Fast, Diverse and Accurate Image Captioning Guided By Part-of-Speech

Aditya Deshpande[1]*, Jyoti Aneja[1]*, Liwei Wang[2], Alexander Schwing[1] and David Forsyth[1]
ardeshp2,janeja2,aschwing,daf@illinois.edu liweiwang@tencent.com
[1]University of Illinois at Urbana Champaign [2]Tencent AI Lab, Seattle

Abstract

*Image captioning is an ambiguous problem, with many suitable captions for an image. To address ambiguity, beam search is the de facto method for sampling multiple captions. However, beam search is computationally expensive and known to produce generic captions [8, 10]. To address this concern, some variational auto-encoder (VAE) [32] and generative adversarial net (GAN) [5, 25] based methods have been proposed. Though diverse, GAN and VAE are less accurate. In this paper, we first predict a meaningful summary of the image, then generate the caption based on that summary. We use part-of-speech as summaries, since our summary should drive caption generation. We achieve the trifecta: (1) High accuracy for the diverse captions as evaluated by standard captioning metrics and user studies; (2) Faster computation of diverse captions compared to beam search and diverse beam search [28]; and (3) High diversity as evaluated by counting novel sentences, distinct n-grams and mutual overlap (*i.e.*, mBleu-4) scores.*

1. Introduction

In this paper we show how to force an image captioning system to generate diverse captions by conditioning on different high-level summaries of the image. Our summaries are quantized part-of-speech (POS) tag sequences. Our system generates captions by (a) predicting different summaries from the image then (b) predicting captions conditioned on each summary. This approach leads to captions that are *accurate, quick to obtain*, and *diverse*. Our system is accurate, because it is able to steer a number of narrow beam searches to explore the space of caption sequences more efficiently. It is fast because each beam is narrow. And the captions are diverse, because depending on the summary (*i.e.*, part-of-speech) the system is forced to produce captions that contain (for example) more or less adjectives. This means we can avoid the tendency to produce minimal or generic captions that is common in systems that try to optimize likelihood without awareness of language priors (like part-of-speech).

Method	Fast	Diverse	Accurate
Beam search	×	×	✓
Diverse beam search [28]	×	×	✓
AG-CVAE [32]	✓	✓	×
Ours (POS)	✓	✓	✓

Table 1: We show that our part-of-speech (POS) based method achieves the trifecta of **high accuracy, fast computation** and **more diversity**. Beam search and diverse beam search are slow. They also produce captions with high mutual overlap and lower distinct n-grams than POS (see mBleu-4, div-1 and div-2 in Tab. 5). POS and AG-CVAE are fast, however POS does better on captioning metrics in Fig. 3 and is therefore more accurate.

A large body of literature has focused on developing predictive image captioning techniques, often using recurrent neural nets (RNN) [20, 29, 34, 12, 2]. More recently [3, 33], demonstrate predictive captioning with accuracy similar to RNNs while using convolutional networks. An essential feature of captioning is that it is ambiguous – many captions can describe the same image. This creates a problem, as image captioning programs trained to maximize some score may do so by producing strongly non-committal captions. It also creates an opportunity for research – how can one produce multiple, diverse captions that still properly describe the image? Our method offers a procedure to do so.

Tractable image captioning involves factoring the sequence model for the caption. Inference then requires beam search, which investigates a set of captions determined by local criteria to find the caption with highest posterior probability. Finding very good captions requires a wide beam, which is slow. Moreover, beam search is also known to generate generic captions that lack diversity [8, 10]. Variational auto-encoder (VAE) [32] and generative adversarial net (GAN) [5, 25, 14] formulations outperform beam search on diversity metrics. VAE and GAN-based methods sample latent vectors from some distribution, then generate captions conditioned on these samples. The latent variables have no exposed semantics, and captions tend not to score as well (on captioning metrics) as those produced by beam

* Denotes equal contribution.

search (*e.g.*, Tab. 1 of [25]).

This paper offers an alternative. First predict a meaningful summary of the image, then generate the caption based on that summary. For this to work, the summary needs to be able to drive language generation (for the caption generator), and must be predictable. We find quantized part of speech tag sequences to be very effective summaries. These sequences can clearly drive language generation (*e.g.*, forcing a captioner to produce adjectives in particular locations). More surprisingly, one can predict quantized tag sequences from images rather well, likely because such sequences do summarize the main action of the image. For example, compare *determiner-noun-verb* with *determiner-adjective-noun-verb-adjective-noun*. In the first case, something appears in the image, in the second, a subject with a noteworthy attribute is doing something to an object with a noteworthy attribute. Consequently, the two images appear quite different.

Contributions: We show that image captioning with POS tag sequences is fast, diverse and accurate (Tab. 1). Our POS methods sample captions faster and with more diversity than techniques based on beam search and its variant diverse beam search [28] (Tab. 5). Our diverse captions are more accurate than their counterparts produced by GANs [25] (Tab. 4) and VAEs [32] (Tab. 3, Fig. 3).

2. Related Work

In the following, we first review works that generate a single (or best-1) caption before discussing diverse image captioning methods which produce k different (or a set of best-k) captions.

2.1. Image Captioning

Most image captioning approaches [12, 29, 34] use a convolutional neural net pre-trained on classification [26] to represent image features. Image features are fed into a recurrent net (often based on long-short-term-memory (LSTM) units) to model the language word-by-word. These networks are trained with maximum likelihood. To obtain high performance on standard image captioning metrics, Yao *et al*. [35] use a network trained on COCO-attributes in addition to image features. Anderson *et al.* [2] develop an attention-based network architecture. Aneja *et al*. [3] change the language decoder from an LSTM-net to a convoluational network and show that they obtain more diversity. Similarly, Wang *et al*. [33] also use a convolutional language decoder. Since diversity is of interest to us, we use the convolutional language decoder similar to [3, 33]. We leave incorporation of techniques such as attribute vectors specific to the COCO dataset, and a sophisticated attention mechanism from [35, 2] for further performance gains to future work.

Apart from exploring different network architectures, some prior works focus on using different training losses. Reinforcement learning has been used in [19, 24, 17], to directly train for non-differentiable evaluation metrics such as BLEU, CIDEr and SPICE. In this paper, we use maximum likelihood training for our methods and baselines to ensure a fair comparison. Training our POS captioning network in a reinforcement learning setup can be investigated as part of future work.

Notable advances have been made in conditioning image captioning on semantic priors of objects by using object detectors [18, 30]. This conditioning is only limited to the objects (or nouns) in the caption and ignores the remainder, while our POS approach achieves coordination for the entire sentence.

2.2. Diverse Image Captioning

Four main techniques have been proposed to generate multiple captions and rank them to obtain a set of best-k captions.

Beam search. Beam search is the classical method to sample multiple solutions given sequence models for neural machine translation and image captioning. We compare to beam search on the same base captioning network as POS, but without part-of-speech conditioning. We find that though beam search is accurate, it is slow (Tab. 3) and lacks diversity (Tab. 5). Our base captioning network uses a convolutional neural net (CNN) [3] and is equivalent to the standard LSTM based captioning network of Karpathy *et al*. [12] in terms of accuracy.

Diverse beam search. Vijayakumar *et al*. [28] augment beam search with an additional diversity function to generate diverse outputs. They propose a hamming diversity function that penalizes expanding a beam with the same word used in an earlier beam. In our results, we compare to this diverse beam search (Div-BS). Note, beam search and diverse beam search are local search procedures which explore the output captioning space word-by-word. While, POS tag sequences act as global probes that permit to sample captions in many different parts of the captioning space.

GAN. More recent work on diverse image captioning focuses on using GANs. Adversarial training has been employed by [5, 14, 25] to generate diverse captions. [5, 14] train a conditional GAN for diverse caption generation. [5] uses a trainable loss which differentiates human annotations from generated captions. Ranking based techniques, which attempt to score human annotated captions higher than generated ones, are demonstrated in [14]. Shetty *et al*. [25] use adversarial training in combination with an approximate Gumbel sampler to match the generated captions to the human annotations.

Generally, GAN based methods improve on diversity, but suffer on accuracy. For example, in Tab. 1 of [25], ME-

TEOR and SPICE scores drop drastically compared to an LSTM baseline. In Tab. 4, we compare GAN [25] and our POS-based method which is more accurate.

VAE. Wang *et al*. [32] propose to generate diverse captions using a conditional variational auto-encoder with an additive Gaussian latent space (AG-CVAE) instead of a GAN. The diversity obtained with their approach is due to sampling from the learned latent space. They demonstrate improvements in accuracy over the conventional LSTM baseline. Due to the computational complexity of beam search they used fewer beams for the LSTM baseline compared to the number of captions sampled from the VAE, *i.e*., they ensured equal computational time. We compare to AG-CVAE [32] and show that we obtain higher best-1 caption accuracy (Tab. 3) and our best-k^{th} caption accuracy ($k = 1$ to 10) outperforms AG-CVAE (Fig. 3). Note, best-k scores in Tab. 3 and Fig. 3 denote the score of the k^{th} ranked caption given the same number of sampled captions (20 or 100) for all methods. For fairness, we use the same ranking procedure (*i.e*., consensus re-ranking proposed by [7] and used in [32]) to rank the sampled captions for all methods.

3. Background

Problem Setup and Notation. The goal of diverse captioning is to generate k sequences $y^1, y^2, \dots, y^k$, given an image. For readability we drop the super-script and focus on a single sequence y. The methods we discuss and develop will sample many such sequences y and rank them to obtain the best-k – $y^1, y^2, \dots, y^k$. A single caption $y = (y_1, \dots, y_N)$ consists of a sequence of words y_i, $i \in \{1, \dots, N\}$ which accurately describe the given image I. For each caption y, the words $y_i, i \in \{1, \dots, N\}$ are obtained from a fixed vocabulary $\mathcal{Y}$, *i.e*., $y_i \in \mathcal{Y}$. Additionally, we assume availability of a part-of-speech (POS) tagger for the sentence y. More specifically, the POS tagger provides a tag sequence $t = (t_1, \dots, t_N)$ for a given sentence, where $t_i \in \mathcal{T}$ is the POS tag for word y_i. The set $\mathcal{T}$ encompasses 12 universal POS tags – *verb (VERB), noun (NOUN), pronoun (PRON), etc.*[1]

To train our models we use a dataset $\mathcal{D} = \{(I, y, t)\}$ which contains tuples (I, y, t) composed of an image I, a sentence y, and the corresponding POS tag sequence t. Since it is not feasible to annotate the $\sim$ .5M captions of MSCOCO with POS tags, we use an automatic part-of-speech tagger.[1]

Classical Image Captioning. Classical techniques factor the joint probabilistic model $p_\theta(y|I)$ over all words into a product of conditionals. They learn model parameters θ^* by

[1] See https://www.nltk.org/book/ch05.html for POS tag and automatic POS tagger details

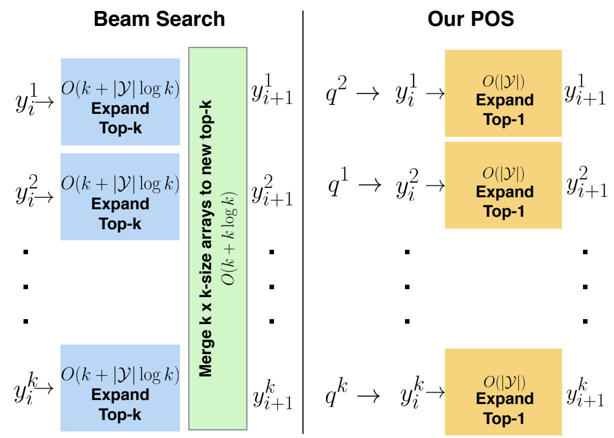


Figure 1: Illustration of beam search and POS-sampling to expand the best-k captions $(y_i^1, y_i^2, \dots y_i^k)$ from word position i to $i+1$. See Sec. 3 for notation and other details.

maximizing the likelihood over the training set $\mathcal{D}$, *i.e*.,

$$\max_\theta \sum_{(I,y)\in\mathcal{D}} \log p_\theta(y|I), \quad \text{where} \;\; p_\theta(y|I) = \prod_{i=1}^{N} p(y_i|y_{<i}, I). \tag{1}$$

The factorization of the joint probability distribution enforces a temporal ordering of words. Hence, word y_i at the i^{th} time-step (or word position) depends only on all previous words $y_{<i}$. This probability model is represented using a recurrent neural network or a feed-forward network with temporal (or masked) convolutions. Particularly the latter, *i.e*., temporal convolutions, have been used recently for different vision and language tasks in place of classical recurrent neural nets, *e.g*., [3, 9, 4].

During training, we learn the optimal parameters θ^*. Then for test image I, conditional word-wise posterior probabilities $p_{\theta^*}(y_i|y_{<i}, I)$ are generated sequentially from $i = 1$ to N. Given these posteriors, beam search is applicable and forms our baseline. Fig. 1 illustrates beam search with a beam width of k from word position y_i to y_{i+1}. Here, beam search maintains best-k (incomplete) captions ordered by likelihood. It expands the best-k captions at every word greedily from start to end of the sentence.

More specifically, for beam search from word position i, we first generate posteriors $p_{\theta^*}^j(y_{i+1}|y_{<(i+1)}^j, I)$ based on the current top-k list containing $y_{<(i+1)}^j$, $j \in \{1, \dots, k\}$. We then obtain new top-k captions by expanding each of the k entries $y_{<(i+1)}^j$ in the list using the computed posterior $p_{\theta^*}^j(y_{i+1}|y_{<(i+1)}^j, I)$. We call this 'expand top-k.' The time complexity for a single expand top-k operation is identical to obtaining the sorted top-k values from an array of size $|\mathcal{Y}|$.[2] The time complexity of all expand top-k operations is $O(k^2 + |\mathcal{Y}|k \log k)$.

[2] https://www.geeksforgeeks.org/k-largestor-smallest-elements-in-an-array/

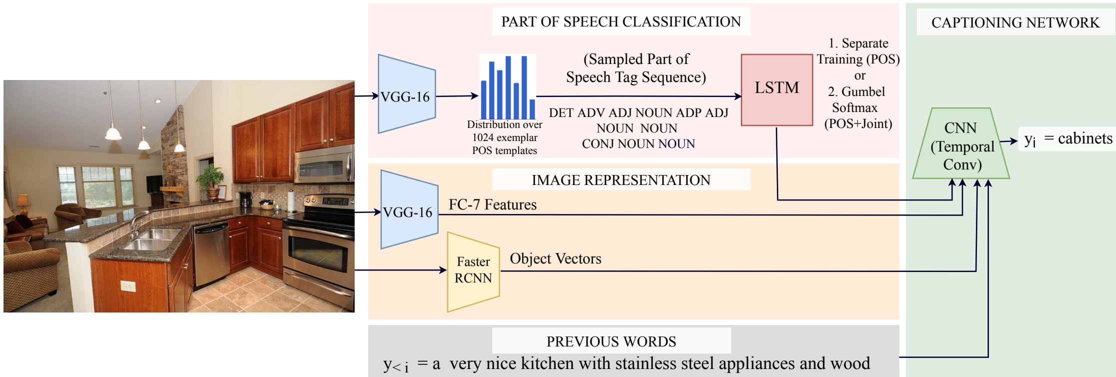


Figure 2: An illustration of our POS captioning method on a test image. For the image representation, fc7 features are extracted from VGG-16 and embedded into 512 dimensional vectors. For object vectors, we use the 80 dimensional class vector from faster rcnn [22] (same as [32]). For part-of-speech classification, we use VGG-16 with two linear layers and a 1024-way softmax. Then, we encode sampled POS via an LSTM-net to a 512 dimensional vector. Our captioning network uses temporal convolutions and operates on image representation, part-of-speech vector, object vector and previous words in the caption ($y_{<i}$) to produce the next word (y_i). The network is trained for 20 epochs using the ADAM optimizer [13] (initial learning rate of $5e^{-5}$ and a decay factor of .1 after 15 epochs). The part of speech classification step can be trained separately (POS) or jointly using a gumbel softmax (POS+Joint). Note, image representation is same for our method and baselines.

We merge all the expanded top-k captions to the final top-k captions using the log sum of the posterior probability at word position $i+1$. We call this operation merge. The merge operation has a complexity of $O(k + k\log k)$, which is identical to merging k sorted arrays.[3] In Sec. 4, we show that our inference with POS has better time complexity.

4. Image Captioning with Part-of-Speech

In our approach for image captioning, we introduce a POS tag sequence t, to condition the recurrent model given in Eq. (1). More formally, we use the distribution

$$p_\theta(y|t, I) = \prod_{i=1}^{N} p_\theta(y_i|t, y_{<i}, I). \tag{2}$$

Following classical techniques, we train our POS-conditioned approach by maximizing the likelihood (similar to Eq. (1)), *i.e.*, we want to find the parameters

$$\theta^* = \operatorname*{argmax}_\theta \sum_{(I,t,y)\in\mathcal{D}} \log p_\theta(y|t, I). \tag{3}$$

Importantly, note that we use the entire POS tag sequence in the conditional above, because it allows global control over the entire sentence structure.

Training involves learning the parameters θ^* for our conditional captioning model (Eq. (3)). During test time, conditioning on POS tags provides a mechanism for diverse image captioning, *i.e.*, given a test image I, we obtain k diverse captions by sampling k POS tag sequences $t^1, t^2, \ldots, t^k$. Note that every sequence is a tuple of POS tags, *i.e.*, $t^i = (t^i_1, t^i_2, \ldots), i \in \{1, \ldots, k\}$.

Since a large number of possible POS tag sequences exists, in Sec. 4.1, we discuss how we obtain quantized POS tag sequences $q^1, q^2, \ldots, q^k$ given the input image. These quantized sequences approximate the actual POS tag sequences $t^1, t^2, \ldots, t^k$.

Concretely, during inference we sample k quantized POS tag sequences given the image. This is shown as the part-of-speech classification step in Fig. 2. Then, we encode each sampled POS tag sequence q using an LSTM model. The encoded POS tag sequence, along with object vector, image features (fc7 of VGG-16) and previous words ($y_{<i}$) forms the input to the temporal convolutions-based captioning network. This captioning network implements our posterior probability $p_\theta(y_i|y_{<i}, q, I)$, which is used to predict the next word $y^*_i = \text{argmax}_{y_i} p_\theta(y_i|y_{<i}, q, I)$.

Fast inference with POS. For every sampled tag sequence $q^j, j \in \{1, 2, \cdots, k\}$ (*i.e.* quantization of tag sequence t^j), we maximize the learned probabilistic model, *i.e.*, $y^j_i = \text{argmax}_y p_{\theta^*}(y_i|y_{<i}, q^j, I)$ greedily. As just discussed, we simply use the maximum probability word at every word position. Fig. 1 compares this computationally much more effective method, which has a time complexity of $O(k|\mathcal{Y}|)$, to the breadth first approach employed by beam search.

Note that POS-based sampling requires only a single max-operation at every step during inference (our effective beam size is 1), making it faster than beam search with wide beams. It is also faster than diverse beam search (with group size parameter set to 1 as in our results) which performs the k 'expand top-k' operations sequentially using an augmented diversity function.

[3] https://www.geeksforgeeks.org/merge-k-sorted-arrays/

Method	Beam size or #samples	Best-1 Oracle Accuracy B4	B3	B2	B1	C	R	M	S	Speed (s/img)	Speed	Accuracy
Beam search	20	0.489✓	0.626✓	0.752✓	0.875✓	1.595✓	0.698✓	0.402✓	0.284✓	3.74×	×	✓
Div-BS [28]		0.383×	0.538×	0.687×	0.837	1.405	0.653	0.357	0.269	3.42	×	×
AG-CVAE [32]		0.471	0.573	0.698	0.834×	1.308×	0.638×	0.309×	0.244×	-	-	×
POS		0.449	0.593	0.737	0.874	1.468	0.678	0.365	0.277	0.21	✓	✓
POS+Joint		0.431	0.581	0.721	0.865	1.448	0.670	0.357	0.271	0.20✓	✓	✓
Beam Search	100	0.641✓	0.742✓	0.835✓	0.931✓	1.904✓	0.772✓	0.482✓	0.332✓	20.33	×	✓
Div-BS [28]		0.402×	0.555×	0.698×	0.846×	1.448×	0.666×	0.372	0.290	19.05	×	×
AG-CVAE [32]		0.557	0.654	0.767	0.883	1.517	0.690	0.345×	0.277×	-	-	×
POS		0.578	0.689	0.802	0.921	1.710	0.739	0.423	0.322	1.29	✓	✓
POS+Joint		0.550	0.672	0.787	0.909	1.661	0.725	0.409	0.311	1.27✓	✓	✓

Table 2: **Best-1 accuracy by oracle re-ranking**. Our POS methods are faster at sampling than beam search and they also generate a higher scoring best-1 caption than AG-CVAE [32] and Div-BS [28]. Beam search obtains the best scores, however it is slow. From all sampled captions (#samples = 20 or 100), we use oracle to pick the best-1 caption for every metric. This gives an estimate of the upper bound on captioning accuracy for each method. We use standard captioning metrics, BLEU (B1-B4) [21], CIDEr (C) [27], ROUGE (R) [15], METEOR (M) [6] and SPICE (S) [1]. Note, ✓indicates good performance on the metric for the corresponding column and × indicates bad performance.

4.1. Image to Part-of-Speech Classification

Because our model conditions sentence probabilities on a POS tag sequence, we need to compute it before performing inference. Several ways exist to obtain the POS tag sequence. *E.g.*, choosing a POS tag sequence by hand, sampling from a distribution of POS tag sequences seen in the dataset $\mathcal{D}$, or predicting POS tag sequences conditioned on the observed image I. The first one is not scalable. The second approach of sampling from $\mathcal{D}$ without considering the provided image is easy, but generates inaccurate captions. We found the third approach to yield most accurate results. While this seems like an odd task at first, our experiments suggest very strongly that image based prediction of POS tag sequences works rather well. Indeed, intuitively, inferring a POS tag sequence from an image is similar to predicting a situation template [36] – one must predict a rough template sketching what is worth to be said about an image.

To capture multi-modality, we use a classification model to compute our POS predictions for a given image I. However, we find that there are $> 210K$ POS tag sequences in our training dataset $\mathcal{D}$ of $|\mathcal{D}| > 500K$ captions. To maintain efficiency, we therefore quantize the space of POS tag sequences to 1024 exemplars as discussed subsequently.

Quantizing POS tag sequences. We perform a hamming distance based k-medoids clustering to obtain 1024-cluster centers. We use concatenated 1-hot encodings (of POS tags) to encode the POS tag sequence. We observe our clusters to be tight, *i.e.*, more than 75% of the clusters have an average hamming distance less than 3. We use the cluster medoids as the quantized POS tag sequences for our classifier. Given an input tag sequence t we represent it using its nearest neighbor in quantized space, which we denote by $q = \mathcal{Q}(t)$. Note, in our notation the quantization function $\mathcal{Q}(t)$, reduces t to its quantized tag sequence q.

Our image to part-of-speech classifier (shown in Fig. 2) learns to predict over quantized POS sequence space by maximizing the likelihood, $p_\phi(q|I)$. Formally, we look for its optimal parameters ϕ^* via

$$\phi^* = \operatorname*{argmax}_{\phi} \sum_{(I,t)\in\mathcal{D}} \log p_\phi(q|I), \tag{4}$$

where $\log p_\phi(q|I) = \sum_{i=1}^{1024} \delta[q^i = \mathcal{Q}(t)] \log p_\phi(q^i|I)$.

4.2. Separate vs. Joint Training

Training involves learning the parameters θ of the captioning network (Eq. (3)) and the parameters ϕ of the POS classification network (Eq. (4)). We can trivially train these two networks separately and we call this method **POS**.

We also experiment with joint training by sampling from the predicted POS posterior $p_\phi(t|I)$ using a Gumbel softmax [11] before subsequently using its output in the captioning network. Inconsistencies between sampled POS sequence and corresponding caption y will introduce noise since the ground-truth caption y may be incompatible with the sampled sequence q. Therefore, during every training iteration, we sample 50 POS tag sequences from the Gumbel soft-max and only pick the one q with the best alignment to POS tagging of caption y. We refer to this form of joint training via **POS+Joint**. In Sec. 5.1 and Sec. 5.2, we show that POS+Joint (*i.e.*, jointly learning θ and ϕ) is useful and produces more accurate captions.

5. Results

In the following, we compare our developed approach for diverse captioning with POS tags to competing baselines for diverse captioning. We first provide information about the dataset, the baselines and the evaluation metrics before presenting our results.

Dataset. We use the **MS COCO** dataset [16] for our experiments. For the train/val/test splits we follow: (1) M-RNN [20] using 118,287 images for training, 4,000 images

Method	Beam size or #samples	Best-1 Consensus Re-ranking Accuracy B4	B3	B2	B1	C	R	M	S	Speed (s/img)	Speed	Accuracy
Beam search (w. Likelihood)	20	0.305	0.402$^{\times}$	0.538$^{\times}$	0.709$^{\times}$	0.947$^{\times}$	0.523	0.248	0.175	3.19	$\times$	$\times$
Beam search		0.319	0.423	0.564	0.733	1.018	0.537$^{\checkmark}$	0.255	0.185	7.41	$\times$	$\checkmark$
Div-BS [28]		0.320$^{\checkmark}$	0.424$^{\checkmark}$	0.562	0.729	1.032$^{\checkmark}$	0.536	0.255$^{\checkmark}$	0.184	7.60$^{\times}$	$\times$	$\checkmark$
AG-CVAE [32]		0.299$^{\times}$	0.402$^{\times}$	0.544	0.716	0.963	0.518$^{\times}$	0.237$^{\times}$	0.173$^{\times}$	-	-	$\times$
POS		0.306	0.419	0.570$^{\checkmark}$	0.744$^{\checkmark}$	1.014	0.531	0.252	0.188$^{\checkmark}$	1.13$^{\checkmark}$	$\checkmark$	$\checkmark$
POS+Joint		0.305	0.415	0.563	0.737	1.020	0.531	0.251	0.185	1.13$^{\checkmark}$	$\checkmark$	$\checkmark$
Beam search (w. Likelihood)	100	0.300$^{\times}$	0.397$^{\times}$	0.532$^{\times}$	0.703$^{\times}$	0.937$^{\times}$	0.519$^{\times}$	0.246	0.174$^{\times}$	18.24	$\times$	$\times$
Beam search		0.317	0.419	0.558	0.729	1.020	0.532	0.253	0.186	40.39$^{\times}$	$\times$	$\checkmark$
Div-BS [28]		0.325$^{\checkmark}$	0.430$^{\checkmark}$	0.569$^{\checkmark}$	0.734	1.034	0.538$^{\checkmark}$	0.255$^{\checkmark}$	0.187	39.71	$\times$	$\checkmark$
AG-CVAE [32]		0.311	0.417	0.559	0.732	1.001	0.528	0.245$^{\times}$	0.179	-	-	$\times$
POS		0.311	0.421	0.567	0.737	1.036	0.530	0.253	0.188$^{\checkmark}$	7.54	$\checkmark$	$\checkmark$
POS+Joint		0.316	0.425	0.569$^{\checkmark}$	0.739$^{\checkmark}$	1.045$^{\checkmark}$	0.532	0.255$^{\checkmark}$	0.188$^{\checkmark}$	7.32	$\checkmark$	$\checkmark$

Table 3: **Best-1 accuracy by consensus re-ranking**. Our POS methods obtain higher scores on captioning metrics than AG-CVAE [32]. This demonstrates our POS natural language prior is more useful than the abstract latent vector used by VAE-based methods. POS methods obtain comparable accuracy to Beam Search and Div-BS [28], and they are more computationally efficient at sampling (*i.e.*, high speed). Note, we also outperform the standard beam search that uses likelihood based ranking. For these results, consensus re-ranking [7] is used to pick the best-1 caption from all sampled captions (unless 'w. Likelihood' is specified). For fair comparison, each method uses the same 80-dimensional object vector from faster rccn [23] and the same image features/parameters for consensus re-ranking. The captioning metrics are the same as in Tab. 2. Note, $\checkmark$ indicates good performance on the metric for the corresponding column and $\times$ indicates bad performance.

Method	#samples	Meteor	Spice
Beam Search (with VGG-16)	5	.247	.175
GAN (with Resnet-152)	5	.236	.166
POS+Joint (with VGG-16) [25]	5	**.247**	**.180**

Table 4: **Comparison to GAN-based method.** To compare to GAN, we train our POS+Joint on another split of MSCOCO by Karpathy *et al.* [12]. Our POS+Joint method samples more accurate best-1 captions than the GAN method. POS+Joint also obtains better SPICE score on this split compared to beam search. Our accuracy may improve with the use of Resnet-152 features. For fair comparison, we use the same 80-dimensional object vectors from faster rcnn [23] and rank the generated captions with likelihood for all methods.

for validation, and 1,000 images for testing; and (2) Karpathy *et al.* [12] using 113,287 images for training, 5,000 images for validation and 5,000 images for testing. The latter split is used to compare to GAN-based results in Tab. 4.

Methods. In the results, we denote our approach by **POS**, and our approach with joint training by **POS+Joint** (see Sec. 4.2 for the differences). We compare to the additive Gaussian conditional VAE-based diverse captioning method of Wang *et al.* [32], denoted by **AG-CVAE**. Our captioning network is based on [3]. For a fair comparison to beam search we also compare to convolutional captioning [3] with beam search. This is referred to as **beam search**. We compare to diverse beam search denoted denoted by **Div-BS**. The abbreviation **GAN** is used to denote the GAN-based method in [25].

Evaluation criteria. We compare all methods using four criteria – accuracy, diversity, speed, human perception:

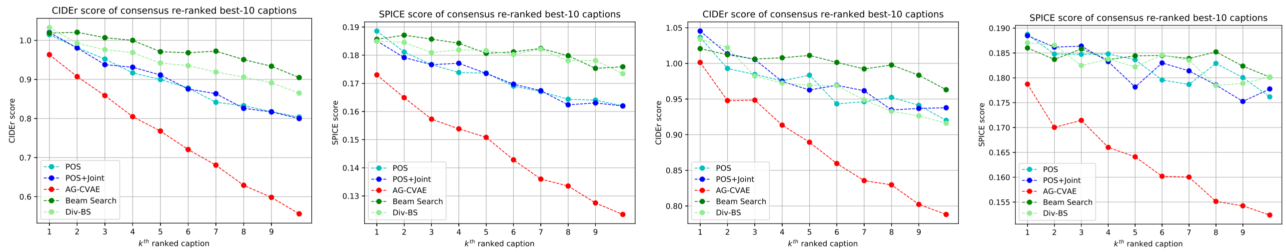
- **Accuracy.** In Sec. 5.1 (Best-1 Accuracy) we compare the accuracy using the standard image captioning task of generating a single caption. Subsequently, in Sec. 5.2 (Best-k^{th} Accuracy), we assess the accuracy of k captions on different image captioning metrics.
- **Diversity.** We evaluate the performance of each method on different diversity metrics in Sec. 5.3.
- **Speed.** In addition to accuracy, in Sec. 5.4, we also measure the computational efficiency of each method for sampling multiple captions.
- **Human perception.** We do a user study in Sec. 5.5.

5.1. Best-1 Accuracy

We use two ranking methods – oracle and consensus re-ranking – on the set of generated captions and pick the best-1 caption. Our results for oracle re-ranking in Tab. 2 and for consensus re-ranking in Tab. 3 show that, beam search and diverse beam search are accurate however slow. POS is both fast and accurate. POS outperforms the accuracy of AG-CVAE.

Oracle re-ranking. The reference captions of the test set are used and the generated caption with the maximum score for each metric is chosen as best-1 (as also used in [32]). This metric permits to assess the best caption for each metric and the score provides an upper bound on the achievable best-1 accuracy. Higher oracle scores are also indicative of the method being a good search method in the space of captions. Results in Tab. 2 show that beam search obtains the best oracle scores. However, it is painfully slow ($\sim$ 20s per image to sample 100 captions). POS, POS+Joint obtain higher accuracy than AG-CVAE and comparable accuracy to beam search with faster runtime.

Consensus re-ranking scores. In a practical test setting,



(a) Best-10 CIDEr from 20 samples (b) Best-10 SPICE from 20 samples (c) Best-10 CIDEr from 100 samples (d) Best-10 SPICE from 100 samples

Figure 3: **Best-10 CIDEr and SPICE accuracy.** Our POS and POS+Joint achieve best-k accuracy comparable to Beam Search and Div-BS [28] with faster computation time. We outperform the best-k scores of AG-CVAE [32], demonstrating part-of-speech conditioning is better than abstract latent variables of a VAE. Note, this figure is best viewed in high-resolution.

Method	Beam size or #samples	Distinct Captions	# Novel sentences (Best-5)	mBleu-4 (Best-5)	n-gram Diversity (Best-5) Div-1	Div-2	Overall Diversity
Beam search		**100%**	2317	0.777	0.21	0.29	×
Div-BS [28]		**100%**	3106	0.813	0.20	0.26	×
AG-CVAE [32]	20	69.8%	3189	0.666	**0.24**	0.34	✓
POS		96.3%	3394	**0.639**	**0.24**	**0.35**	✓
POS+Joint		77.9%	**3409**	0.662	0.23	0.33	✓
Beam search		**100%**	2299	0.781	0.21	0.28	×
Div-BS [28]		**100%**	3421	0.824	0.20	0.25	×
AG-CVAE [32]	100	47.4%	3069	0.706	**0.23**	0.32	✓
POS		91.5%	**3446**	**0.673**	**0.23**	**0.33**	✓
POS+Joint		58.1%	3427	0.703	0.22	0.31	✓
Human	5	99.8%	-	0.510	0.34	0.48	

Table 5: **Diversity statistics.** For each method, we report the number of novel sentences (*i.e.*, sentences not seen in the training set) out of at most best-5 sentences after consensus re-ranking. Though Beam Search showed high accuracy in Tab. 2, 3 and Fig. 3, here, we see that it produces less number of novel sentences than our POS methods. Therefore, beam search is more prone to regurgitating training data. Low mBleu-4 indicates lower 4-gram overlap between generated captions and more diversity in generated captions. POS has the lowest mBleu-4 and therefore high diversity in generated captions. For details on other metrics see Sec. 5.3.

reference captions of the test set won't be available to rank the best k captions and obtain best-1. Therefore, in consensus re-ranking, the reference captions of training images similar to the test image are retrieved. The generated captions are ranked via the CIDEr score computed with respect to the retrieved reference set [7].

We use the same image features [31] and parameters for consensus re-ranking as [32]. Tab. 3 shows that our methods POS and POS+Joint outperform the AG-CVAE baseline on all metrics. Moreover, our methods are faster than beam search and diverse beam search. They produce higher CIDEr, Bleu-1,2, METEOR and SPICE scores. Other scores are comparable and differ in the 3rd decimal. Note, our POS+Joint achieves better scores than POS, especially for 100 samples. This demonstrates that joint training is useful.

We also train our POS+Joint method on the train/test split of Karpathy *et al.* [12] used by the GAN method [25]. In Tab. 4, we show that we obtain higher METEOR and SPICE scores than those reported in [25].

Baseline Method	POS Wins	Baseline Method Wins
Beam search	**57.7%**	42.2%
Diverse beam search [28]	45.3%	**54.6%**
AG-CVAE [32]	**64.8%**	35.1%

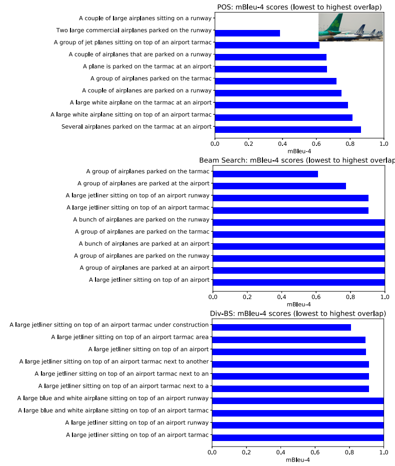
Table 6: We show the user captions sampled from best-k (same k^{th} ranked, $k = 1$ to 5) for baseline methods and our POS. The user is allowed to pick the caption that best describes the image. Note, user is not aware of the method that generated the caption. Here, we observe that our POS method outperforms Beam search and AG-CVAE on our user study. Our user study has 123 participants with on average 23.3 caption pairs annotated by each user.

5.2. Best-k^{th} Accuracy

Our captioning method can be conditioned on different part-of-speech tags to generate diverse captions. For diverse image captioning, in addition to best-1 accuracy, best-k^{th} accuracy should also be measured. Best-k^{th} accuracy is the score of the k^{th} ranked caption, therefore it is lower than the best-1 score. All k generated captions should be accurate and therefore it is desirable to have high best-k^{th} scores. This metric has not been reported previously [25, 32].

	**POS:** - two people are standing on the back of an elephant. - a man and a woman are on the back of an elephant. - two people standing on top of an elephant. - a group of people riding an elephant in a park. - two people are riding an elephant on a dirt road. **Diverse Beam Search:** - two people are standing next to an elephant. - a couple of people standing next to an elephant. - a couple of people standing next to an elephant on a dirt road. - a couple of people that are standing next to an elephant. - a couple of people standing next to an elephant on top of a.	**Beam Search:** - a couple of people standing on top of an elephant. - a couple of people are standing on top of an elephant. - a man and a woman standing next to an elephant. - a man and woman standing next to an elephant. - a group of people standing next to an elephant. **AG-CVAE:** - a group of people riding on top of an elephant. - a man and a man riding on top of an elephant. - a large group of people riding on top of an elephant. - a man riding on the back of a elephant. - a group of people standing on top of an elephant.
	**POS:** - a rear view mirror on the side of a car window. - a side view mirror of a car with a bird on the window. - a rear view mirror hanging on the side of a car. - a side view of a car with birds in the side mirror. - a view of a mirror of a car looking at a mirror. **Diverse Beam Search:** - a close up of a bird on a car mirror. - a bird is sticking its head out of a car window. - a close up of a bird on a car. - a close up of a bird on the back of a car. - a bird that is sitting in the back of a car.	**Beam Search:** - a reflection of a bird on the back of a truck. - a close up of a bird on the back of a vehicle. - a bird is perched on the back of a car. - a bird is sitting in the seat of a car. - a bird that is sitting in the back seat of a car. **AG-CVAE:** - a dog is looking out the window of a car. - a dog is sitting in the window of a car. - a small bird sitting on the side of a car. - a dog sitting on the side of a car. - a bird sitting on the back of a car.

(a) Qualitative Comparison



(b) Diversity (or Overlap) Comparison

Figure 4: In figure on left, notice POS captions contain things like rear/side view mirror, dirt road, the quantifier 'two' which is less common in other methods. The inaccuracies are highlighted in red and the novel parts in green. In figure on right, we compare the diversity (or overlap) of captions. The mBleu-4 score measures 4-gram overlap between one generated caption and the rest. Lower is better, *e.g.*, 0 means caption has no 4-gram overlap to other sentences. POS is better than BS and Div-BS in the plots above (lower mBleu-4 scores). Note, ground-truth 5 captions all have 0 overlap to each other for this example. On our 1000 image test set with 10 captions generated per image, POS generates 10.94% sentences with 0 overlap; in contrast Div-BS generates 1.02% and Beam Search 2.4%. Figure best viewed in high-resolution.

In Fig. 3, we compare best-k^{th} ($k = 1$ to 10) scores for all methods. Note, the accuracy of AG-CVAE drops drastically on both CIDEr and Spice, while our POS methods maintain accuracy comparable to beam search. This proves that our POS image summaries are better at sampling accurate captions than the abstract latent variables of a VAE.

5.3. Evaluation of Diversity

In Tab. 5 we compare methods on diversity metrics.
(1) **Uniqueness.** The number of unique sentences generated after sampling. Beam search and diverse beam search always sample a unique sentence. Note, our POS also samples a high number of unique sentences 19.26 (96.3%) out of 20, 91.55 out of 100. The uniqueness reduces for joint training. This is because, generation of a caption while training POS+Joint is based on a noisy POS tag sequence sampled from the Gumbel softmax. Therefore, the caption may not be compatible with this noisy POS tag sequence which leads to an overly smooth latent representation for the POS tag. Therefore, different POS tags may produce the same latent code and hence the same caption.
(2) **Novel sentences.** We measure the number of novel sentences (not seen in train), and find that our POS-based methods produce more novel sentences than all other methods. Beam search produces the least number of novel sentences.
(3) **Mutual overlap.** We also measure the mutual overlap between generated captions. This is done by taking one caption out of k generated captions and evaluating the average Bleu-4 with respect to all other $k-1$ captions. Lower value indicates higher diversity. POS is the most diverse. Note, the average score is computed by picking every caption *vs.* the remaining $k-1$ captions.
(4) **n-gram diversity (div-n).** We measure the ratio of distinct n-grams per caption to the total number of words generated per image. POS outperforms other methods.

5.4. Speed

In Fig. 1 we showed that our POS based methods have better time complexity than beam search and diverse beam search. The time complexity of our POS-based approach is the same as sampling from a VAE or GAN, provided the max probability word is chosen at each word position (as we do). The empirical results in Tab. 2 and Tab. 3 show that POS methods are $5\times$ faster than beam search methods.

5.5. User Study

Fig. 4 compares the captions generated by different methods and in Tab. 6, we provide the results of a user study. A user is shown two captions sampled from two different methods. The user is asked to pick the more appropriate image caption. Tab. 6 summarizes our results. We observe POS outperforms AG-CVAE and Beam search.

6. Conclusion

The developed diverse image captioning approach conditions on part-of-speech. It obtains higher accuracy (best-1 and best-10) than GAN and VAE-based methods and is computationally more efficient than the classical beam search. It performs better on different diversity metrics compared to other methods.

Acknowledgments: Supported by NSF Grant No. 1718221 and ONR MURI Award N00014-16-1-2007.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 5

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *arXiv preprint arXiv:1707.07998*, 2017. 1, 2

[3] J. Aneja, A. Deshpande, and A. Schwing. Convolutional image captioning. In *Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 6

[4] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, abs/1803.01271, 2018. 3

[5] B. Dai, S. Fidler, R. Urtasun, and D. Lin. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*, 2017. 1, 2

[6] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014. 5

[7] J. Devlin, S. Gupta, R. B. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. *CoRR*, abs/1505.04467, 2015. 3, 6, 7

[8] J. R. Finkel, C. D. Manning, and A. Y. Ng. Solving the problem of cascading errors: Approximate bayesian inference for linguistic annotation pipelines. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 2006. 1

[9] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. *CoRR*, abs/1705.03122, 2017. 3

[10] K. Gimpel, D. Batra, C. Dyer, and G. Shakhnarovich. A systematic exploration of diversity in machine translation. In *In Proc. of EMNLP*, 2013. 1

[11] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. 2017. 5

[12] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, June 2015. 1, 2, 6, 7

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. abs/1412.6980, 2014. 4

[14] D. Li, X. He, Q. Huang, M.-T. Sun, and L. Zhang. Generating diverse and accurate visual captions by comparative adversarial learning. *arXiv preprint arXiv:1804.00861*, 2018. 1, 2

[15] C.-Y. Lin. Rouge: a package for automatic evaluation of summaries. July 2004. 5

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. *Microsoft COCO: Common Objects in Context*, pages 740–755. Springer International Publishing, Cham, 2014. 5

[17] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. *arXiv preprint arXiv:1612.00370*, 2016. 2

[18] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk, 2018. 2

[19] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich. Discriminability objective for training descriptive captions. *arXiv preprint arXiv:1803.04376*, 2018. 2

[20] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*, 2015. 1, 5

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 5

[22] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NIPS)*, 2015. 4

[23] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 91–99, Cambridge, MA, USA, 2015. MIT Press. 6

[24] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 2

[25] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 3, 6, 7

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

[27] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575. IEEE Computer Society, 2015. 5

[28] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. J. Crandall, and D. Batra. Diverse beam search for improved description of complex scenes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018. 1, 2, 5, 6, 7

[29] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, Apr. 2017. 1, 2

[30] J. Wang, P. S. Madhyastha, and L. Specia. Object counts! bringing explicit detections back into image captioning. In *Proceedings of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies (NAACL HLT)*. Association for Computational Linguistics, 2018. 2

[31] L. Wang, Y. Li, and S. Lazebnik. Learning two-branch neural networks for image-text matching tasks. *CoRR*, abs/1704.03470, 2017. 7

[32] L. Wang, A. G. Schwing, and S. Lazebnik. Diverse and accurate image description using a variational auto-encoder with

an additive gaussian encoding space. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5758–5768, 2017. 1, 2, 3, 4, 5, 6, 7

[33] Q. Wang and A. B. Chan. Cnn+cnn: Convolutional decoders for image captioning. *CoRR*, abs/1805.09019, 2018. 1, 2

[34] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2048–2057. JMLR.org, 2015. 1, 2

[35] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4904–4912, 2017. 2

[36] M. Yatskar, L. Zettlemoyer, and A. Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*, 2016. 5