# From Captions to Visual Concepts and Back

Zhang, Liqiang

July 6, 2018

## Abstract

*In this paper, the author propose a novel approach for automatically generating image descriptions: visual detectors, language models, and multimodal similarity models learn directly from a dataset of image captions. They 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. This team capture global semantics by re-ranking caption candidates using sentence-level features and a deep multimodal similarity model. The system that the author proposed is state-of-the-art on the official Microsoft COCO benchmark, producing a BlEU-4 score of 29.1%.*

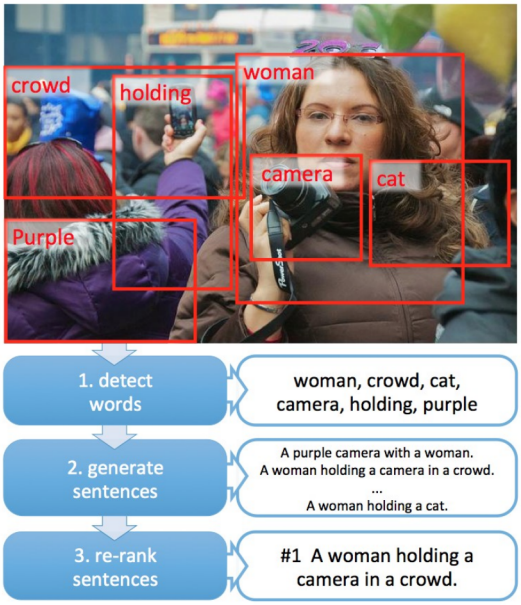


Figure 1. An illustrative example of our pipeline.

## 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 salinent 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 [1].

This paper describes a novel approach for generating image captions from samples. The author train their caption generator from a dataset of images and corresponding image descriptions. Previous approaches to generating image captions relied. Previous approaches to generating image captions relied on object, attribute, and relation detectors learned from separate hand-labeled training data [3].

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

An overview of the new approach is shown in Fig. 1 [2]. Firs, the author 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 hace access to supervisory signals, such as object bounding boxes, that are found in other data sets. Next, they featurize each of these regions using rich convolutional neural network (CNN) features, fine-tuned on our training data. Finally, this team map the features of each region to words likely to be contained in the caption. They train this map using multiple instance learning (MIL) [4] which learns discriminative visual signature

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

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

for each word.

## 2. Word Detection

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. This paper use a noisy-OR version of MIL, where the probability of bag $b_i$ containing word $w$ is calculated from the probabilities of individual instances in the bag as Eq. 1 [2]:

$$1 - \prod_{j \in b_i} (1 - p_{ij}^w) \tag{1}$$

where $\phi(b_{ij})$ is the probability that a given image region $j$ in image $i$ corresponds to word $w$. The author computing a logistic function on top of the fc7 layer as Eq. 3 [2]:

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

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 $w$.

## 3. Language Generation

Table 1 in [2] has summarized the basic discrete ME features. These features form the "baseline" system. It has proven effective to the log-likelihood of a word according to the corresponding visual detector. The author ha also experimented with distant bigram features and continuous space log-bilinear features, 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.:

$$L(\Lambda) = \sum_{s=1}^{S} \sum_{l=1}^{\#(s)} \log \Pr(\bar{w}_l^{(s)} | \bar{w}_{l-1}^{(s)}, \cdots, \bar{w}_1^{(s)}, <s>, \widetilde{\mathcal{V}}_{l-1}^{(s)} \tag{3}$$

where the superscript $(s)$ denotes the index of sentences in the training data, and $\#(s)$ denotes the length of the sentence. In the generation process, the author use the unnormalized NCE likelihood estimates which are far more efficient than the exact likelihoods, and produce very similar outputs.

## References


[1] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2014. 1

[2] H. Fang, J. C. Platt, C. L. Zitnick, G. Zweig, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, and J. Gao. From captions to visual concepts and back. In *CVPR*, 2015. 1, 2

[3] G. Kulkarni, V. Premraj, S. Dhar, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2013. 1

[4] O. Maron. A framework for multiple-instance learning. In *NIPS*, 1998. 1