham, and M. Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015. 1, 6
- [22] R. Lebre, P. O. Pinheiro, and R. Collobert. Simple image description generator via a linear phrase-based approach. *arXiv preprint arXiv:1412.8419*, 2014. 1, 6
- [23] R. Lebre, P. O. Pinheiro, and R. Collobert. Phrase-based image captioning. *arXiv preprint arXiv:1502.03671*, 2015. 1, 6
- [24] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 2
- [25] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632, 2014. 2, 4
- [26] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014. 1, 2, 6
- [27] R. Mason and E. Charniak. Domain-specific image captioning. In *CoNLL*, 2014. 1, 2
- [28] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006. 1, 2
- [29] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 1, 2
- [30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 1, 2, 3, 4
- [31] P. Phillips and E. Newton. Meta-analysis of face recognition algorithms. *ICAFGR*, 2002. 1
- [32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2
- [33] R. Socher, Q. Le, C. Manning, and A. Ng. Grounded compositional semantics for finding and describing images with sentences. In *NIPS Deep Learning Workshop*, 2013. 2, 6
- [34] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. *arXiv preprint arXiv:1411.5726*, 2014. 1, 2, 4, 6
- [35] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CVPR*, 2015. 1, 4, 6
- [36] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015. 1, 4
- [37] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 1