Language Models for Image Captioning: The Quirks and What Works

Jacob Devlin★, Hao Cheng♠, Hao Fang♠, Saurabh Gupta♣,
Li Deng, Xiaodong He★, Geoffrey Zweig★, Margaret Mitchell★

Microsoft Research

★ Corresponding authors: {jdevlin,xiaohe,gzweig,memitc}@microsoft.com
♠ University of Washington
♣ University of California at Berkeley

Abstract

Two recent approaches have achieved state-of-the-art results in image captioning. The first uses a pipelined process where a set of candidate words is generated by a convolutional neural network (CNN) trained on images, and then a maximum entropy (ME) language model is used to arrange these words into a coherent sentence. The second uses the penultimate activation layer of the CNN as input to a recurrent neural network (RNN) that then generates the caption sequence. In this paper, we compare the merits of these different language modeling approaches for the first time by using the same state-of-the-art CNN as input. We examine issues in the different approaches, including linguistic irregularities, caption repetition, and data set overlap. By combining key aspects of the ME and RNN methods, we achieve a new record performance over previously published results on the benchmark COCO dataset. However, the gains we see in BLEU do not translate to human judgments.

1 Introduction

Recent progress in automatic image captioning has shown that an image-conditioned language model can be very effective at generating captions. Two leading approaches have been explored for this task. The first decomposes the problem into an initial step that uses a convolutional neural network to predict a bag of words that are likely to be present in a caption; then in a second step, a maximum entropy language model (ME LM) is used to generate a sentence that covers a minimum number of the detected words (Fang et al., 2015). The second approach uses the activations from final hidden layer of an object detection CNN as the input to a recurrent neural network language model (RNN LM). This is referred to as a Multimodal Recurrent Neural Network (MRNN) (Karpathy and Fei-Fei, 2015; Mao et al., 2015; Chen and Zitnick, 2015). Similar in spirit is the the log-bilinear (LBL) LM of Kiros et al. (2014).

In this paper, we study the relative merits of these approaches. By using an identical state-of-the-art CNN as the input to RNN-based and ME-based models, we are able to empirically compare the strengths and weaknesses of the language modeling components. We find that the approach of directly generating the text with an MRNN[1] outperforms the ME LM when measured by BLEU on the COCO dataset (Lin et al., 2014),[2] but this recurrent model tends to reproduce captions in the training set. In fact, a simple k-nearest neighbor approach, which is common in earlier related work (Farhadi et al., 2010; Mason and Charniak, 2014), performs similarly to the MRNN. In contrast, the ME LM generates the most novel captions, and does the best at captioning images for which there is no close match in the training data. With a Deep Multimodal Similarity Model (DMSM) incorporated,[3] the ME LM significantly outperforms other methods according to human judgments. In sum, the contributions of this paper are as follows:

1. We compare the use of discrete detections and continuous valued CNN activations as the conditioning information for language models trained to generate image captions.
2. We show that a simple k-nearest neighbor retrieval method performs at near state-of-the-art for this task and dataset.
3. We demonstrate that a state-of-the-art

[1] In our case, a gated recurrent neural network (GRNN) is used (Cho et al., 2014), similar to an LSTM.

[2] This is the largest image captioning dataset to date.

[3] As described by Fang et al. (2015).

MRNN-based approach tends to reconstruct previously seen captions; in contrast, the two stage ME LM approach achieves similar or better performance while generating relatively novel captions.

4. We advance the state-of-the-art BLEU scores on the COCO dataset.
5. We present human evaluation results on the systems with the best performance as measured by automatic metrics.
6. We explore several issues with the statistical models and the underlying COCO dataset, including linguistic irregularities, caption repetition, and data set overlap.

2 Models

All language models compared here are trained using output from the same state-of-the-art CNN. The CNN used is the 16-layer variant of VGGNet (Simonyan and Zisserman, 2014) which was initially trained for the ILSVRC2014 classification task (Russakovsky et al., 2015), and then fine-tuned on the Microsoft COCO data set (Fang et al., 2015; Lin et al., 2014).

2.1 Detector Conditioned Models

We study the effect of leveraging an explicit detection step to find key objects/attributes in images before generation, examining both an ME LM approach as reported in previous work (Fang et al., 2015), and a novel LSTM approach introduced here. Both use a CNN trained to output a bag of words indicating the words that are likely to appear in a caption, and both use a beam search to find a top-scoring sentence that contains a subset of the words. This set of words is dynamically adjusted to remove words as they are mentioned.

We refer the reader to Fang et al. (2015) for a full description of their ME LM approach, whose 500-best outputs we analyze here.[4] We also include the output from their ME LM that leverages scores from a Deep Multimodal Similarity Model (DMSM) during n-best re-ranking. Briefly, the DMSM is a non-generative neural network model which projects both the image pixels and caption text into a comparable vector space, and scores their similarity.

In the LSTM approach, similar to the ME LM approach, we maintain a set of likely words $\mathcal{D}$ that have not yet been mentioned in the caption under construction. This set is initialized to all the words predicted by the CNN above some threshold α.[5] The words already mentioned in the sentence history h are then removed to produce a set of conditioning words $\mathcal{D} \setminus \{\mathbf{h}\}$. We incorporate this information within the LSTM by adding an additional input encoded to represent the remaining visual attributes $\mathcal{D} \setminus \{\mathbf{h}\}$ as a continuous valued auxiliary feature vector (Mikolov and Zweig, 2012). This is encoded as $f(\mathbf{s}_{h_{-1}} + \sum_{v \in \mathcal{D} \setminus \{\mathbf{h}\}} \mathbf{g}_v + \mathbf{U}\mathbf{q}_{\mathbf{h},\mathcal{D}})$, where $\mathbf{s}_{h_{-1}}$ and $\mathbf{g}_v$ are respectively the continuous-space representations for last word h_{-1} and detector $v \in \mathcal{D} \setminus \{\mathbf{h}\}$, $\mathbf{U}$ is learned matrix for recurrent histories, and $f(\cdot)$ is the sigmoid transformation.

2.2 Multimodal Recurrent Neural Network

In this section, we explore a model directly conditioned on the CNN activations rather than a set of word detections. Our implementation is very similar to captioning models described in Karpathy and Fei-Fei (2015), Vinyals et al. (2014), Mao et al. (2015), and Donahue et al. (2014). This joint vision-language RNN is referred to as a Multimodal Recurrent Neural Network (MRNN).

In this model, we feed each image into our CNN and retrieve the 4096-dimensional final hidden layer, denoted as `fc7`. The `fc7` vector is then fed into a hidden layer H to obtain a 500-dimensional representation that serves as the initial hidden state to a gated recurrent neural network (GRNN) (Cho et al., 2014). The GRNN is trained jointly with H to produce the caption one word at a time, conditioned on the previous word and the previous recurrent state. For decoding, we perform a beam search of size 10 to emit tokens until an END token is produced. We use a 500-dimensional GRNN hidden layer and 200-dimensional word embeddings.

2.3 k-Nearest Neighbor Model

Both Donahue et al. (2015) and Karpathy and Fei-Fei (2015) present a 1-nearest neighbor baseline. As a first step, we replicated these results using the cosine similarity of the `fc7` layer between each test set image t and training image r. We randomly emit one caption from t's most similar training image as the caption of t. As reported in previous results, performance is quite poor, with a BLEU

[4] We will refer to this system as D-ME.

[5] In all experiments in this paper, α=0.5.

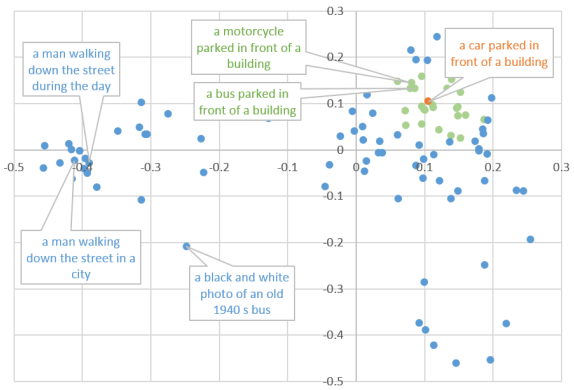

Figure 1: Example of the set of candidate captions for an image, the highest scoring m captions (green) and the consensus caption (orange). This is a real example visualized in two dimensions.

score of 11.2%.

However, we explore the idea that we may be able to find an optimal k-nearest neighbor *consensus caption*. We first select the $k = 90$ nearest training images of a test image t as above. We denote the union of training captions in this set as $C = c_1, ..., c_{5k}$.[6] For each caption c_i, we compute the n-gram overlap F-score between c_i and each other caption in C. We define the consensus caption c^* to be caption with the highest mean n-gram overlap with the other captions in C. We have found it is better to only compute this average among c_i's $m = 125$ most similar captions, rather than all of C. The hyperparameters k and m were obtained by a grid search on the validation set.

A visual example of the consensus caption is given in Figure 1. Intuitively, we are choosing a single caption that may describe *many* different images that are similar to t, rather than a caption that describes the *single* image that is most similar to t. We believe that this is a reasonable approach to take for a retrieval-based method for captioning, as it helps ensure incorrect information is not mentioned. Further details on retrieval-based methods are available in, e.g., (Ordonez et al., 2011; Hodosh et al., 2013).

3 Experimental Results

3.1 The Microsoft COCO Dataset

We work with the Microsoft COCO dataset (Lin et al., 2014), with 82,783 training images, and the validation set split into 20,243 validation images and 20,244 testval images. Most images contain multiple objects and significant contextual information, and each image comes with 5 human-annotated captions. The images create a challenging testbed for image captioning and are widely used in recent automatic image captioning work.

[6]Each training image has 5 captions.

LM	PPLX	BLEU	METEOR
D-ME†	18.1	23.6	22.8
D-LSTM	14.3	22.4	22.6
MRNN	13.2	25.7	22.6
k-Nearest Neighbor	-	26.0	22.5
1-Nearest Neighbor	-	11.2	17.3

Table 1: Model performance on testval. †: From (Fang et al., 2015).

D-ME+DMSM	a plate with a sandwich and a cup of coffee
MRNN	a close up of a plate of food
D-ME+DMSM+MRNN	a plate of food and a cup of coffee
k-NN	a cup of coffee on a plate with a spoon

D-ME+DMSM	a black bear walking across a lush green forest
MRNN	a couple of bears walking across a dirt road
D-ME+DMSM+MRNN	a black bear walking through a wooded area
k-NN	a black bear that is walking in the woods

D-ME+DMSM	a gray and white cat sitting on top of it
MRNN	a cat sitting in front of a mirror
D-ME+DMSM+MRNN	a close up of a cat looking at the camera
k-NN	a cat sitting on top of a wooden table

Table 2: Example generated captions.

3.2 Metrics

The quality of generated captions is measured automatically using BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). BLEU roughly measures the fraction of N-grams (up to 4 grams) that are in common between a hypothesis and one or more references, and penalizes short hypotheses by a brevity penalty term.[7] METEOR (Denkowski and Lavie, 2014) measures unigram precision and recall, extending exact word matches to include similar words based on WordNet synonyms and stemmed tokens. We also report the perplexity (PPLX) of studied detection-conditioned LMs. The PPLX is in many ways the natural measure of a statistical LM, but can be loosely correlated with BLEU (Auli et al., 2013).

3.3 Model Comparison

In Table 1, we summarize the generation performance of our different models. The discrete detection based models are prefixed with "D". Some example generated results are show in Table 2.

We see that the detection-conditioned LSTM LM produces much lower PPLX than the detection-conditioned ME LM, but its BLEU score is no better. The MRNN has the lowest PPLX, and highest BLEU among all LMs stud-

[7]We use the length of the reference that is closest to the length of the hypothesis to compute the brevity penalty.

Re-Ranking Features	BLEU	METEOR
D-ME †	23.6	22.8
+ DMSM †	25.7	23.6
+ MRNN	26.8	23.3
+ DMSM + MRNN	**27.3**	23.6

Table 3: Model performance on testval after re-ranking. †: previously reported and reconfirmed BLEU scores from (Fang et al., 2015). +DMSM had resulted in the highest score yet reported.

ied in our experiments. It significantly improves BLEU by 2.1 absolutely over the D-ME LM baseline. METEOR is similar across all three LM-based methods.

Perhaps most surprisingly, the k-nearest neighbor algorithm achieves a higher BLEU score than all other models. However, as we will demonstrate in Section 3.5, the generated captions perform significantly better than the nearest neighbor captions in terms of human quality judgements.

3.4 n-best Re-Ranking

In addition to comparing the ME-based and RNN-based LMs independently, we explore whether combining these models results in an additive improvement. To this end, we use the 500-best list from the D-ME and add a score for each hypothesis from the MRNN.[8] We then re-rank the hypotheses using MERT (Och, 2003). As in previous work (Fang et al., 2015), model weights were optimized to maximize BLEU score on the validation set. We further extend this combination approach to the D-ME model with DMSM scores included during re-ranking (Fang et al., 2015).

Results are show in Table 3. We find that combining the D-ME, DMSM, and MRNN achieves a 1.6 BLEU improvement over the D-ME+DMSM.

3.5 Human Evaluation

Because automatic metrics do not always correlate with human judgments (Callison-Burch et al., 2006; Hodosh et al., 2013), we also performed human evaluations using the same procedure as in Fang et al. (2015). Here, human judges were presented with an image, a system generated caption, and a human generated caption, and were asked which caption was "better".[9] For each condition, 5 judgments were obtained for 1000 images from the testval set.

[8]The MRNN does not produce a diverse n-best list.

[9]The captions were randomized and the users were not informed which was which.

Results are shown in Table 4. The D-ME+DMSM outperforms the MRNN by 5 percentage points for the "Better Or Equal to Human" judgment, despite both systems achieving the same BLEU score. The k-Nearest Neighbor system performs 1.4 percentage points worse than the MRNN, despite achieving a slightly higher BLEU score. Finally, the combined model does not outperform the D-ME+DMSM in terms of human judgments despite a 1.6 BLEU improvement.

Although we cannot pinpoint the exact reason for this mismatch between automated scores and human evaluation, a more detailed analysis of the difference between systems is performed in Sections 4 and 5.

	Human Judgements		
Approach	Better	Better or Equal	BLEU
D-ME+DMSM	7.8%	34.0%	25.7
MRNN	8.8%	29.0%	25.7
D-ME+DMSM+MRNN	5.7%	34.2%	**27.3**
k-Nearest Neighbor	5.5%	27.6%	26.0

Table 4: Results when comparing produced captions to those written by humans, as judged by humans. These are the percent of captions judged to be "better than" or "better than or equal to" a caption written by a human.

4 Language Analysis

Examples of common mistakes we observe on the testval set are shown in Table 5. The D-ME system has difficulty with anaphora, particularly within the phrase "on top of it", as shown in examples (1), (2), and (3). This is likely due to the fact that is maintains a local context window. In contrast, the MRNN approach tends to generate such anaphoric relationships correctly.

However, the D-ME LM maintains an explicit coverage state vector tracking which attributes have already been emitted. The MRNN *implicitly* maintains the full state using its recurrent layer, which sometimes results in multiple emission mistakes, where the same attribute is emitted more than once. This is particularly evident when coordination ("and") is present (examples (4) and (5)).

4.1 Repeated Captions

All of our models produce a large number of captions seen in the training and repeated for different images in the test set, as shown in Table 6 (also observed by Vinyals et al. (2014) for their LSTM-based model). There are at least two potential causes for this repetition.

	D-ME+DMSM	MRNN
1	a slice of pizza sitting on top of it	a bed with a red blanket on top of it
2	a black and white bird perched on top of it	a birthday cake with candles on top of it
3	a little boy that is brushing his teeth with a toothbrush in her mouth	a little girl brushing her teeth with a toothbrush
4	a large bed sitting in a bedroom	a bedroom with a bed and a bed
5	a man wearing a bow tie	a man wearing a tie and a tie

Table 5: Example errors in the two basic approaches.

System	Unique Captions	Seen In Training
Human	99.4%	4.8%
D-ME+DMSM	47.0%	30.0%
MRNN	33.1%	60.3%
D-ME+DMSM+MRNN	28.5%	61.3%
k-Nearest Neighbor	36.6%	100%

Table 6: Percentage unique (Unique Captions) and novel (Seen In Training) captions for testval images. For example, 28.5% unique means 5,776 unique strings were generated for all 20,244 images.

First, the systems often produce generic captions such as "a close up of a plate of food", which may be applied to many publicly available images. This may suggest a deeper issue in the training and evaluation of our models, which warrants more discussion in future work. Second, although the COCO dataset and evaluation server[10] has encouraged rapid progress in image captioning, there may be a lack of diversity in the data. We also note that although caption duplication is an issue in all systems, it is a greater issue in the MRNN than the D-ME+DMSM.

5 Image Diversity

The strong performance of the k-nearest neighbor algorithm and the large number of repeated captions produced by the systems here suggest a lack of diversity in the training and test data.[11]

We believe that one reason to work on image captioning is to be able to caption *compositionally* novel images, where the individual *components* of the image may be seen in the training, but the entire composition is often not.

In order to evaluate results for only compositionally novel images, we bin the test images based on visual overlap with the training data. For each test image, we compute the `fc7` cosine similarity with each training image, and the mean value of the 50 closest images. We then compute BLEU on the 20% least overlapping and 20% most overlapping subsets.

Condition	Train/Test Visual Overlap BLEU		
	Whole Set	20% Least	20% Most
D-ME+DMSM	25.7	20.9	29.9
MRNN	25.7	18.8	32.0
D-ME+DMSM+MRNN	27.3	21.7	32.0
k-Nearest Neighbor	26.0	18.4	33.2

Table 7: Performance for different portions of testval, based on visual overlap with the training.

Results are shown in Table 7. The D-ME+DMSM outperforms the k-nearest neighbor approach by 2.5 BLEU on the "20% Least" set, even though performance on the whole set is comparable. Additionally, the D-ME+DMSM outperforms the MRNN by 2.1 BLEU on the "20% Least" set, but performs 2.1 BLEU *worse* on the "20% Most" set. This is evidence that D-ME+DMSM generalizes better on novel images than the MRNN; this is further supported by the relatively low percentage of captions it generates seen in the training data (Table 6) while still achieving reasonable captioning performance. We hypothesize that these are the main reasons for the strong human evaluation results of the D-ME+DMSM shown in Section 3.5.

6 Conclusion

We have shown that a gated RNN conditioned directly on CNN activations (an MRNN) achieves better BLEU performance than an ME LM or LSTM conditioned on a set of discrete activations; and a similar BLEU performance to an ME LM combined with a DMSM. However, the ME LM + DMSM method significantly outperforms the MRNN in terms of human quality judgments. We hypothesize that this is partially due to the lack of novelty in the captions produced by the MRNN. In fact, a k-nearest neighbor retrieval algorithm introduced in this paper performs similarly to the MRNN in terms of both automatic metrics and human judgements.

When we use the MRNN system alongside the DMSM to provide additional scores in MERT reranking of the n-best produced by the image-conditioned ME LM, we advance by 1.6 BLEU points on the best previously published results on the COCO dataset. Unfortunately, this improvement in BLEU does not translate to improved human quality judgments.

[10] http://mscoco.org/dataset/

[11] This is partially an artifact of the manner in which the Microsoft COCO data set was constructed, since each image was chosen to be in one of 80 pre-defined object categories.

References

Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig. 2013. Joint language and translation modeling with recurrent neural networks. In *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*, pages 1044–1054.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *EACL*, volume 6, pages 249–256.

Xinlei Chen and C. Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proc. Conf. Comput. Vision and Pattern Recognition (CVPR)*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: language specific translation evaluation for any target language. In *Proc. EACL 2014 Workshop Statistical Machine Translation*.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2014. Long-term recurrent convolutional networks for visual recognition and description. *arXiv:1411.4389 [cs.CV]*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. Conf. Comput. Vision and Pattern Recognition (CVPR)*.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollá, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captionons to visual concepts and back. In *Proc. Conf. Comput. Vision and Pattern Recognition (CVPR)*.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: generating sentences from images. In *Proc. European Conf. Comput. Vision (ECCV)*, pages 15–29.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: data models and evaluation metrics. *J. Artificial Intell. Research*, pages 853–899.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. Conf. Comput. Vision and Pattern Recognition (CVPR)*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. 2014. Multimodal neural language models. In *Proc. Int. Conf. Machine Learning (ICML)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. *arXiv:1405.0312 [cs.CV]*.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-RNN). In *Proc. Int. Conf. Learning Representations (ICLR)*.

Rebecca Mason and Eugene Charniak. 2014. Domain-specific image captioning. In *CoNLL*.

Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *SLT*, pages 234–239.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*, ACL '03.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing images using 1 million captioned photogrphs. In *Proc. Annu. Conf. Neural Inform. Process. Syst. (NIPS)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. Assoc. for Computational Linguistics (ACL)*, pages 311–318.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: a neural image caption generator. In *Proc. Conf. Comput. Vision and Pattern Recognition (CVPR)*.