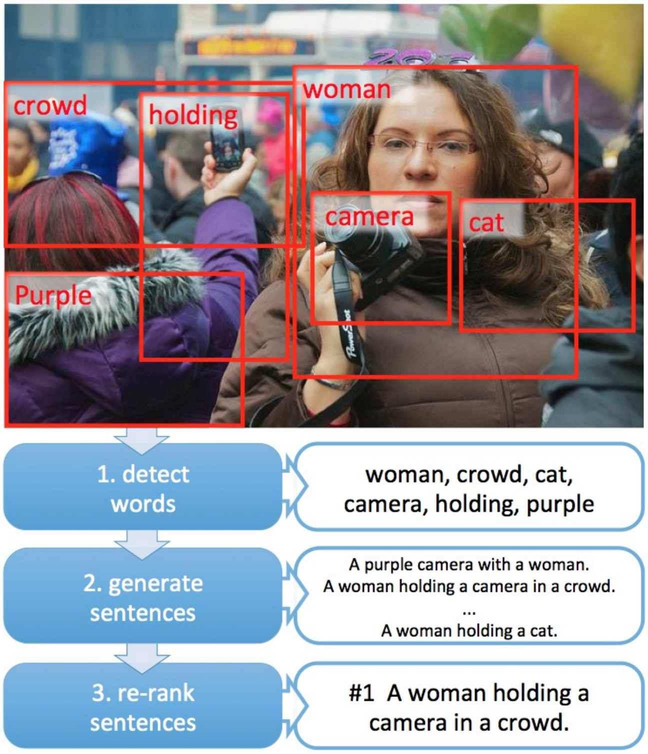
From Captions to Visual Concepts and Back

Hao Fang[1,*], Saurabh Gupta[2,*], Forrest Iandola[2,*], Rupesh K. Srivastava[3,*], Li Deng[4], Piotr Dollár[5]
Jianfeng Gao[4], Xiaodong He[4], Margaret Mitchell[4], John C. Platt[6], C. Lawrence Zitnick[4], Geoffrey Zweig[4]
[*]Equal Contribution, [1]U. of Washington, [2]UC Berkeley, [3]IDSIA, USI-SUPSI, [4]Microsoft Research, [5]Facebook AI Research, [6]Google



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.

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 relations between objects. 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.

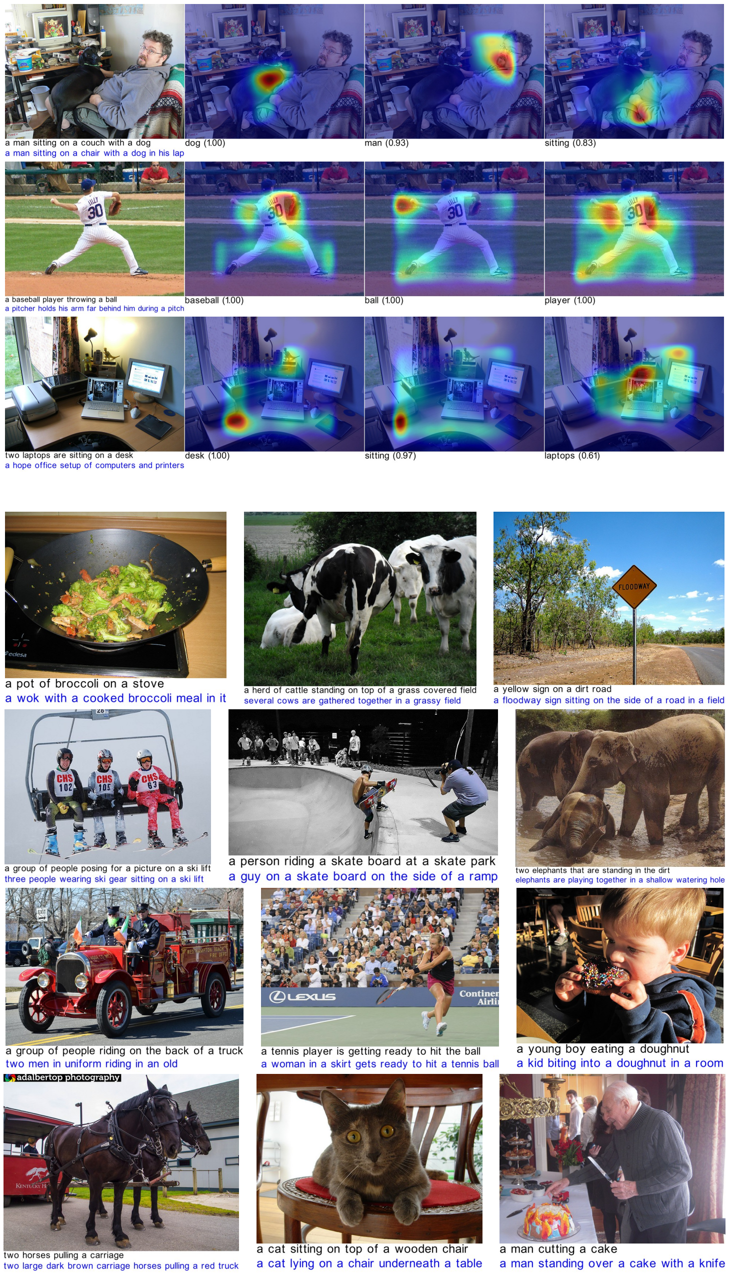
This paper describes a novel approach for generating image captions from samples. Previous approaches to generating image captions relied on object, attribute, and relation detectors learned from separate hand-labeled training data [22, 47]. Instead, we train our caption generator from a data set of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

Figure 1 shows the overview of our approach. We first use Multiple Instance Learning to learn detectors for words that occur in image captions. We then train a maximum entropy language model conditioned on the set of detected words and use it to generate candidate captions for the image. Finally, we learn to re-rank these generated sentences and output the highest scoring sentence.

Our evaluation was performed on the challenging MS COCO dataset [4, 28] containing complex images with multiple objects. Examples results are shown in Figure 2. We use the multiple metrics and *better/worse/equal* comparisons by human subjects on Amazon's Mechanical Turk to evaluate the quality of our automatic captions on a subset of the validation set.

We also submitted generated captions for the test set to the official COCO evaluation server (results in Fig. 3). Surprisingly, our generated captions match or outperform humans on 12 out of 14 official metrics. We also outperform other public results on all official metrics. When evaluated by human subjects, our captions were judged to be of the same or better quality than humans 34% of the time. Our results demonstrate the utility of training both visual detectors and LMs directly on image captions, as well as using a global semantic model for re-ranking the caption candidates.

This is an extended abstract. Full paper available at the Computer Vision Foundation page.


a man sitting on a couch with a dog
a man sitting on a chair with a dog in his lap

a baseball player throwing a ball
a pitcher holds his arm far behind him during a pitch

two laptops are sitting on a desk
a hope office setup of computers and printers

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 person riding a skate board at a skate park
a guy on a skate board on the side of a ramp
two elephants that are standing in the dirt
elephants are playing together in a shallow watering hole
a group of people riding on the back of a truck
two men in uniform riding in an old
a tennis player is getting ready to hit the ball
a woman in a skirt gets ready to hit a tennis ball
a young boy eating a doughnut
a kid biting into a doughnut in a room
two horses pulling a carriage
two large dark brown carriage horses pulling a red truck
a cat sitting on top of a wooden chair
a cat lying on a chair underneath a table
a man cutting a cake
a man standing over a cake with a knife

Figure 2: Qualitative results for several randomly chosen images on the MS COCO dataset, with our generated caption (black) and a human caption (blue) for each image. Top rows also show MIL based localization. More examples can be found on the project website: http://research.microsoft.com/image_captioning.

	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)

Figure 3: 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.