From Captions to Visual Concepts and Back

Hao Fang* Saurabh Gupta* Forrest Iandola* Rupesh K. Srivastava*
Li Deng Piotr Dollár† Jianfeng Gao Xiaodong He
Margaret Mitchell John C. Platt‡ C. Lawrence Zitnick Geoffrey Zweig

Microsoft Research

Abstract

This paper presents a novel approach for automatically generating image descriptions: visual detectors, language models, and multimodal similarity models learnt directly from a dataset of image captions. We use multiple instance learning to train visual detectors for words that commonly occur in captions, including many different parts of speech such as nouns, verbs, and adjectives. The word detector outputs serve as conditional inputs to a maximum-entropy language model. The language model learns from a set of over 400,000 image descriptions to capture the statistics of word usage. We capture global semantics by re-ranking caption candidates using sentence-level features and a deep multimodal similarity model. Our system is state-of-the-art on the official Microsoft COCO benchmark, producing a BLEU-4 score of 29.1%. When human judges compare the system captions to ones written by other people on our held-out test set, the system captions have equal or better quality 34% of the time.

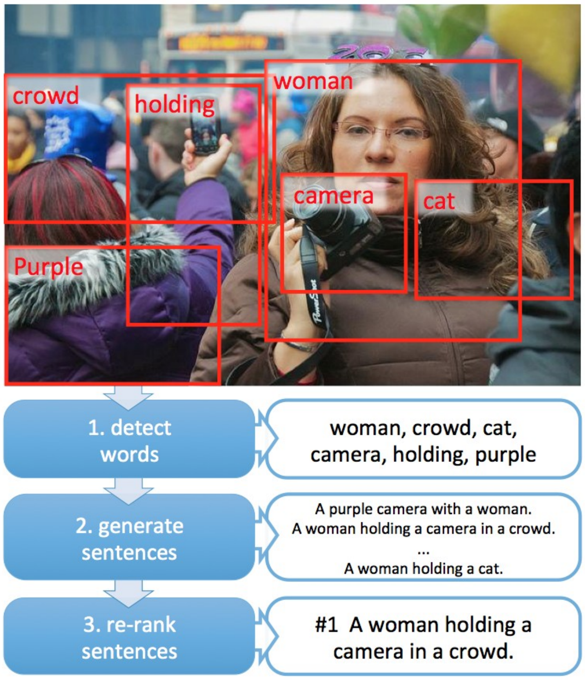

crowd
holding
woman
camera
cat
Purple
1. detect words
woman, crowd, cat, camera, holding, purple
2. generate sentences
A purple camera with a woman.
A woman holding a camera in a crowd.
...
A woman holding a cat.
3. re-rank sentences
#1 A woman holding a camera in a crowd.

Figure 1. An illustrative example of our pipeline.

1. Introduction

When does a machine "understand" an image? One definition is when it can generate a novel caption that summarizes the salient content within an image. This content may include objects that are present, their attributes, or their relations with each other. Determining the salient content requires not only knowing the contents of an image, but also deducing which aspects of the scene may be interesting or novel through commonsense knowledge [51, 5, 8].

This paper describes a novel approach for generating image captions from samples. We train our caption generator from a dataset of images and corresponding image descriptions. Previous approaches to generating image captions relied on object, attribute, and relation detectors learned from separate hand-labeled training data [47, 22].

The direct use of captions in training has three distinct advantages. First, captions only contain information that is inherently salient. For example, a dog detector trained from images with captions containing the word `dog` will be biased towards detecting dogs that are salient and not those that are in the background. Image descriptions also contain variety of word types, including nouns, verbs, and adjectives. As a result, we can learn detectors for a wide variety of concepts. While some concepts, such as `riding` or `beautiful`, may be difficult to learn in the abstract, these

*H. Fang, S. Gupta, F. Iandola and R. K. Srivastava contributed equally to this work while doing internships at Microsoft Research. Current affiliations are H. Fang: University of Washington; S. Gupta and F. Iandola: University of California at Berkeley; R. K. Srivastava: IDSIA, USI-SUPSI.

†P. Dollár is currently at Facebook AI Research.

‡J. Platt is currently at Google.

terms may be highly correlated to specific visual patterns (such as a person on a horse or mountains at sunset).

Second, training a language model (LM) on image captions captures commonsense knowledge about a scene. A language model can learn that a person is more likely to sit on a chair than to stand on it. This information disambiguates noisy visual detections.

Third, by learning a joint multimodal representation on images and their captions, we are able to measure the global similarity between images and text, and select the most suitable description for the image.

An overview of our approach is shown in Figure 1. First, we use weakly-supervised learning to create detectors for a set of words commonly found in image captions. Learning directly from image captions is difficult, because the system does not have access to supervisory signals, such as object bounding boxes, that are found in other data sets [11, 7]. Many words, e.g., `crowded` or `inside`, do not even have well-defined bounding boxes. To overcome this difficulty, we use three ideas. First, the system reasons with image sub-regions rather than with the full image. Next, we featurize each of these regions using rich convolutional neural network (CNN) features, fine-tuned on our training data [21, 42]. Finally, we map the features of each region to words likely to be contained in the caption. We train this map using multiple instance learning (MIL) [30, 49] which learns discriminative visual signature for each word.

Generating novel image descriptions from a bag of likely words requires an effective LM. In this paper, we view caption generation as an optimization problem. In this view, the core task is to take the set of word detection scores, and find the highest likelihood sentence that covers each word exactly once. We train a maximum entropy (ME) LM from a set of training image descriptions [2, 40]. This training captures commonsense knowledge about the world through language statistics [3]. An explicit search over word sequences is effective at finding high-likelihood sentences.

The final stage of the system (Figure 1) re-ranks a set of high-likelihood sentences by a linear weighting of sentence features. These weights are learned using Minimum Error Rate Training (MERT) [35]. In addition to several common sentence features, we introduce a new feature based on a Deep Multimodal Similarity Model (DMSM). The DMSM learns two neural networks that map images and text fragments to a common vector representation in which the similarity between sentences and images can be easily measured. As we demonstrate, the use of the DMSM significantly improves the selection of quality sentences.

To evaluate the quality of our automatic captions, we use three easily computable metrics and *better/worse/equal* comparisons by human subjects on Amazon's Mechanical Turk (AMT). The evaluation was performed on the challenging Microsoft COCO dataset [28, 4] containing complex images with multiple objects. Each of the 82,783 training images has 5 human annotated captions. For measuring the quality of our sentences we use the popular BLEU [37], METEOR [1] and perplexity (PPLX) metrics. Surprisingly, we find our generated captions outperform humans based on the BLEU metric; and this effect holds when evaluated on unseen test data from the COCO dataset evaluation server, reaching 29.1% BLEU-4 vs. 21.7% for humans. Human evaluation on our held-out test set has our captions judged to be of the same quality or better than humans 34% of the time. We also compare to previous work on the PASCAL sentence dataset [38], and show marked improvements over previous work. Our results demonstrate the utility of training both visual detectors and LMs directly on image captions, as well as using a global multimodal semantic model for re-ranking the caption candidates.

2. Related Work

There are two well-studied approaches to automatic image captioning: retrieval of existing human-written captions, and generation of novel captions. Recent retrieval-based approaches have used neural networks to map images and text into a common vector representation [43]. Other retrieval based methods use similarity metrics that take predefined image features [15, 36]. Farhadi et al. [12] represent both images and text as linguistically-motivated semantic triples, and compute similarity in that space. A similar fine-grained analysis of sentences and images has been done for retrieval in the context of neural networks [19].

Retrieval-based methods always return well-formed human-written captions, but these captions may not be able to describe new combinations of objects or novel scenes. This limitation has motivated a large body of work on generative approaches, where the image is first analyzed and objects are detected, and then a novel caption is generated. Previous work utilizes syntactic and semantic constraints in the generation process [32, 48, 26, 23, 22, 47], and we compare against prior state of the art in this line of work. We focus on the Midge system [32], which combines syntactic structures using maximum likelihood estimation to generate novel sentences; and compare qualitatively against the Baby Talk system [22], which generates descriptions by filling sentence template slots with words selected from a conditional random field that predicts the most likely image labeling. Both of these previous systems use the same set of test sentences, making direct comparison possible.

Recently, researchers explored purely statistical approaches to guiding language models using images. Kiros et al. [20] use a log-bilinear model with bias features derived from the image to model text conditioned on the image. Also related are several contemporaneous papers [29, 45, 6, 18, 9, 46, 25]. Among these, a common theme [29, 45, 6, 18] has been to utilize a recurrent neural network

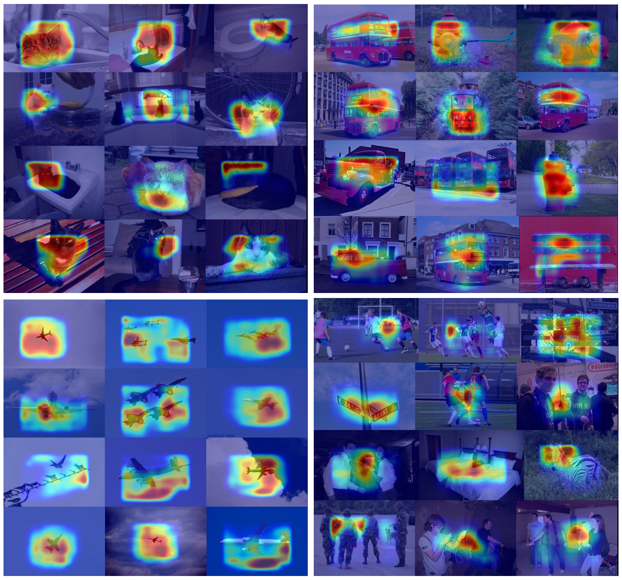


Figure 2. Multiple Instance Learning detections for `cat`, `red`, `flying` and `two` (left to right, top to bottom). View in color.

for generating images captions by conditioning its output on image features extracted by a convolutional neural network. More recently, Donahue et al. [9] also applied a similar model to video description. Lebret et al. [25] have investigated the use of a phrase-based model for generating captions, while Xu et al. [46] have proposed a model based on visual attention.

Unlike these approaches, in this work we detect words by applying a CNN to image regions [13] and integrating the information with MIL [49]. We minimize *a priori* assumptions about how sentences should be structured by training directly from captions. Finally, in contrast to [20, 29], we formulate the problem of generation as an optimization problem and search for the most likely sentence [40].

3. Word Detection

The first step in our caption generation pipeline detects a set of words that are likely to be part of the image's description. These words may belong to any part of speech, including nouns, verbs, and adjectives. We determine our vocabulary $\mathcal{V}$ using the 1000 most common words in the training captions, which cover over 92% of the word occurrences in the training data (available on project webpage [1]).

3.1. Training Word Detectors

Given a vocabulary of words, our next goal is to detect the words from images. We cannot use standard supervised learning techniques for learning detectors, since we do not know the image bounding boxes corresponding to the words. In fact, many words relate to concepts for which bounding boxes may not be easily defined, such as `open` or `beautiful`. One possible approach is to use image classifiers that take as input the entire image. As we show in Section 6, this leads to worse performance since many words or concepts only apply to image sub-regions. Instead, we learn our detectors using the weakly-supervised approach of Multiple Instance Learning (MIL) [30, 49].

For each word $w \in \mathcal{V}$, MIL takes as input sets of "positive" and "negative" bags of bounding boxes, where each bag corresponds to one image i. A bag b_i is said to be positive if word w is in image i's description, and negative otherwise. Intuitively, MIL performs training by iteratively selecting instances within the positive bags, followed by retraining the detector using the updated positive labels.

We use a noisy-OR version of MIL [49], where the probability of bag b_i containing word w is calculated from the probabilities of individual instances in the bag:

$$p_i^w \longrightarrow \quad 1 - \prod_{j \in b_i} \left(1 - p_{ij}^w\right) \qquad (1)$$

where p_{ij}^w is the probability that a given image region j in image i corresponds to word w. We compute p_{ij}^w using a multi-layered architecture [21, 42][2], by computing a logistic function on top of the `fc7` layer (this can be expressed as a fully connected `fc8` layer followed by a sigmoid layer):

$$\frac{1}{1 + \exp\left(-(\mathbf{v_w^t} \phi(b_{ij}) + u_w)\right)}, \qquad (2)$$

where $\phi(b_{ij})$ is the `fc7` representation for image region j in image i, and $\mathbf{v_w}$, u_w are the weights and bias associated with word w.

第i张图片包含w的概率=1-第i张图片中所有box都不含w的概率。

box包含w的概率通过一个线性模型计算：输入为box的特征。每个w都有自己的一套参数。

We express the fully connected layers (`fc6`, `fc7`, `fc8`) of these networks as convolutions to obtain a fully convolutional network. When this fully convolutional network is run over the image, we obtain a coarse spatial response map. Each location in this response map corresponds to the response obtained by applying the original CNN to overlapping shifted regions of the input image (thereby effectively scanning different locations in the image for possible objects). We up-sample the image to make the longer side to be 565 pixels which gives us a 12 × 12 response map at `fc8` for both [21, 42] and corresponds to sliding a 224×224 bounding box in the up-sampled image with a stride of 32. The noisy-OR version of MIL is then implemented on top of this response map to generate a single probability p_i^w for each word for each image. We use a cross entropy loss and optimize the CNN end-to-end for this task with stochastic gradient descent. We use one image in each batch and train for 3 epochs. For initialization, we use the network pretrained on ImageNet [7].

[1] http://research.microsoft.com/image_captioning

[2] We denote the CNN from [21] as AlexNet and the 16-layer CNN from [42] as VGG for subsequent discussion. We use the code base and models available from the Caffe Model Zoo https://github.com/BVLC/caffe/wiki/Model-Zoo [17].

3.2. Generating Word Scores for a Test Image

Given a novel test image i, we up-sample and forward propagate the image through the CNN to obtain p_i^w as described above. We do this for all words w in the vocabulary $\mathcal{V}$. Note that all the word detectors have been trained independently and hence their outputs need to be calibrated. To calibrate the output of different detectors, we use the image level likelihood p_i^w to compute precision on a held-out subset of the training data [14]. We threshold this precision value at a global threshold τ, and output all words $\tilde{\mathcal{V}}$ with a precision of τ or higher along with the image level probability p_i^w, and raw score $\max_j p_{ij}^w$.

Figure 2 shows some sample MIL detections. For each image, we visualize the spatial response map p_{ij}^w. Note that the method has not used any bounding box annotations for training, but is still able to reliably localize objects and also associate image regions with more abstract concepts.

4. Language Generation

We cast the generation process as a search for the likeliest sentence conditioned on the set of visually detected words. The language model is at the heart of this process because it defines the probability distribution over word sequences. Note that despite being a statistical model, the LM can encode very meaningful information, for instance that `running` is more likely to follow `horse` than `talking`. This information can help identify false word detections and encodes a form of commonsense knowledge.

4.1. Statistical Model

To generate candidate captions for an image, we use a maximum entropy (ME) LM conditioned on the set of visually detected words. The ME LM estimates the probability of a word w_l conditioned on the preceding words $w_1, w_2, \cdots, w_{l-1}$, as well as the set of words with high likelihood detections $\tilde{\mathcal{V}}_l \subset \tilde{\mathcal{V}}$ that have yet to be mentioned in the sentence. The motivation of conditioning on the unused words is to encourage all the words to be used, while avoiding repetitions. The top 15 most frequent closed-class words[3] are removed from the set $\tilde{\mathcal{V}}$ since they are detected in nearly every image (and are trivially generated by the LM). It should be noted that the detected words are usually somewhat noisy. Thus, when the end of sentence token is being predicted, the set of remaining words may still contain some words with a high confidence of detection.

Following the definition of an ME LM [2], the word probability conditioned on preceding words and remaining objects can be written as:

$$\begin{aligned}&\Pr(w_l = \bar{w}_l | \bar{w}_{l-1}, \cdots, \bar{w}_1, \text{<}s\text{>}, \tilde{\mathcal{V}}_{l-1}) = \\ &\frac{\exp\left[\sum_{k=1}^{K} \lambda_k f_k(\bar{w}_l, \bar{w}_{l-1}, \cdots, \bar{w}_1, \text{<}s\text{>}, \tilde{\mathcal{V}}_{l-1})\right]}{\sum_{v \in \mathcal{V} \cup \text{</}s\text{>}} \exp\left[\sum_{k=1}^{K} \lambda_k f_k(v, \bar{w}_{l-1}, \cdots, \bar{w}_1, \text{<}s\text{>}, \tilde{\mathcal{V}}_{l-1})\right]}\end{aligned} \tag{3}$$

where <s> denotes the start-of-sentence token, $\bar{w}_j \in \mathcal{V} \cup$ </s>, and $f_k(w_l, \cdots, w_1, \tilde{\mathcal{V}}_{l-1})$ and λ_k respectively denote the k-th max-entropy feature and its weight. The basic discrete ME features we use are summarized in Table 1. These features form our "baseline" system. It has proven effective to extend this with a "score" feature, which evaluates to the log-likelihood of a word according to the corresponding visual detector. We have also experimented with distant bigram features [24] and continuous space log-bilinear features [33, 34], but while these improved PPLX significantly, they did not improve BLEU, METEOR or human preference, and space restrictions preclude further discussion.

To train the ME LM, the objective function is the log-likelihood of the captions conditioned on the corresponding set of detected objects, i.e.:

$$L(\Lambda) = \sum_{s=1}^{S} \sum_{l=1}^{\#(s)} \log \Pr(\bar{w}_l^{(s)} | \bar{w}_{l-1}^{(s)}, \cdots, \bar{w}_1^{(s)}, \text{<}s\text{>}, \tilde{\mathcal{V}}_{l-1}^{(s)}) \tag{4}$$

where the superscript (s) denotes the index of sentences in the training data, and $\#(s)$ denotes the length of the sentence. The noise contrastive estimation (NCE) technique is used to accelerate the training by avoiding the calculation of the exact denominator in (3) [34]. In the generation process, we use the unnormalized NCE likelihood estimates, which are far more efficient than the exact likelihoods, and produce very similar outputs. However, all PPLX numbers we report are computed with exhaustive normalization. The ME features are implemented in a hash table as in [31]. In our experiments, we use N-gram features up to 4-gram and 15 contrastive samples in NCE training.

4.2. Generation Process

During generation, we perform a left-to-right beam search similar to the one used in [39]. This maintains a stack of length l partial hypotheses. At each step in the search, every path on the stack is extended with a set of likely words, and the resulting length $l+1$ paths are stored. The top k length $l+1$ paths are retained and the others pruned away.

We define the possible extensions to be the end of sentence token </s>, the 100 most frequent words, the set of attribute words that remain to be mentioned, and all the words in the training data that have been observed to follow the last word in the hypothesis. Pruning is based on the likelihood of the partial path. When </s> is generated, the full path to </s> is removed from the stack and set aside as a completed sentence. The process continues until a maximum sentence length L is reached.

[3] The top 15 frequent closed-class words are `a`, `on`, `of`, `the`, `in`, `with`, `and`, `is`, `to`, `an`, `at`, `are`, `next`, `that` and `it`.

Table 1. Features used in the maximum entropy language model.

Feature	Type	Definition	Description
Attribute	0/1	$\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	Predicted word is in the attribute set, i.e. has been visually detected and not yet used.
N-gram+	0/1	$\bar{w}_{l-N+1}, \cdots, \bar{w}_l = \kappa$ and $\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	N-gram ending in predicted word is κ and the predicted word is in the attribute set.
N-gram-	0/1	$\bar{w}_{l-N+1}, \cdots, \bar{w}_l = \kappa$ and $\bar{w}_l \notin \tilde{\mathcal{V}}_{l-1}$	N-gram ending in predicted word is κ and the predicted word is not in the attribute set.
End	0/1	$\bar{w}_l = \kappa$ and $\tilde{\mathcal{V}}_{l-1} = \emptyset$	The predicted word is κ and all attributes have been mentioned.
Score	$\mathbb{R}$	score($\bar{w}_l$) when $\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	The log-probability of the predicted word when it is in the attribute set.

Table 2. Features used by MERT.

1. The log-likelihood of the sequence.
2. The length of the sequence.
3. The log-probability per word of the sequence.
4. The logarithm of the sequence's rank in the log-likelihood.
5. 11 binary features indicating whether the number of mentioned objects is x ($x = 0, \ldots, 10$).
6. The DMSM score between the sequence and the image.

After obtaining the set of completed sentences $\mathcal{C}$, we form an M-best list as follows. Given a target number of T image attributes to be mentioned, the sequences in $\mathcal{C}$ covering at least T objects are added to the M-best list, sorted in descending order by the log-likelihood. If there are less than M sequences covering at least T objects found in $\mathcal{C}$, we reduce T by 1 until M sequences are found.

5. Sentence Re-Ranking

Our LM produces an M-best set of sentences. Our final stage uses MERT [35] to re-rank the M sentences. MERT uses a linear combination of features computed over an entire sentence, shown in Table 2. The MERT model is trained on the M-best lists for the validation set using the BLEU metric, and applied to the M-best lists for the test set. Finally, the best sequence after the re-ranking is selected as the caption of the image. Along with standard MERT features, we introduce a new multimodal semantic similarity model, discussed below.

5.1. Deep Multimodal Similarity Model

To model global similarity between images and text, we develop a Deep Multimodal Similarity Model (DMSM). The DMSM learns two neural networks that map images and text fragments to a common vector representation. We measure similarity between images and text by measuring cosine similarity between their corresponding vectors. This cosine similarity score is used by MERT to re-rank the sentences. The DMSM is closely related to the unimodal Deep Structured Semantic Model (DSSM) [16, 41], but extends it to the multimodal setting. The DSSM was initially proposed to model the semantic relevance between textual search queries and documents, and is extended in this work to replace the query vector in the original DSSM by the image vector computed from the deep convolutional network.

The DMSM consists of a pair of neural networks, one for mapping each input modality to a common semantic space, which are trained jointly. In training, the data consists of a set of image/caption pairs. The loss function minimized during training represents the negative log posterior probability of the caption given the corresponding image.

Image model: We map images to semantic vectors using the same CNN (AlexNet / VGG) as used for detecting words in Section 3. We first finetune the networks on the COCO dataset for the full image classification task of predicting the words occurring in the image caption. We then extract out the `fc7` representation from the finetuned network and stack three additional fully connected layers with *tanh* non-linearities on top of this representation to obtain a final representation of the same size as the last layer of the text model. We learn the parameters in these additional fully connected layers during DMSM training.

Text model: The text part of the DMSM maps text fragments to semantic vectors, in the same manner as in the original DSSM. In general, the text fragments can be a full caption. Following [16] we convert each word in the caption to a letter-trigram count vector, which uses the count distribution of context-dependent letters to represent a word. This representation has the advantage of reducing the size of the input layer while generalizing well to infrequent, unseen and incorrectly spelled words. Then following [41], this representation is forward propagated through a deep convolutional neural network to produce the semantic vector at the last layer.

Objective and training: We define the relevance R as the cosine similarity between an image or query (Q) and a text fragment or document (D) based on their representations y_Q and y_D obtained using the image and text models: $R(Q,D) = \text{cosine}(y_Q, y_D) = ({y_Q}^T y_D)/\|y_Q\|\|y_D\|$. For a given image-text pair, we can compute the posterior probability of the text being relevant to the image via:

$$P(D|Q) = \frac{\exp(\gamma R(Q,D))}{\Sigma_{D' \in \mathbb{D}} \exp(\gamma R(Q,D'))} \tag{5}$$

Here γ is a smoothing factor determined using the validation set, which is 10 in our experiments. $\mathbb{D}$ denotes the set of all candidate documents (captions) which should be compared to the query (image). We found that restricting $\mathbb{D}$ to one matching document D^+ and a fixed number N of randomly selected non-matching documents D^- worked reasonably well, although using noise-contrastive estimation could further improve results. Thus, for each image we select one relevant text fragment and N non-relevant fragments to compute the posterior probability. N is set to 50 in our experiments. During training, we adjust the model parameters Λ to minimize the negative log posterior probability that the relevant captions are matched to the images:

$$L(\Lambda) = -\log \prod_{(Q,D^+)} P(D^+|Q) \tag{6}$$

6. Experimental Results

We next describe the datasets used for testing, followed by an evaluation of our approach for word detection and experimental results on sentence generation.

6.1. Datasets

Most of our results are reported on the Microsoft COCO dataset [28, 4]. The dataset contains 82,783 training images and 40,504 validation images. The images create a challenging testbed for image captioning since most images contain multiple objects and significant contextual information. The COCO dataset provides 5 human-annotated captions per image. The test annotations are not available, so we split the validation set into validation and test sets[4].

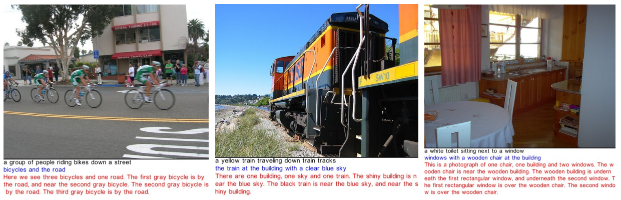
For experimental comparison with prior papers, we also report results on the PASCAL sentence dataset [38], which contains 1000 images from the 2008 VOC Challenge [11], with 5 human captions each.

6.2. Word Detection

To gain insight into our weakly-supervised approach for word detection using MIL, we measure its accuracy on the word classification task: If a word is used in at least one ground truth caption, it is included as a positive instance. Note that this is a challenging task, since conceptually similar words are classified separately; for example, the words cat/cats/kitten, or run/ran/running all correspond to different classes. Attempts at adding further supervision, e.g., in the form of lemmas, did not result in significant gains.

Average Precision (AP) and Precision at Human Recall (PHR) [4] results for different parts of speech are shown in Table 3. We report two baselines. The first (Chance) is the result of randomly classifying each word. The second (Classification) is the result of a whole image classifier which uses features from AlexNet or VGG CNN [21, 42]. These features were fine-tuned for this word classification task using a logistic regression loss.

As shown in Table 3, the MIL NOR approach improves over both baselines for all parts of speech, demonstrating that better localization can help predict words. In fact, we observe the largest improvement for nouns and adjectives, which often correspond to concrete objects in an image subregion. Results for both classification and MIL NOR are lower for parts of speech that may be less visually informative and difficult to detect, such as adjectives (e.g., `few`, which has an AP of 2.5), pronouns (e.g., `himself`, with an AP of 5.7), and prepositions (e.g., `before`, with an AP of 1.0). In comparison words with high AP scores are typically either visually informative (`red`: AP 66.4, `her`: AP 45.6) or associated with specific objects (`polar`: AP 94.6, `stuffed`: AP 74.2). Qualitative results demonstrating word localization are shown in Figures 2 and 3.



a group of people riding bikes down a street
bicycles and the road
Here we see three bicycles and one road. The first gray bicycle is by the road, and near the second gray bicycle. The second gray bicycle is by the road. The third gray bicycle is by the road.

a yellow train traveling down train tracks
the train at the building with a clear blue sky
There are one building, one sky and one train. The shiny building is near the blue sky. The black train is near the blue sky, and near the shiny building.

a white toilet sitting next to a window
windows with a wooden chair at the building
This is a photograph of one chair, one building and two windows. The wooden chair is near the wooden building. The wooden building is underneath the first rectangular window, and underneath the second window. The first rectangular window is over the wooden chair. The second window is over the wooden chair.

Figure 4. Qualitative results for images on the PASCAL sentence dataset. Captions using our approach (black), Midge [32] (blue) and Baby Talk [22] (red) are shown.

6.3. Caption Generation

We next describe our caption generation results, beginning with a short discussion of evaluation metrics.

Metrics: The sentence generation process is measured using both automatic metrics and human studies. We use three different automatic metrics: PPLX, BLEU [37], and METEOR [1]. PPLX (perplexity) measures the uncertainty of the language model, corresponding to how many bits on average would be needed to encode each word given the language model. A lower PPLX indicates a better score. BLEU [37] is widely used in machine translation and measures the fraction of N-grams (up to 4-gram) that are in common between a hypothesis and a reference or set of references; here we compare against 4 randomly selected references. METEOR [1] measures unigram precision and recall, extending exact word matches to include similar words based on WordNet synonyms and stemmed tokens. We additionally report performance on the metrics made available from the MSCOCO captioning challenge,[5] which includes scores for BLEU-1 through BLEU-4, METEOR, CIDEr [44], and ROUGE-L [27].

All of these automatic metrics are known to only roughly correlate with human judgment [10]. We therefore include human evaluation to further explore the quality of our models. Each task presents a human (Mechanical Turk worker) with an image and two captions: one is automatically generated, and the other is a human caption. The human is asked to select which caption better describes the image, or to choose a "same" option when they are of equal quality. In each experiment, 250 humans were asked to compare

[4] We split the COCO train/val set ito 82,729 train/20243 val/20244 test. Unless otherwise noted, test results are reported on the 20444 images from the validation set.

[5] http://mscoco.org/dataset/#cap2015

Table 3. Average precision (AP) and Precision at Human Recall (PHR) [4] for words with different parts of speech (NN: Nouns, VB: Verbs, JJ: Adjectives, DT: Determiners, PRP: Pronouns, IN: Prepositions). Results are shown using a chance classifier, full image classification, and Noisy OR multiple instance learning with AlexNet [21] and VGG [42] CNNs.

	Average Precision								Precision at Human Recall							
	NN	VB	JJ	DT	PRP	IN	Others	All	NN	VB	JJ	DT	PRP	IN	Others	All
Count	616	176	119	10	11	38	30	1000								
Chance	2.0	2.3	2.5	23.6	4.7	11.9	7.7	2.9								
Classification (AlexNet)	32.4	16.7	20.7	31.6	16.8	21.4	15.6	27.1	39.0	27.7	37.0	37.3	26.2	31.5	25.0	35.9
Classification (VGG)	37.0	19.4	22.5	32.9	19.4	22.5	16.9	30.8	45.3	31.0	37.1	40.2	29.6	33.9	25.5	40.6
MIL (AlexNet)	36.9	18.0	22.9	31.7	16.8	21.4	15.2	30.4	46.0	29.4	40.1	37.9	25.9	31.5	21.6	40.8
MIL (VGG)	41.4	20.7	24.9	32.4	19.1	22.8	16.3	34.0	51.6	33.3	44.3	39.2	29.4	34.3	23.9	45.7
Human Agreement									63.8	35.0	35.9	43.1	32.5	34.3	31.6	52.8

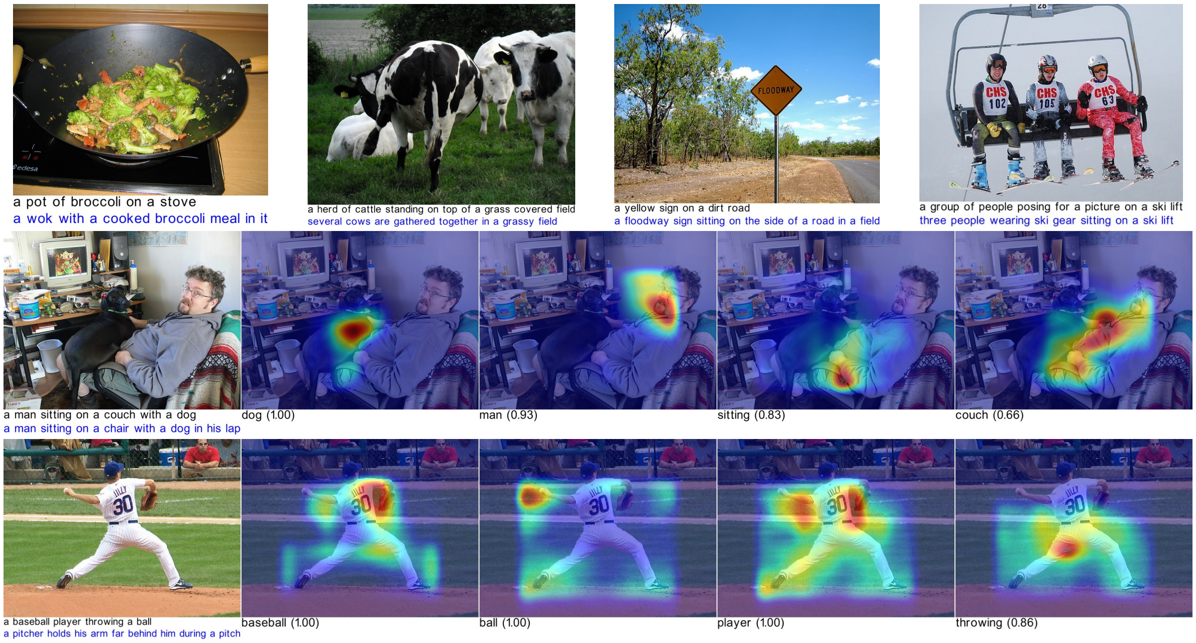

a pot of broccoli on a stove
a wok with a cooked broccoli meal in it
a herd of cattle standing on top of a grass covered field
several cows are gathered together in a grassy field
a yellow sign on a dirt road
a floodway sign sitting on the side of a road in a field
a group of people posing for a picture on a ski lift
three people wearing ski gear sitting on a ski lift
a man sitting on a couch with a dog
a man sitting on a chair with a dog in his lap
dog (1.00)
man (0.93)
sitting (0.83)
couch (0.66)
a baseball player throwing a ball
a pitcher holds his arm far behind him during a pitch
baseball (1.00)
ball (1.00)
player (1.00)
throwing (0.86)

Figure 3. Qualitative results for several randomly chosen images on the Microsoft COCO dataset, with our generated caption (black) and a human caption (blue) for each image. In the bottom two rows we show localizations for the words used in the sentences. More examples can be found on the project website[1].

20 caption pairs each, and 5 humans judged each caption pair. We used Crowdflower, which automatically filters out spammers. The ordering of the captions was randomized to avoid bias, and we included four check-cases where the answer was known and obvious; workers who missed any of these were excluded. The final judgment is the majority vote of the judgment of the 5 humans. In ties, one-half of a count is distributed to the two best answers. We also compute errors bars on the human results by taking 1000 bootstrap resamples of the majority vote outcome (with ties), then reporting the difference between the mean and the 5th or 95th percentile (whichever is farther from the mean).

Generation results: Table 4 summarizes our results on the Microsoft COCO dataset. We provide several baselines for experimental comparison, including two baselines that measure the complexity of the dataset: Unconditioned, which generates sentences by sampling an N-gram LM without knowledge of the visual word detectors; and Shuffled Human, which randomly picks another human generated caption from another image. Both the BLEU and METEOR scores are very low for these approaches, demonstrating the variation and complexity of the Microsoft COCO dataset.

We provide results on seven variants of our end-to-end approach: Baseline is based on visual features from AlexNet and uses the ME LM with all the discrete features as described in Table 1. Baseline+Score adds the feature for the word detector score into the ME LM. Both of these versions use the same set of sentence features (excluding the DMSM score) described in Section 5 when re-ranking the captions using MERT. Baseline+Score+DMSM uses the same ME LM as Baseline+Score, but adds the DMSM score as a feature for re-ranking. Baseline+Score+DMSM+ft adds finetuning. VGG+Score+ft and VGG+Score+DMSM+ft are analogous to Baseline+Score and Baseline+Score+DMSM but use

Table 4. Caption generation performance for seven variants of our system on the Microsoft COCO dataset. We report performance on our held out test set (half of the validation set). We report Perplexity (PPLX), BLEU and METEOR, using 4 randomly selected caption references. Results from human studies of subjective performance are also shown, with error bars in parentheses. Our final System "VGG+Score+DMSM+ft" is "same or better" than human 34% of the time.

System	PPLX	BLEU	METEOR	$\approx$human	$>$human	$\geq$human
1. Unconditioned	24.1	1.2%	6.8%			
2. Shuffled Human	–	1.7%	7.3%			
3. Baseline	20.9	16.9%	18.9%	9.9% (±1.5%)	2.4% (±0.8%)	12.3% (±1.6%)
4. Baseline+Score	20.2	20.1%	20.5%	16.9% (±2.0%)	3.9% (±1.0%)	20.8% (±2.2%)
5. Baseline+Score+DMSM	20.2	21.1%	20.7%	18.7% (±2.1%)	4.6% (±1.1%)	23.3% (±2.3%)
6. Baseline+Score+DMSM+ft	19.2	23.3%	22.2%	–	–	–
7. VGG+Score+ft	18.1	23.6%	22.8%	–	–	–
8. VGG+Score+DMSM+ft	18.1	25.7%	23.6%	26.2% (±2.1%)	7.8% (±1.3%)	**34.0%** (±2.5%)
Human-written captions	–	19.3%	24.1%			

Table 5. Official COCO evaluation server results on test set (40,775 images). First row show results using 5 reference captions, second row, 40 references. Human results reported in parentheses.

	CIDEr	BLEU-4	BLEU-1	ROUGE-L	METEOR
[5]	.912 (.854)	.291 (.217)	.695 (.663)	.519 (.484)	.247 (.252)
[40]	.925 (.910)	.567 (.471)	.880 (.880)	.662 (.626)	.331 (.335)

finetuned VGG features. Note: the AlexNet baselines without finetuning are from an early version of our system which used object proposals from [50] instead of dense scanning.

As shown in Table 4, the PPLX of the ME LM with and without the word detector score feature is roughly the same. But, BLEU and METEOR improve with addition of the word detector scores in the ME LM. Performance improves further with addition of the DMSM scores in re-ranking. Surprisingly, the BLEU scores are actually above those produced by human generated captions (25.69% vs. 19.32%). Improvements in performance using the DMSM scores with the VGG model are statistically significant as measured by 4-gram overlap and METEOR per-image (Wilcoxon signed-rank test, $p < .001$).

We also evaluated an approach (not shown) with whole-image classification rather than MIL. We found this approach to under-perform relative to MIL in the same setting (for example, using the VGG+Score+DMSM+ft setting, PPLX=18.9, BLEU=21.9%, METEOR=21.4%). This suggests that integrating information about words associated to image regions with MIL leads to improved performance over image classification alone.

The VGG+Score+DMSM approach produces captions that are judged to be of the same or better quality than human-written descriptions 34% of the time, which is a significant improvement over the Baseline results. Qualitative results are shown in Figure 3, and many more are available on the project website.

COCO evaluation server results: We further generated the captions for the images in the actual COCO test set consisting of 40,775 images (human captions for these images are not available publicly), and evaluated them on the COCO evaluation server. These results are summarized in Table 5. Our system gives a BLEU-4 score of 29.1%, and equals or surpasses human performance on 12 of the 14 metrics reported – the only system to do so. These results are also state-of-the-art on all 14 reported metrics among the four other results available publicly at the time of writing this paper. In particular, our system is the only one exceeding human CIDEr scores, which has been specifically proposed for evaluating image captioning systems [44].

To enable direct comparison with previous work on automatic captioning, we also test on the PASCAL sentence dataset [38], using the 847 images tested for both the Midge [32] and Baby Talk [22] systems. We show significantly improved results over the Midge [32] system, as measured by both BLEU and METEOR (2.0% vs. 17.6% BLEU and 9.2% vs. 19.2% METEOR).[6] To give a basic sense of the progress quickly being made in this field, Figure 4 shows output from the system on the same images.[7]

7. Conclusion

This paper presents a new system for generating novel captions from images. The system trains on images and corresponding captions, and learns to extract nouns, verbs, and adjectives from regions in the image. These detected words then guide a language model to generate text that reads well and includes the detected words. Finally, we use a global deep multimodal similarity model introduced in this paper to re-rank candidate captions

At the time of writing, our system is state-of-the-art on all 14 official metrics of the COCO image captioning task, and equal to or exceeding human performance on 12 out of the 14 official metrics. Our generated captions have been judged by humans (Mechanical Turk workers) to be equal to or better than human-written captions 34% of the time.

[6] Baby Talk generates long, multi-sentence captions, making comparison by BLEU/METEOR difficult; we thus exclude evaluation here.

[7] Images were selected visually, without viewing system captions.

References

[1] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005. 2, 6

[2] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 1996. 2, 4

[3] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010. 2

[4] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 6, 7

[5] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013. 1

[6] X. Chen and C. L. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. *CVPR*, 2015. 2

[7] 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, 3

[8] S. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014. 1

[9] 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. 2, 3

[10] D. Elliott and F. Keller. Comparing automatic evaluation measures for image description. In *ACL*, 2014. 6

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, June 2010. 2, 6

[12] 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. 2

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 3

[14] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *CVPR*, 2014. 4

[15] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 47:853–899, 2013. 2

[16] P. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*, 2013. 5

[17] 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. 3

[18] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *CVPR*, 2015. 2

[19] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv preprint arXiv:1406.5679*, 2014. 2

[20] R. Kiros, R. Zemel, and R. Salakhutdinov. Multimodal neural language models. In *NIPS Deep Learning Workshop*, 2013. 2, 3

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 3, 6, 7

[22] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 1, 2, 6, 8

[23] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In *ACL*, 2012. 2

[24] R. Lau, R. Rosenfeld, and S. Roukos. Trigger-based language models: A maximum entropy approach. In *ICASSP*, 1993. 4

[25] R. Lebret, P. O. Pinheiro, and R. Collobert. Phrase-based image captioning. *arXiv preprint arXiv:1502.03671*, 2015. 2, 3

[26] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *CoNLL*, 2011. 2

[27] C.-Y. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. 6

[28] 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. 2, 6

[29] 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. 2, 3

[30] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *NIPS*, 1998. 2, 3

[31] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Cernocky. Strategies for training large scale neural network language models. In *ASRU*, 2011. 4

[32] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III. Midge: Generating image descriptions from computer vision detections. In *EACL*, 2012. 2, 6, 8

[33] A. Mnih and G. Hinton. Three new graphical models for statistical language modelling. In *ICML*, 2007. 4

[34] A. Mnih and Y. W. Teh. A fast and simple algorithm for training neural probabilistic language models. In *ICML*, 2012. 4

[35] F. J. Och. Minimum error rate training in statistical machine translation. In *ACL*, 2003. 2, 5

[36] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 2

[37] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 2, 6

[38] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's mechanical turk. In *NAACL HLT Workshop Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010. 2, 6, 8

[39] A. Ratnaparkhi. Trainable methods for surface natural language generation. In *NAACL*, 2000. 4

[40] A. Ratnaparkhi. Trainable approaches to surface natural language generation and their application to conversational dialog systems. *Computer Speech & Language*, 16(3):435–455, 2002. 2, 3

[41] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. A latent semantic model with convolutional-pooling structure for information retrieval. In *CIKM*, 2014. 5

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2, 3, 6, 7

[43] 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

[44] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014. 6, 8

[45] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CVPR*, 2015. 2

[46] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015. 2, 3

[47] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011. 1, 2

[48] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2T: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010. 2

[49] C. Zhang, J. C. Platt, and P. A. Viola. Multiple instance boosting for object detection. In *NIPS*, 2005. 2, 3

[50] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 8

[51] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013. 1