Sequential Latent Spaces for Modeling the Intention During Diverse Image Captioning

Jyoti Aneja*[1], Harsh Agrawal*[2], Dhruv Batra[2,3], Alexander Schwing[1]

[1]University of Illinois, Urbana-Champaign, [2]Georgia Institute of Technology, [3]Facebook AI Research
[1]{janeja2, aschwing}@illinois.edu , [2]{hagrawal9, dbatra}@gatech.edu

Abstract

Diverse and accurate vision+language modeling is an important goal to retain creative freedom and maintain user engagement. However, adequately capturing the intricacies of diversity in language models is challenging. Recent works commonly resort to latent variable models augmented with more or less supervision from object detectors or part-of-speech tags [10, 40]. In common to all those methods is the fact that the latent variable either only initializes the sentence generation process or is identical across the steps of generation. Both methods offer no fine-grained control. To address this concern, we propose ***Seq-CVAE*** *which learns a latent space for every word position. We encourage this temporal latent space to capture the 'intention' about how to complete the sentence by mimicking a representation which summarizes the future. We illustrate the efficacy of the proposed approach to anticipate the sentence continuation on the challenging MSCOCO dataset, significantly improving diversity metrics compared to baselines while performing on par w.r.t. sentence quality.*

1. Introduction

Diverse yet accurate image captioning is an important goal towards augmenting present-day editing and auto-response tools with technology that maintains creative freedom while providing meaningful and inspiring suggestions. On the quest to succeed in this tightrope walk, methods need to maintain accuracy of the provided descriptions while elaborately managing the intricacies of the respective language. This balancing act to aesthetically craft short yet precise statements that hit the point is an art.

Despite best efforts, any description is always and inherently targeted towards a group of readers. Because words are overloaded, the crisp picture that we intend to draw blurs rapidly if we don't use language and associations that readers are familiar with. Without the right language the message of the description is diluted, remains hard to access or even inaccessible.

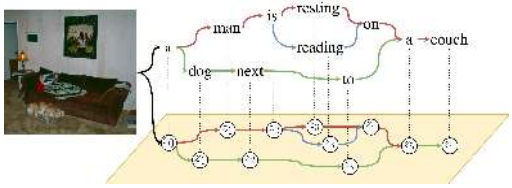


Figure 1: Meaningful diverse captions generated (blue arrows) for a given image by linearly interpolating from one latent vector (green arrows) to another (red arrows).

Going forward, imagine your description to automatically adjust depending on the background knowledge of the reader. Obviously we are far from this idea being remotely feasible. Nonetheless, in recent years, remarkable progress has been made in image captioning [3, 6, 7, 12–14, 21, 22, 24, 25, 30, 36, 39, 41] and particularly controllable [10, 40] and diverse [9, 10, 26, 34, 38, 40] image captioning. Many of the proposed mechanisms for image captioning rely on long-short-term-memory (LSTM) [18] nets where words are generated one at a time. For diversity, LSTM based variational auto-encoders [23] or generative adversarial nets [16] and their conditional counterparts [37] are employed. For high-level control, one-hot encodings that represent observed objects or groups of objects are injected at the first step of the LSTM [40]. Very recently [10], more low-level control has also been discussed by conditioning on abstract representations of part-of-speech tags. Again, the conditioning was achieved by changing the initial LSTM input.

Because of this single initial conditioning input, none of the aforementioned methods provide the fine-grained diversity that is desirable for adjusting individual words of a sentence. We address this shortcoming by developing **Seq-CVAE**, a method which learns a latent space for every word. Importantly, we want the latent space to be predictive of

*First two authors contributed equally.

the subsequent parts of the sentence, *i.e.*, the future of the sentence. We achieve this by employing a data-dependent transition model which captures the 'intention,' *i.e.*, a representation which encodes the remaining part of the sentence. During training the intention model is tasked to fit the representations of a backward LSTM.

This proposed approach enables to distinctly modify descriptions starting from a particular position. We demonstrate this fine-grained diversity by sampling a diverse set of captions and linearly interpolating between the chosen latent representations. As illustrated in Fig. 1, a convex combination of latent vectors permits to gradually transition from one caption to another.

We evaluate the proposed approach quantitatively on the challenging MSCOCO [27] dataset and significantly outperform competing methods w.r.t. diversity metrics: among 5000 sampled captions more than 4200 are novel and not part of the training set while the runner-up baselines produces just short of 3500 novel ones. Despite this diversity the proposed approach is on par w.r.t. accuracy metrics. These results are particularly remarkable because all competing baselines use additional information like object detectors [32] during inference, while the proposed approach does not use any additional information.

Contributions: We develop an image captioning model with a sequential latent space to capture the intention, *i.e.*, the future of the sentence (Sec. 3). We show that sampling with sequential latent spaces results in significantly more diverse captions than baselines (Tab. 2) despite not using any additional information. Perceptual metrics of our diverse captions are on par with baselines (Tab. 1). The sequential latent space permits distinct access to sentences starting from any specific position (Fig. 1).

2. Related Work

Image captioning and paragraph generation [3, 4, 6, 7, 12–14, 19–22, 24, 25, 30, 36, 39, 41] have attracted a significant amount of work. Early classical approaches are based on sentence retrieval [3]: the best fitting sentence from a set of possible descriptions is recovered by matching sentence representations with image representations. Those representations are learned from a set of available captions. However, firstly, this matching procedure is computationally expensive and, secondly, it is demanding to construct a database of captions which is sufficiently large to describe a reasonably comprehensive set of images accurately.

Image Captioning: Therefore, more recently, recurrent neural networks (RNNs) and variants like long-short-term-memory (LSTM) [18] networks decompose the caption space into a product space of individual words. Specifically, image representations are first extracted via a convolutional deep network which are subsequently used to prime the LSTM based recurrent network. The latter is trained via maximum likelihood to predict the next word given current and past words. Extensions involve object detectors [42], attention-based deep networks [1], and convolutional approaches [2]. Beyond maximum likelihood, reinforcement learning based techniques have also been discussed to produce a single caption, directly optimizing perceptual metrics [28, 33]. All these methods have demonstrated compelling results and have consequently been adopted widely. However, predicting a single caption does not allow for modeling ambiguity that is inherent. Consequently, diversity based methods have very recently been discussed.

Diversity in Image Captioning: To achieve diversity, four techniques have been investigated. Among the first was *beam search*, a classic approach to sample multiple captions which are assigned a high probability by the underlying model. While multiple captions are readily available, results usually only differ slightly because single word changes affect the sentence probability minimally.

To address this concern, *diverse beam search* [38] augments beam search. It advocates for more drastic changes by encouraging to recover different modes of a probability distribution rather than high-probability configurations.

To avoid sampling from a distribution defined over a high-dimensional output space, *generative adversarial networks* (GANs) have been proposed [9, 26, 34]. While GAN based methods improve diversity, they tend to suffer on perceptual metrics.

Variational auto-encoders (VAEs) are a fourth direction that has been explored [40]. The intuition is identical to the one of GANs, *i.e.*, avoid sampling from a distribution defined over a high-dimensional output space. However, in contrast to GAN-based methods, VAE-based image captioning techniques tend to produce high-quality captions when evaluated on perceptual metrics.

Similar to the aforementioned approaches we develop an approach based on VAEs. However, different from all the aforementioned techniques we also aim at incorporating more fine-grained diversity.

Controllability in Image Captioning: Beyond diversity, controllability of captions has become an important topic very recently. In particular Wang *et al.* [40] use a variational auto-encoder conditioned on object detections to control diversity. While intuitive, control remains indirect as the sentence generating decoder is only influenced at its first timestep. Influencing subsequent generation of words did not significantly change the result. Even more recently POSCap [10] was developed. While also only priming the decoder at its first step, use of clustered part-of-speech tags was proposed and shown to improve diversity. However, due to use of *encoded and clustered* part-of-speech tags, controllability was limited.

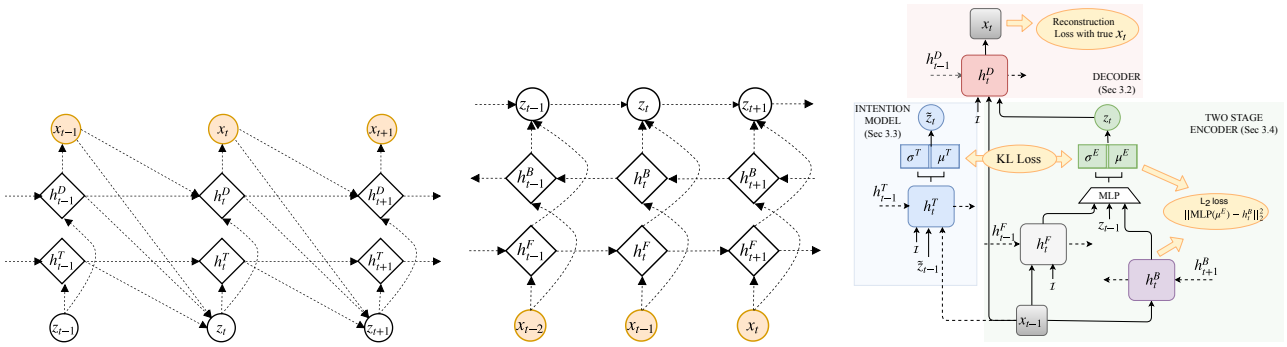


Figure 2: **(a)**: Computation graph for the generation network. The hidden states of the intention model LSTM and the decoder LSTM are h_t^T and h_t^D respectively. At a given time step t, the latent sample z_t depends on all prior words $x_{<t}$ and all prior latent samples $z_{<t}$. The sample z_t along with all prior words $x_{<t}$ predicts x_t. **(b)**: Computation graph for the encoder network. The hidden states of the forward LSTM and the backward LSTM are h_t^F and h_t^B respectively. At a given time t, the latent sample z_t depends on the entire caption x through the forward and backward models. **(c)**: Our proposed **Seq-CVAE** training architecture, shown for a single time slice t. Our model includes three components during training: (1) Two-stage encoder; (2) Intention Model; and (3) Decoder. Details for each of the components are provided in Sec. 3. At test time only decoder and intention model are used.

In contrast to the aforementioned techniques, we develop a VAE-based technique which learns a latent space for every word position. While enabling diversity, direct control over words emitted at a particular position is also possible as illustrated in Fig. 1.

Sequential VAE: Our proposed approach is related to a sequence of papers on sequential recurrent nets. Fraccaro *et al*. [15] develop SRNN, Chung *et al*. [8] devise VRNN, and Goyal *et al*. [17] propose Z-forcing. Although VRNN [8], SRNN [15], and Z-forcing [17] have similar intuition, *i.e*., maximizing a lower bound of the data likelihood, these models differ in the assumptions made on the graphical model for data generation, the choice of the prior, approximate posterior used for amortized variational inference and the decoder networks:

(1) *VRNN* uses a filtering posterior, *i.e*., the latent distribution at each time step depends on (a) all the previous latent vectors and (b) the *previous* input data. Instead we use a smoothing posterior, where the latent distribution at a given time depends on (a) the latent vector from previous time steps and (b) input data from *all* time steps, past and future. This leads to better models, since all context is provided.

(2) *SRNN* uses a smoothing posterior via a backward RNN as we do. However, decoder and prior differ: (a) unlike us, SRNN doesn't use latent variables in the autoregressive decoder, hence intention isn't available; (b) SRNN uses a Markovian prior, while we include the entire history of latent variables for the prior at time t.

(3) *Z-forcing* assumptions are similar, but differ architecturally: their prior, decoder and the approximating posterior share the same forward LSTM. This is undesirable since different distributions are best served by their own individual representation.

More crucially, these methods assess test set log-likelihood for sequential modeling, or perplexity on the IMDB dataset. In contrast, along with accuracy, we also care about **diversity** for image captioning. Hence, we are the first to extensively study these models on various measures of diversity (Tab. 2, Fig. 3, Fig. 7).

3. Approach

We first outline the proposed approach before discussing individual components.

3.1. Overview

Given an image I we are interested in generating a diverse set of captions x^k, $k \in \{1, \ldots, K\}$. For readability we drop the index k henceforth and note that the developed method will produce many captions that are ranked subsequently. Every generated caption $x = (x_1, \ldots, x_T)$ is a tuple consisting of words $x_t \in \mathcal{X}$, $t \in \{1, \ldots, T\}$, each from a discrete vocabulary $\mathcal{X}$. Given an image I, we devise a probabilistic model $p_\theta(x|I)$ which depends on parameters θ and assigns a probability to every caption x.

To effectively sample from this probabilistic space, we assume the probability distribution $p_\theta(x|I)$, jointly defined over all words x_t of a caption, to factorize into a product of word-conditionals, *i.e*.,

$$p_\theta(x|I) = \prod_{t \in \{1,\ldots,T\}} p_\theta(x_t|x_{<t}, I).$$

This factorization enforces a temporal ordering, *i.e*., the probability distribution for word x_t is conditioned on all preceding words $x_{<t}$. Importantly, because the conditional's domain is the vocabulary space $\mathcal{X}$ and not a product space thereof, as it is the case for the joint distribution $p_\theta(x)$, ancestral sampling is a suitable and effective technique to generate a diverse set of captions.

In practice, the conditional probability distributions $p_\theta(x_t|x_{<t}, I)$ are modeled via recurrent LSTM nets or

masked convolutions, where we use h^D_{t-1} to refer to its hidden state which summarizes $x_{<t-1}$, while x_{t-1} directly influences the distribution. However, given the preceding words $x_{<t}$ the conditional distribution $p_\theta(x_t|x_{<t}, I)$ has to cover many suitable options to complete the sentence. While LSTM-nets can model complex distributions, we hypothesize that a latent variable z_t which captures the current intention about how to complete the sentence can significantly reduce the complexity of the conditional.

Therefore, instead of directly modeling $p_\theta(x_t|x_{<t}, I)$ which marginalizes across all possible intentions, we are interested in modeling and sampling from the decoder distribution $p_\theta(x_t|x_{<t}, z_{\leqslant t}, I)$, *i.e.*, a distribution conditioned on the current and all previous intentions, which are however unobserved.

During inference, as illustrated in Fig. 2(a), we ensure effective sampling of an intention z_t by modeling the transition through the tuple of all intentions $z = (z_1, \dots, z_T)$ again via a product of conditionals $p_\theta(z_t|z_{<t}, x_{<t}, I)$, *i.e.*,

$$\text{(intention model)} \qquad \prod_{t\in\{1,\dots,T\}} p_\theta(z_t|z_{<t}, x_{<t}, I).$$

To obtain a description for a given image, as shown in Fig. 2(a), we alternatingly sample from $p_\theta(z_t|z_{<t}, x_{<t}, I)$, which is sometimes referred to as the 'prior,' and from the decoder $p(x_t|x_{<t}, z_{\leqslant t}, I)$. Different from classical approaches we employ a parametric 'intention model' which decomposes temporally. Note that the 'intention distribution' at time t, *i.e.*, $p_\theta(z_t|z_{<t}, x_{<t}, I)$ is dependent on $x_{<t}$. This is technically correct due to the temporal decomposition, *i.e.*, the distribution at time t can depend on all previously available data. Similar to the decoder, we use an LSTM net to capture the recurrence of the intention model and refer to its latent state via h^T_t.

To model the intentions z, during training, as illustrated in Fig. 2(b), we encourage the intention model to fit the approximate posterior $q_\phi(z|x, I)$ which we model using again a product of conditionals, *i.e.*,

$$\text{(approx. posterior)} \quad q_\phi(z|x, I) = \prod_{t\in\{1,\dots,T\}} q_\phi(z_t|z_{t-1}, x, I).$$

The distribution $q_\phi(z|x, I)$ is commonly referred to as the encoder. To adequately capture the intention on how to complete the sentence, as illustrated in Fig. 2(b), we develop a two-stage encoder consisting of a forward stage to model language and a backward stage to summarize intention, *i.e.*, the future of the sentence. We discuss details in Sec. 3.4.

During training we are given a dataset $\mathcal{D} = \{(I, x)\}$ consisting of pairs (I, x), each containing an image I and a corresponding caption x. We maximize the data log-likelihood $\sum_{(I,x)\in\mathcal{D}} \ln p_\theta(x|I)$. For readability we drop the summation over samples in the dataset subsequently. By using the aforementioned decompositions we note that the data log-likelihood $\ln p_\theta(x|I)$ is obtained by marginalizing over the space of intentions, *i.e.*,

$$\begin{aligned}\ln p_\theta(x|I) &= \ln \sum_z p_\theta(x, z|I) \\ &= \ln \sum_z \prod_t \overbrace{\underbrace{p_\theta(x_t|x_{<t}, z_{\leqslant t}, I)}_{\text{decoder}} \underbrace{p_\theta(z_t|z_{<t}, x_{<t}, I)}_{\text{intention}}}^{p_\theta(x_t, z_t|x_{<t}, z_{<t}, I)}.\end{aligned}$$

Marginalization over the space of intentions makes maximization of this objective computationally expensive. It is therefore common to utilize an approximate posterior and apply Jensen's inequality which gives the lower bound

$$\ln p_\theta(x|I) \geqslant \mathbb{E}_{z\sim q_\phi(z|x,I)}\left[\ln \frac{p_\theta(x, z|I)}{q_\phi(z|x, I)}\right].$$

Combined with the employed temporal decomposition, this yields the objective

$$\begin{aligned}\mathbb{E}_{z\sim q_\phi(z|x,I)}\Bigg[\sum_t \big(&\ln p_\theta(x_t|x_{<t}, z_{\leqslant t}, I) \\ &+\ln p_\theta(z_t|z_{<t}, x_{<t}, I) - \ln q_\phi(z_t|z_{t-1}, x, I)\big)\Bigg],\end{aligned} \tag{1}$$

which we maximize w.r.t. parameters θ and ϕ. In the following we discuss decoder, prior (intention model) and encoder in more detail. Notice, all their parameters are subsumed in the vectors θ and ϕ and jointly trained end-to-end. A single timestep of all three recurrent models is illustrated in Fig. 2(c).

3.2. Decoder

As illustrated in the top part of Fig. 2(c), the decoder is a classical LSTM net. At time t the decoder yields a multinomial probability distribution $p_\theta(x_t|x_{<t}, z_{\leqslant t}, I)$ defined over words $x_t \in \mathcal{X}$.

While representations of z_t and x_{t-1} are concatenated before being provided as input to the LSTM net, we encode dependence on $x_{<t-1}$ and $z_{<t}$ via its hidden representation h^D_{t-1}. Dependence on the image is encoded into the LSTM net via an image embedding obtained from the fc7 layer of a VGG16 network [35], pre-trained on the ImageNet dataset. The image embedding is fed as input at every time step of the LSTM, concatenated with the input word embedding and the sampled vector from the latent space. For all our experiments, we found a 512-dimensional latent vector to give the best results (Tab. 4).

3.3. Intention Model

Similar to the decoder we model the intention transition model $p_\theta(z_t|z_{<t}, x_{<t}, I)$ as an LSTM net. However, different from the decoder, given $z_{<t}$ and $x_{<t}$, we model $p_\theta(z_t|z_{<t}, x_{<t}, I)$ as a Gaussian distribution with

	Method	Best-1 Oracle Accuracy on M-RNN Test Split							
		B4	B3	B2	B1	C	R	M	S
Beam size #samples: 20	Beam search	**0.489**	**0.626**	**0.752**	**0.875**	**1.595**	**0.698**	**0.402**	**0.284**
	Div-BS [38]	0.383	0.538	0.687	0.837	1.405	0.653	0.357	0.269
	AG-CVAE [40]	0.471	0.573	0.698	0.834	1.259	0.638	0.309	0.244
	POS [10]	0.449	0.593	0.737	0.874	1.468	0.678	0.365	0.277
	Seq-CVAE	0.445	0.591	0.727	0.870	1.448	0.671	0.356	0.279
Beam size #samples: 100	Beam Search	**0.641**	**0.742**	**0.835**	**0.931**	**1.904**	**0.772**	**0.482**	**0.332**
	Div-BS [38]	0.402	0.555	0.698	0.846	1.448	0.666	0.372	0.290
	AG-CVAE [40]	0.557	0.654	0.767	0.883	1.517	0.690	0.345	0.277
	POS [10]	0.578	0.689	0.802	0.921	1.710	0.739	0.423	0.322
	Seq-CVAE	0.575	0.691	0.803	0.922	1.695	0.733	0.410	0.320

Table 1: **Best-1-Oracle Accuracy.** Our **Seq-CVAE** method obtains high scores on standard captioning metrics. We obtain an accuracy comparable to both the very recently proposed POS approach [10] which uses a part-of-speech prior and also to the AG-CVAE method [40]. Both these methods use additional information in the form of object vectors from a Faster-RCNN [32] during inference. In contrast, we do not use any additional information during inference. To calculate the best-1-accuracy, we report the caption with highest CIDEr score from all the sampled captions (# samples = 20 or 100). Beam search, although obtaining the highest CIDEr score, is known to be extremely slow and significantly less diverse.

time-dependent mean $\mu_t^T(z_{<t}, x_{<t}, I)$ and standard deviation $\sigma_t^T(z_{<t}, x_{<t}, I)$ obtained from an LSTM net. The LSTM net input z_{t-1}^T and x_{t-1}^T directly influence μ_t^T and σ_t^T. Dependence on $z_{<t-1}$ and $x_{<t-1}$ is encoded via the hidden representation h_{t-1}^T. Dependence on the image is encoded into the LSTM net via an image embedding obtained from the fc7 layer of a VGG16 network [35], pre-trained on the ImageNet dataset. The 512 dimensional image embedding is fed as input at every time step of the intention model LSTM, concatenated with the output from the previous time step and the word embedding of the previous word. The image embedding, the word embedding and the latent vector are all 512-dimensional.

During inference, at time t we use a sample z_t from the modeled Gaussian with mean $\mu_t^T(z_{<t}, x_{<t}, I)$ and standard deviation $\sigma_t^T(z_{<t}, x_{<t}, I)$ as input for the decoder. However, during training, as illustrated in Fig. 2(c), we use a sample from the encoder. This discrepancy is justified by the fact that part of the training objective given in Eq. (1) maximizes the negative KL-divergence

$$\sum_t -\mathbb{E}_{z_t \sim q_\phi(z_t|z_{t-1},x,I)}\left[\ln \frac{q_\phi(z_t|z_{t-1},x,I)}{p_\theta(z_t|z_{<t},x_{<t},I)}\right]$$

between the intention model and the encoder at each time-step. This is highlighted in Fig. 2(c). Therefore, upon training we want those distributions to be adequately close. This ensures that samples used during testing are suitable.

More importantly however, note that the encoder $q_\phi(z_t|z_{t-1}, x, I)$ depends on the entire sentence x while the intention model $p_\theta(z_t|z_{<t}, x_{<t}, I)$ only depends on the past observations $x_{<t}$. Consequently, if we construct an adequate encoder and if $p_\theta(z_t|z_{<t}, x_{<t}, I)$ is accurate, we are able to capture the intention about how to complete the sentence using samples from the intention model. We discuss an encoder structure that yielded encouraging results next.

	Method	Distinct Captions	# Novel Sentences	mBleu-4	n-gram Diversity	
					Div-1	Div-2
Beam size #samples: 20	Beam search	**100%**	2317	0.77	0.21	0.29
	Div-BS [38]	**100%**	3106	0.81	0.20	0.26
	AG-CVAE [40]	69.8%	3189	0.66	0.24	0.34
	POS [10]	96.3%	3394	0.64	0.24	0.35
	Seq-CVAE	94.0%	**4266**	**0.52**	**0.25**	**0.54**
Beam size #samples: 100	Beam search	**100%**	2299	0.78	0.21	0.28
	Div-BS [38]	**100%**	3421	0.82	0.20	0.25
	AG-CVAE [40]	47.4%	3069	0.70	0.23	0.32
	POS [10]	91.5%	3446	0.67	0.23	0.33
	Seq-CVAE	84.2%	**4215**	**0.64**	**0.33**	**0.48**
5	Human	99.8%	-	0.51	0.34	0.48

Table 2: **Diversity Statistics.** We report the number of novel sentences (sentences never seen during training) for each method. Beam search and diverse beam search (Div-BS) produce the least number of novel sentences. POS [10] uses additional information in the form of part-of-speech tokens and object detections from Faster-RCNN [32]. AG-CVAE [40] also uses additional information in the form of object vectors. Our **Seq-CVAE** with ELMo doesn't use any additional information during inference and produces 4278/5000 novel sentences. Our method also yields significant improvements on 2-gram diversity, producing $\approx 20\%$ more unique 2-grams for 20 samples and a $\approx 15\%$ improvement for 100 samples when compared to the runner-up, *i.e.*, POS [10]. The model also provides the lowest m-Bleu-4, which shows that for each image the diverse captions are most different from each other. Beam search has the highest m-Bleu-4, which shows that all the distinct captions don't differ from each other at many word locations.

3.4. Encoder to Expose Intention

To adequately encode the intention, *i.e.*, the future of a sentence, we need to construct a model which contains at time t information about the entire sentence rather than only its past. To achieve this, we develop a two-stage encoder $q_\phi(z_t|z_{t-1}, x, I)$ which models at time t a Gaussian distribution with mean $\mu_t^E(z_{t-1}, x, I)$ and standard deviation $\sigma_t^E(z_{t-1}, x, I)$ modulated by multiplying with the exponentiated function value $F_\phi(\mu_t^E, x, I) \in \mathbb{R}$, *i.e.*,

$$q_\phi(z_t|z_{t-1}, x, I) \propto \mathcal{N}(z_t|\mu_t^E, \sigma_t^E) \cdot \exp F_\phi(\mu_t^E, x, I).$$

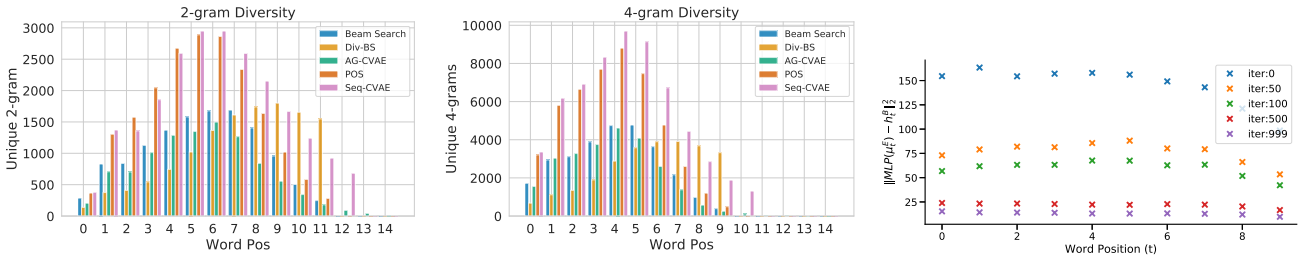
Note that multiplication with the exponentiated function value $F_\phi(\mu_t^E, x, I)$ doesn't change the fact that q_ϕ is a valid distribution over the latent space. Importantly however, multiplication permits to add the term $-F_\phi(\mu_t^E, x, I)$ to the objective given in Eq. (1). As detailed below, we will use F_ϕ to encourage μ_t^E to better capture the future of a sentence.

The first stage of the encoder captures the past via a classical forward LSTM net with hidden states referred to as h_t^F. The second stage of the encoder captures the future via a backward LSTM net with hidden states referred to as h_t^B. We subsequently combine both via a multi-layer perceptron (MLP) net which yields mean $\mu_t^E(z_{t-1}, x, I)$ and standard deviation $\sigma_t^E(z_{t-1}, x, I)$ of the Gaussian distribution.

To ensure that the mean $\mu_t^E(z_{t-1}, x, I)$ more closely resembles the information obtained from the backward pass we choose

$$F(\mu_t^E, x, I) = \lambda \|g(\mu_t^E(z_{t-1}, x, I)) - h_t^B\|_2^2, \qquad (2)$$

where g is another MLP which maps μ_t^E to fit h_t^B. The latter



(a) (b) (c)

Figure 3: (a & b): n-gram diversity across word positions. **Seq-CVAE** improves significantly upon many baselines. (c): L_2 distance between the ELMo hidden state and the representation inferred by passing the encoder mean μ_t^E through an MLP. The latter matches h_t^B better as the training progresses. This indicates that the latent space learned by the encoder at a given time t is trained to better regress to word representations which summarize future words.

is 512-dimensional in our case. λ is a hyper-parameter set to $5e^{-4}$. For the backward LSTM we use the pre-trained ELMo [31] model, with a hidden dimension of 512. ELMo is a deep bidirectional language model trained on 1 Billion Word Language Model Benchmark [5]. We only use the backward part of the model. Word representations taken from this backward pass at any time t are a good encoding of the future $x_{>t}$. ELMo is not fine-tuned through training.

4. Results

In the following, we first describe the dataset along with the competitive baselines for diverse captioning and the evaluation metrics used. We then present our results.

Dataset: We use the challenging **MSCOCO** dataset [27] for our experiments. Following the approach in [10, 40], we perform our analysis on the split of M-RNN [30] which has 118,287 train, 4,000 val and 1,000 test images. Additional results are deferred to the supplementary material.

Methods: We refer to our proposed approach as **Seq-CVAE**. We also provide results of our proposed approach without the ELMo based backward LSTM and without the data-dependent intention model.

We compare the results to the diverse captioning approach of [10] which uses part-of-speech as a prior and refer to the method via **POS**. We also compare to the additive Gaussian conditional VAE-based diverse captioning method of Wang *et al*. [40], denoted by **AG-CVAE** which uses object detections as additional information. Moreover, we compare to beam search denoted as **Beam search** and diverse beam search [38] referred to as **Div-BS** applied to standard captioning methods based on convolutions [2] and LSTM nets [22].

Evaluation criteria: We compare all aforementioned methods via the following accuracy and diversity metrics:

- **Accuracy.** In Sec. 4.1, we report the Top-1-Accuracy evaluated on the standard image captioning metrics (Bleu-1 to Bleu-4, CIDEr, ROUGE, METEOR and SPICE, each abbreviated with its initial).
- **Diversity.** Our diversity evaluation is presented in Sec. 4.2

4.1. Top-1-Accuracy

We use CIDEr as an oracle to pick the top-1 caption from a set of generated diverse captions for a given image.

Following the approach of Deshpande *et al*. [10] and Wang *et al*. [40], the top-1 caption is chosen as the caption with the maximum score calculated with the ground-truth test captions as references. The oracle metric provides the maximally possible top-1 accuracy that a given model can achieve. In Tab. 1 we show that our **Seq-CVAE** performs on par with the best baselines on the M-RNN split [30].

Specifically, in Tab. 1 we show for 20 and 100 samples, that the proposed approach obtains a CIDEr score of 1.448 and 1.695 respectively, on par with POS. Importantly, we emphasize that the proposed approach doesn't use any additional information in the form of part-of speech tags or object vectors from a Faster-RCNN [32] during inference. Note that all the other scores, are comparable to the POS [10] approach while improving upon the AG-CVAE [40] method. The latter is the only other VAE based method which exhibits stochasticity when producing diverse captions. Note that although beam search obtains the best scores, it is known to be slow and less diverse as shown in Sec. 4.2.

4.2. Diversity Evaluation

To ensure comparability to the baselines, our diversity numbers are calculated on the M-RNN split [30]. In Tab. 2 we show the diversity results using the following metrics:

(1) Uniqueness. The number of distinct captions generated by sampling from the latent space. We show that the proposed method produces 18.48/20 (92.4%) and 80.9/100 (80.9%) unique sentences. Note that beam search and Div-BS are deterministic and are guaranteed to generate 100% unique captions. Similarly, POS is completely deterministic and ensures a large number of unique captions via a strong connection between generated words and a hard-coded 'latent space' which depends on part-of-speech tags and is learned in a fully supervised manner. In contrast, AG-CVAE, just like the proposed approach, has a flexible latent space. Compared to AG-CVAE, the proposed approach generates significantly more distinct captions.

Method	ELMo	Distinct Captions	# Novel Sentences	mBleu-4	n-gram Diversity		CIDEr
					Div-1	Div-2	
POS	-	**96.3%**	3394	.64	.24	.35	**1.468**
Z-forcing [17]	✓	47.7%	**4361**	.79	**.25**	.37	1.140
CVAE	×	12.1%	1991	.52	.16	.29	0.959
CVAE	✓	11.9%	1923	.51	.25	.29	0.952
Seq-CVAE+$\mathcal{N}$	×	19.7%	2888	.63	.24	.35	1.057
Seq-CVAE+$\mathcal{N}$	✓	52.8%	4162	.69	.25	.43	1.244
Seq-CVAE(BRNN)	×	91.8%	4267	.65	**.25**	**.52**	1.348
Seq-CVAE	✓	94.0%	4266	**.52**	**.25**	**.54**	1.448
Human	-	99.8%	-	.510	.34	.48	-

Table 3: Diversity and best-1 oracle accuracy on M-RNN test split for different models calculated using top-5 captions and consensus reranking.

(2) Novel Sentences. Novel sentences are those sentences which were never observed in the training data. We see that **Seq-CVAE** produces significantly more novel sentences than any other baseline. This is remarkable and illustrates the ability to emit novel words that form reasonable sentences, particularly when considering that accuracy metrics are on par with the best performing baselines. In Tab. 2, we show that our approach produces >4000 novel captions among the 5000 captions chosen. We choose the top-5 generated captions per image, ranked by CIDEr, using consensus re-ranking following the approach in [11, 40].

(3) Mutual Overlap – (mBleu-4). m-Bleu-4 measures the difference between predicted diverse captions. Specifically, for a given image, we calculate the Bleu-4 metric for every one of the K diverse captions w.r.t. the remaining $K-1$ and average across all test images. A lower value of m-Bleu indicates more diversity. Again we observe that the proposed approach significantly improves upon all baselines. As before, we use top-5 generated captions, ranked by CIDEr, using consensus re-ranking [11, 40].

(4) n-gram diversity – (Div-n): For Div-n scores, we measure the ratio of distinct n-grams to the total number of words generated per set of diverse captions. Higher is better. Again we observe that our approach significantly improves upon the baselines, particularly when considering 2-grams. For instance, we improve from 0.35 to 0.54 when considering 20 samples and from 0.33 to 0.48 when considering 100 samples. This is encouraging because it illustrates the ability of our approach to produce fitting yet diverse descriptions without using any additional information.

(5) Unique n-grams. We measure the unique 2-grams and the unique 4-grams produced by our model in Fig. 3(a, b). We observe that our model produces the largest number of 4-grams for all word positions until position 8. We produce a comparable number of unique 2-grams as POS [10]. To compute the numbers we use 20 samples from the latent space for each of the 1000 test images. This higher number of unique 4-grams, indicates that a model is not just producing unique words, but also unique combinations of words.

4.3. Ablation Study

(1) Is ELMo the reason for good performance? To analyze if the improvements are only due to a strong language model like ELMo, we replaced ELMo with a backward RNN trained on MSCOCO training data. The performance in both diversity metrics and CIDEr, are comparable to our ELMo based model, indicating the high performance gains are not just from using a strong pretrained language model (Tab. 3 row 7, Seq-CVAE(BRNN)).

Method	ELMo	Intention	Latent	C@20	C@100
Seq-CVAE	×	$\mathcal{N}$	512	1.015 ± 0.002	1.082 ± 0.001
Seq-CVAE	×	$z_t \vert z_{<t}$		1.016 ± 0.002	1.089 ± 0.002
Seq-CVAE	✓	$z_t \vert z_{<t}$	512	1.332 ± 0.002	1.568 ± 0.004
Seq-CVAE	✓	$z_t \vert z_{<t}, x_{<t}$		1.332 ± 0.002	1.573 ± 0.002
Seq-CVAE	×	$z_t \vert z_{<t}$		0.998 ± 0.014	1.059 ± 0.017
Seq-CVAE	✓	$z_t \vert z_{<t}$	256	1.339 ± 0.003	1.559 ± 0.003
Seq-CVAE	✓	$z_t \vert z_{<t}, x_{<t}$		1.335 ± 0.002	1.575 ± 0.004
Seq-CVAE	×	$z_t \vert z_{<t}$		0.957 ± 0.001	1.000 ± 0.002
Seq-CVAE	✓	$z_t \vert z_{<t}$	128	1.328 ± 0.008	1.544 ± 0.006
Seq-CVAE	✓	$z_t \vert z_{<t}, x_{<t}$		1.324 ± 0.006	1.571 ± 0.002

Table 4: **Ablation Analysis.** We observe that the ELMo based representation improves the oracle CIDEr @100 by ∼0.5. Using ELMo along with a data dependent intention model gives the best performance with CIDEr ∼1.573. The low value of the standard deviation calculated over 10 runs for all the models is indicative that the learned latent space is robustly structured. Using a constant Gaussian intention model ($\mathcal{N}$) performs on par with using a parametric LSTM based intention model without ELMo, clearly showing the efficacy of the proposed approach.

(2) Using a single latent variable. Accuracy and diversity drop when using a single z (both with and without ELMo) to encode the entire sentence (Tab. 3 rows 3, 4; CVAE), due to posterior collapse. Also, the latent space differs per word (Fig. 5, Fig. 6). A single z doesn't efficiently encode this.

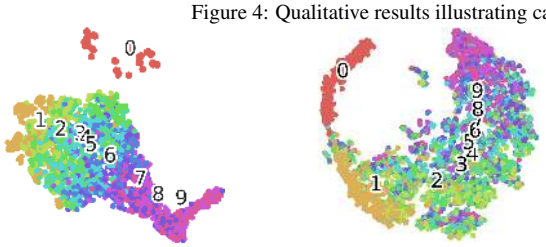
(3) Using a single LSTM for Encoder, Decoder and Transition Network. Different distributions are best served by their own individual representation. Following the approach in [17] using the same LSTM for all networks leads to inferior results (Tab. 3 row 2, Z-forcing).

(4) Using a constant Gaussian distribution per word. Replacing the LSTM based learnable intention model with a constant Gaussian reduces performance (both with and without ELMo) indicating the importance to distill intent via the backward LSTM into a sequential latent space (Tab. 3 rows 5, 6; Seq-CVAE+$\mathcal{N}$).

(5) Conditioning over different z and x. The results are summarized in Tab. 4 and averaged over 10 runs. We show results without using the ELMo based backward LSTM in the encoder (see column titled ELMo), without using a data-dependent intention model ($z_t|z_{<t}$, *i.e.*, the intention model isn't conditioned on $x_{<t}$), and without using any intention (*i.e.*, the intention model is a zero mean unit variance Gaussian $\mathcal{N}$ for all word positions t). Note, based on the CIDEr metric, data dependent intention doesn't contribute much when sampling 20 captions. However, data dependent intention has a slight edge when sampling 100 captions, irrespective of the chosen latent dimension. Note that the standard deviations shown in Tab. 4 are fairly small.

Image	Seq-CVAE	POS	AG-VAE	Div-BS	Beam Search
	• a cat is sitting on a suitcase on a bed • cat sitting on a piece of luggage • a small cat sitting on the back of a suitcase	• a cat is sitting in a suitcase on a bed • a cat sitting on top of a suitcase on a bed • a cat that is laying down on a suitcase	• a small gray and white cat sitting in a suitcase • a white and white cat with a suitcase • a cat sitting on a piece of luggage.	• a black and white cat is sitting in a suitcase • a black and white cat sitting on top of a piece of luggage • a cat is sitting in a suitcase	• a black and white cat is laying in a suitcase • a close up of a cat laying on a luggage bag • a cat that is laying down on a suitcase
	• the birds are swimming in the water and one is on the top • two birds are standing in the water and drinking • a group of birds on some water near water	• two white birds are standing in the water • two large white birds standing in the water • two birds are standing in the water together	• a large white and black bird in a body of water • a white and white bird standing on top of a body of water • a red and white photo of some birds in a pond.	• a couple of birds standing on top of a river • a couple of birds standing on top of a pond • a couple of birds that are standing next to each other	• couple of birds standing in the water • a couple of birds that are standing in the water • a couple of birds standing on a body of water

Figure 4: Qualitative results illustrating captions obtained from different image captioning methods.



(a) (b)

Figure 5: t-SNE plots of the means μ_t^T obtained from the intention model, learned **with ELMo (a)** and **without ELMo (b)**. Notice that with ELMo representation, the model better disentangles the means per word.

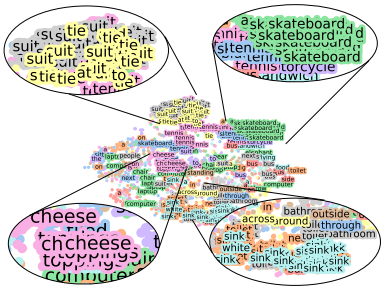


Figure 6: t-SNE plot of the means μ_t^T learned by the intention model, mapped to the words produced by the decoder. Notice that similar words like 'suit,' 'cheese,' *etc*. are grouped in tight clusters.

4.4. Latent Space Analysis

To further understand the intricacies of the learned latent space we analyze its behavior in the following.

In Fig. 3(c) we illustrate for different training iterations (see legend) and word positions t the averaged $F(\mu_t^E, x, I)$ given in Eq. (2), *i.e*., the L2 distance between the ELMo representation h_t^B obtained via the backward LSTM and the ELMo representation inferred by passing the encoder mean μ_t^E through the MLP g. Intuitively we observe models at later iterations to better match the ELMo representation h_t^B.

To further investigate whether the mean μ_t^T of the intention model used during inference captures meaningful transitions, we illustrate in Fig. 5(a, b) t-SNE [29] plots of means obtained from different images and colored by word position t. We can clearly observe that the word positions of μ_t^T are much better grouped when using ELMo representation (Fig. 5 (a)) whereas they are more cluttered when training **Seq-CVAE** without ELMo representations. We verified this analysis across multiple runs.

α		
0.0	a horse is running down a dirt path	a man and woman are playing video games together
0.2	a horse is being led by a man on a horse	people are sitting on a couch with their feet up to the wii
0.4	a man riding a horse through a field	a group of people are sitting on a couch with a wii remote in
0.6	a woman walking across a dirt field with a horse	people are sitting on a couch while playing a video game
0.8	a man riding a horse on a dirt field	a group of people are playing a video game
1.0	a woman riding a horse on a dirt field	a group of people playing a video game together

Figure 7: Diversity of sentences controlled by linear interpolation between two samples. We observe meaningful sentences across all interpolated positions. Here α is the coefficient of linear interpolation.

In Fig. 6 we illustrate a t-SNE plot of μ_t^T based on words emitted by the decoder. We clearly observe clusters of words like 'woman,' 'man,' 'dog,' 'horse,' 'group,' 'bathroom,' 'toilet,' *etc*. This grouping is encouraging as it illustrates how we can control individual emitted words by transitioning from one representation to another. Results for this transition are illustrated in Fig. 1.

4.5. Qualitative Results

We show a transition between two sampled captions in Fig. 7. We linearly interpolate the latent vectors at all word positions between two sampled descriptions.

5. Conclusion

We propose **Seq-CVAE** which learns a word-wise latent space that captures the future of the sentence, *i.e*., the 'intention' about how to complete the image description. This differs from existing techniques which generally learn a single latent space to initialize sentence generation or to identically bias word generation throughout the process. We demonstrate the proposed approach on the standard dataset and illustrate results on par w.r.t. baseline accuracies while significantly improving a large variety of diversity metrics.

Acknowledgments. Supported in part by NSF grant 1718221, Samsung, and 3M. We thank NVIDIA for GPUs.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *Proc. CVPR*, 2017. 2

[2] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional image captioning. In *Proc. CVPR*, 2018. 2, 6

[3] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *JMLR*, 2003. 1, 2

[4] Moitreya Chatterjee and Alexander G. Schwing. Diverse and coherent paragraph generation from images. In *Proc. ECCV*, 2018. 2

[5] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013. 6

[6] Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proc. CVPR*, 2015. 1, 2

[7] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Trans. on Multimedia*, 2015. 1, 2

[8] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Proc. NIPS*, 2015. 3

[9] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proc. ICCV*, 2017. 1, 2

[10] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G. Schwing, and David Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proc. CVPR*, 2019. 1, 2, 5, 6, 7

[11] Jacob Devlin, Saurabh Gupta, Ross B. Girshick, Margaret Mitchell, and C. Lawrence Zitnick. Exploring nearest neighbor approaches for image captioning. *CoRR abs/1505.04467*, 2015. 7

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. CVPR*, 2015. 1, 2

[13] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proc. CVPR*, 2015. 1, 2

[14] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Proc. ECCV*, 2010. 1, 2

[15] Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential neural models with stochastic layers. In *Proc. NIPS*, 2016. 3

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. In *Proc. NIPS*, 2014. 1

[17] Anirudh Goyal ALIAS PARTH Goyal, Alessandro Sordoni, Marc-Alexandre Côté, Nan Rosemary Ke, and Yoshua Bengio. Z-forcing: Training stochastic recurrent networks. In *Proc. NIPS*, 2017. 3, 7

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 1, 2

[19] Unnat Jain, Svetlana Lazebnik, and Alexander G. Schwing. Two can play this game: visual dialog with discriminative question generation and answering. In *Proc. CVPR*, 2018. 2

[20] Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. Creativity: Generating diverse questions using variational autoencoders. In *Proc. CVPR*, 2017. 2

[21] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proc. CVPR*, 2016. 1, 2

[22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, 2015. 1, 2, 6

[23] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Proc. NIPS*, 2014. 1

[24] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 1, 2

[25] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *PAMI*, 2013. 1, 2

[26] Dianqi Li, Xiaodong He, Qiuyuan Huang, Ming-Ting Sun, and Lei Zhang. Generating diverse and accurate visual captions by comparative adversarial learning. *arXiv preprint arXiv:1804.00861*, 2018. 1, 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, 2014. 2, 6

[28] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. *arXiv preprint arXiv:1612.00370*, 2016. 2

[29] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 8

[30] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-rnn). In *Proc. ICLR*, 2015. 1, 2, 6

[31] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. NAACL*, 2018. 6

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. NIPS*, 2015. 2, 5, 6

[33] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *Proc. CVPR*, 2017. 2

[34] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proc. ICCV*, 2017. 1, 2

[35] Karen Simonyan and Andrew Zisserman. Very deep convo-

lutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 5

[36] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2014. 1, 2

[37] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Proc. NIPS*, 2015. 1

[38] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *Proc. AAAI*, 2018. 1, 2, 5, 6

[39] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proc. CVPR*, 2015. 1, 2

[40] Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Proc. NIPS*, 2017. 1, 2, 5, 6, 7

[41] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, 2015. 1, 2

[42] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proc. ICCV*, 2017. 2