

# Creativity: Generating Diverse Questions using Variational Autoencoders

Unnat Jain\*  
 UIUC

uj2@illinois.edu

Ziyu Zhang\*  
 Northwestern University

z Zhang@u.northwestern.edu

Alexander Schwing  
 UIUC

aschwing@illinois.edu

## Abstract

Generating diverse questions for given images is an important task for computational education, entertainment and AI assistants. Different from many conventional prediction techniques is the need for algorithms to generate a diverse set of plausible questions, which we refer to as “creativity”. In this paper we propose a creative algorithm for visual question generation which combines the advantages of variational autoencoders with long short-term memory networks. We demonstrate that our framework is able to generate a large set of varying questions given a single input image.

## 1. Introduction

Creativity is a mindset that can be cultivated, and ingenuity is not a gift but a skill that can be trained. Pushing this line of thought one step further, in this paper, we ask: why are machines and in particular computers not creative? Is it due to the fact that our environment is oftentimes perfectly predictable at the large scale? Is it because the patterns of our life provide convenient shortcuts, and creativity is not a necessity to excel in such an environment?

Replicating human traits on machines is a long-standing goal, and remarkable recent steps to effectively extract representations from data [7, 41] have closed the gap between human-level performance and ‘computer-level’ accuracy on a large variety of tasks such as object classification [39], speech-based translation [61], and language-modeling [28]. There seems no need for computers to be creative. However, creativity is crucial if existing knowledge structures fail to yield the desired outcome. We cannot hope to encode all logical rules into algorithms, or all observations into features, or all data into representations. Hence, we need novel frameworks to automatically mine and implicitly characterize knowledge databases, and we need algorithms which are creative at combining those database entries.

Generative modeling tools can be used for those tasks

\* indicates equal contributions.



Figure 1: Questions generated by our approach - *literal to inferential, i.e.*, questions following visual content and questions requiring scene understanding and prior information about objects.

since they aim at characterizing the distribution from which datapoints are sampled. Classical generative models such as restricted Boltzmann machines [27], probabilistic semantic indexing [30] or latent Dirichlet allocation [8] sample from complex distributions. Instead, in recent years, significant progress in generative modeling suggests to sample from simple distributions and to subsequently transform the sample via function approximators to yield the desired output. Variational autoencoders [36, 35] and adversarial nets [25] are among algorithms which follow this paradigm. Both, variational autoencoders and adversarial nets have successfully been applied to a variety of tasks such as image generation, sentence generation *etc.*

In this work we will use those algorithms for the novel task of visual question generation as opposed to visual question answering. Visual question generation is useful in a variety of areas where it is important to engage the user, *e.g.*, in computational education, for AI assistants, or for entertainment. Retaining continued interest and curiosity is crucial in all those domains, and can only be achieved if the developed system is continuously exposing novel aspects rather than repeating a set of handcrafted traits. Visual question generation closes the loop to question-answering and diverse questions enable engaging conversation, helping AI systems such as driving assistants, chatbots, *etc.*, to perform better on Turing tests. Concretely, consider a program that aims at teaching kids to describe images. Even if

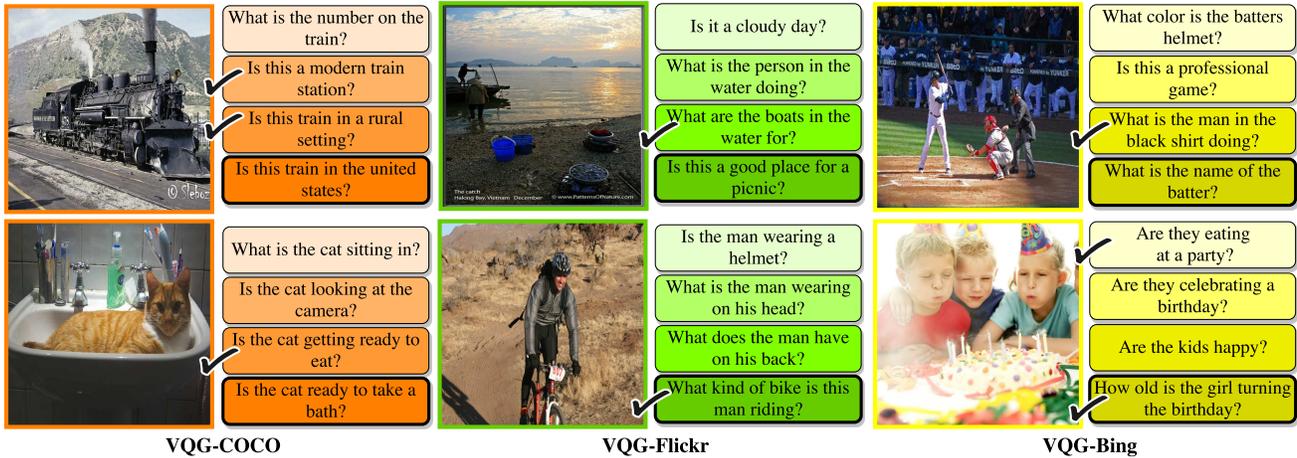


Figure 2: Examples of questions generated by our VQG algorithm. Darker colored boxes contain questions which are more inferential. Our questions include queries about numbers and scanty clouds showing its visual recognition strength. Questions on events, type of sport and motion demonstrate an ability to understand scenes and actions. Unlike questions on colors, counts and shapes, the questions in **bold box** are exemplars of how diverse our model is. It fuses visual information with context to ask questions which cannot be answered simply by looking at the image. Its answer requires prior (human-like) understanding of the objects or scene. The questions with **bold ticks** (✓) are generated by our VQG model which never occurred during training (what we refer to as ‘unseen’ questions).

100 exciting questions are provided, the program will eventually exhaust all of them and we quickly put the program aside mocking about repetitive ‘behavior.’ To alleviate this issue, we argue that creative mechanisms are important, particularly in domains such as question generation.

In this paper we propose a technique for generating diverse questions that is based on generative models. More concretely, we follow the variational autoencoder paradigm rather than adversarial nets, because training seems oftentimes more stable. We learn to embed a given question and the features from a corresponding image into a low-dimensional latent space. During inference, *i.e.*, when we are given a new image, we generate a question by sampling from the latent space and subsequently decode the sample together with the feature embedding of the image to obtain a novel question. We illustrate some images and a subset of the generated questions in Fig. 1 and Fig. 2. Note the diversity of the generated questions some of which are more literal while others are more inferential.

In this paper we evaluate our approach on the VQG - COCO, Flickr and Bing datasets [50]. We demonstrate that the proposed technique is able to ask a series of remarkably diverse questions given only an image as input.

## 2. Related Work

When considering generation of text from images, caption and paragraph generation [32, 4, 11, 18, 19, 20, 13, 33, 37, 40, 48, 60, 64, 69], as well as visual question answering [2, 22, 47, 52, 57, 67, 68, 71, 21, 34, 74, 1, 15, 31, 72, 73, 66, 46, 45, 31] come to mind immediately. We first re-

view those tasks before discussing work related to visual question generation and generative modeling in greater detail.

**Visual Question Answering and Captioning** are two tasks that have received a considerable amount of attention in recent years. Both assume an input image to be available during inference. For visual question answering we also assume a question to be provided. For both tasks a variety of different models have been proposed and attention mechanisms have emerged as a valuable tool because they permit to catch a glimpse on what the generally hardly interpretable neural net model is concerned about.

**Visual Question Generation** is a task that has been proposed very recently and is still very much an open-ended topic. Ren *et al.* [52] proposed a rule-based algorithm to convert a given sentence into a corresponding question that has a single word answer. Mostafazadeh *et al.* [50] were the first to learn a question generation model using human-authored questions instead of machine-generated captions. They focus on creating a ‘natural and engaging’ question. Recently, Vijayakumar *et al.* [63] have shown preliminary results for this task as well.

We think that visual question generation is an important task for two reasons. First, the task is dual to visual question answering and by addressing both tasks we can close the loop. Second, we think the task is in spirit similar to ‘future prediction’ in that a reasonable amount of creativity has to be encoded in the model. Particularly the latter is rarely addressed in the current literature. For example, Mostafazadeh *et al.* [50] obtain best results by generating

a single question per image using a forward pass of image features through a layer of LSTMs or gated recurrent units (GRUs). Vijayakumar *et al.* [63] show early results of question generation by following the same image caption generative model [64] as COCO-QA, but by adding a diverse beam search step to boost diversity.

Both techniques yield encouraging results. However in [50] only a single question is generated per image, while the approach discussed in [63] generates diverse questions by sampling from a complicated energy landscape, which is intractable in general [23, 5]. In contrast, in this paper, we follow more recent generative modeling paradigms by sampling from a distribution in an encoding space. The encodings are subsequently mapped to a high-dimensional representation using, in our case, LSTM nets, which we then use to generate the question.

**Generative Modeling** of data is a longstanding goal. First attempts such as k-means clustering [43] and the Gaussian mixture models [16] restrict the class of considered distributions severely, which leads to significant modeling errors when considering complex distributions required to model objects such as sentences. Hidden Markov models [6], probabilistic latent semantic indexing [30], latent Dirichlet allocation [9] and restricted Boltzmann machines [59, 27] extend the classical techniques. Those extensions work well when carefully tuned to specific tasks but struggle to model the high ambiguity inherently tied to images.

More recently deep nets have been used as function approximators for generative modeling, and, similar to deep net performance in many other areas, they produced extremely encouraging results [25, 17, 51]. Two very successful approaches are referred to as generative adversarial networks (GANs) [25] and variational auto-encoders (VAEs) [36]. However their success relies on a variety of tricks for successful training [53, 25, 51, 10].

Variational auto-encoders (VAEs) were first introduced by Kingma and Welling [36] and they were quickly adopted across different areas. They were further shown to be useful in the semi-supervised setting [35]. Conditional VAEs were recently considered by Yan *et al.* [70]. Moreover, it was also shown by Krishnan *et al.* [38] and Archer *et al.* [3] how to combine VAEs with continuous state-space models. In addition, Gregor *et al.* [26] and Chung *et al.* [14] demonstrated how to extend VAEs to sequential modeling, where they focus on RNNs.

### 3. Approach

For the task of visual question generation, demonstrated in Fig. 2, we rely on variational autoencoders (VAEs). Therefore, in the following, we first provide background on VAEs before presenting the proposed approach.

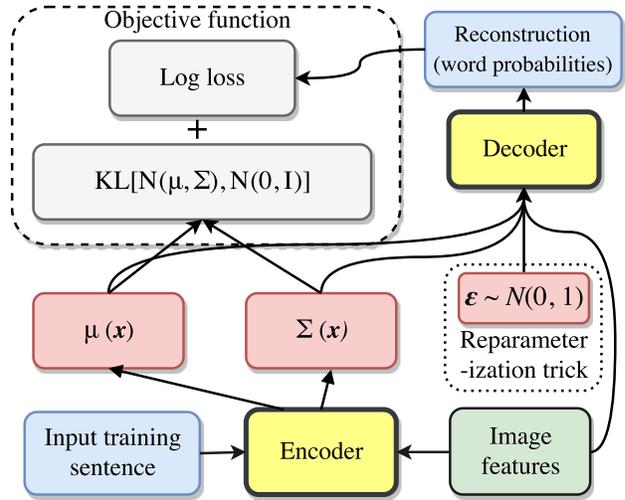


Figure 3: High level VAE overview of our approach.

### 3.1. Background on Variational Autoencoders

Following common techniques for latent variable models, VAEs assume that it is easier to optimize a parametric distribution  $p_\theta(x, z)$  defined over both the variables  $x$ , in our case the words of a sentence, as well as a latent representation  $z$ . By introducing a data-conditional latent distribution  $q_\phi(z|x)$  the log-likelihood of a datapoint  $x$ , *i.e.*,  $\ln p_\theta(x)$ , can be re-written as follows:

$$\begin{aligned} \ln p_\theta(x) &= \sum_z q_\phi(z|x) \ln p_\theta(x) \\ &= \sum_z \left[ q_\phi(z|x) \ln \frac{p_\theta(x, z)}{q_\phi(z|x)} - q_\phi(z|x) \ln \frac{p_\theta(z|x)}{q_\phi(z|x)} \right] \\ &= \mathcal{L}(q_\phi(z|x), p_\theta(x, z)) + \text{KL}(q_\phi(z|x), p_\theta(z|x)). \quad (1) \end{aligned}$$

Since the KL-divergence is non-negative,  $\mathcal{L}$  is a lower bound on the log-likelihood  $\ln p_\theta(x)$ . Note that computation of the KL-divergence is not possible because of the unknown and generally intractable posterior  $p_\theta(z|x)$ . However when choosing a parametric distribution  $q_\phi(z|x)$  with capacity large enough to fit the posterior  $p_\theta(z|x)$ , the log-likelihood w.r.t.  $\theta$  is optimized by instead maximizing the lower bound w.r.t. both  $\theta$ , and  $\phi$ . Note that the maximization of  $\mathcal{L}$  w.r.t.  $\phi$  reduces the difference between the lower bound  $\mathcal{L}$  and the log-likelihood  $\ln p_\theta(x)$ . Instead of directly maximizing the lower bound  $\mathcal{L}$  given in Eq. (1) w.r.t.  $\theta, \phi$ , dealing with a joint distribution  $p_\theta(x, z)$  can be avoided via

$$\begin{aligned} \mathcal{L}(q_\phi, p_\theta) &= \sum_z q_\phi(z|x) \ln \frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)} \\ &= \sum_z q_\phi(z|x) \ln \frac{p_\theta(z)}{q_\phi(z|x)} + \sum_z q_\phi(z|x) \ln p_\theta(x|z) \\ &= -\text{KL}(q_\phi(z|x), p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)} [\ln p_\theta(x|z)]. \quad (2) \end{aligned}$$

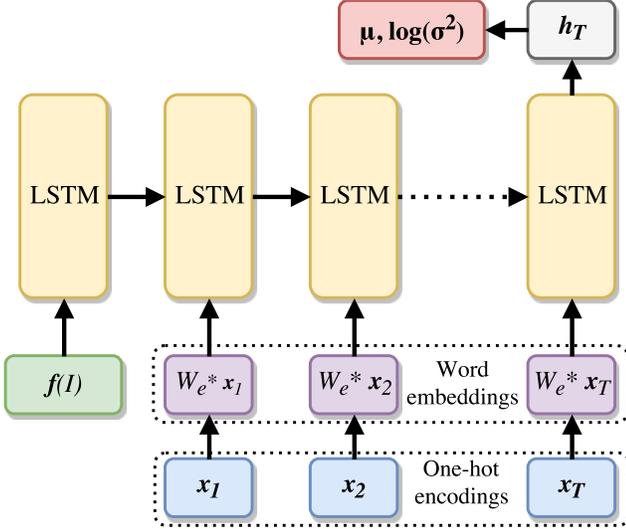


Figure 4: Q-distribution: The  $V$ -dimensional 1-hot encoding of the vocabulary (blue) gets embedded linearly via  $W_e \in \mathbb{R}^{E \times V}$  (purple). Embedding and  $F$ -dimensional image feature (green) are the LSTM inputs, transformed to fit the  $H$  dimensional hidden space. We transform the final hidden representation via two linear mappings to estimate mean and log-variance.

Note that  $p_\theta(z)$  is a prior distribution over the latent space and  $q_\phi(z|x)$  is modeling the intractable and unknown posterior  $p_\theta(z|x)$ . Intuitively the model distribution is used to guide the likelihood evaluation by focusing on highly probable regions.

In a next step the expectation over the model distribution  $q_\phi$  is approximated with  $N$  samples  $z_i \sim q_\phi$ , i.e., after abbreviating  $\text{KL}(q_\phi(z|x), p_\theta(z))$  with  $\text{KL}(q_\phi, p_\theta)$  we obtain:

$$\min_{\phi, \theta} \text{KL}(q_\phi, p_\theta) - \frac{1}{N} \sum_{i=1}^N \ln p_\theta(x|z^i), \text{ s.t. } z^i \sim q_\phi. \quad (3)$$

In order to solve this program in an end-to-end manner, i.e., to optimize w.r.t. both the model parameters  $\theta$  and the parameters  $\phi$  which characterize the distribution over the latent space, it is required to differentiate through the sampling process. To this end Kingma and Welling [36] propose to make use of the ‘reparameterization trick.’ For example, if we restrict  $q_\phi(z|x)$  to be an independent Gaussian with mean  $\mu_j$  and variance  $\sigma_j$  for each component  $z_j$  in  $z = (z_1, \dots, z_M)$ , then we can sample easily via  $z_j^i = \mu_j + \sigma_j \cdot e^i$  where  $e^i \sim \mathcal{N}(0, 1)$ . The means  $\mu_j(x, \phi)$  and variances  $\sigma_j(x, \phi)$  are parametric functions which are provided by the encoder. A general overview of VAEs is provided in Fig. 3.

### 3.2. Visual Question Generation

In the following we describe our technique for learning a high-dimensional embedding and for inference in greater

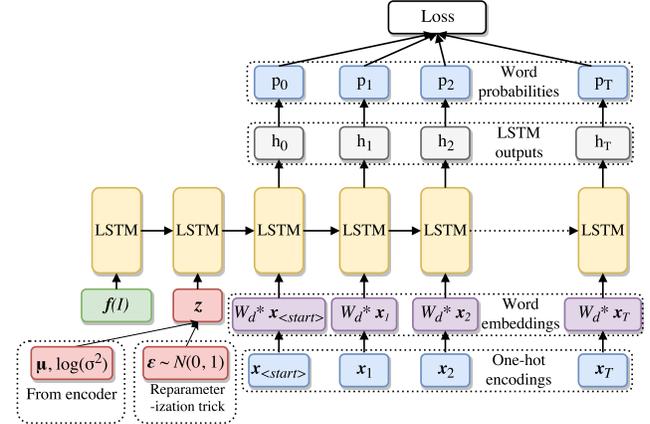


Figure 5: P-distribution: Input to the LSTM units are the  $F$ -dimensional image feature  $f(I)$ , the  $M$ -dimensional sample  $z$  (transformed during training), and the  $E$ -dimensional word embeddings. To obtain a prediction we transform the  $H$ -dimensional latent space into the  $V$ -dimensional logits  $p_i$ .

detail. We start with the learning setting before diving into the details regarding inference.

**Learning:** As mentioned before, when using a variational autoencoder, choosing appropriate  $q$  and  $p$  distributions is of crucial importance. We show a high-level overview of our method in Fig. 3 and choose LSTM models for the encoder ( $q$ -distribution) and decoder ( $p$ -distribution). Learning amounts to finding the parameters  $\phi$  and  $\theta$  of both modules. We detail our choice for both distributions in the following and provide more information regarding the trainable parameters of the model.

**Q-distribution:** The  $q$ -distribution encodes a given sentence and a given image signal into a latent representation. Since this embedding is only used during training we can assume images and questions to be available in the following. Our technique to encode images and questions is based on long short-term memory (LSTM) networks [29]. We visualize the computations in Fig. 4.

Formally, we compute an  $F$ -dimensional feature  $f(I) \in \mathbb{R}^F$  of the provided image  $I$  using a neural net, e.g., the VGG net discussed by Simonyan and Zisserman [58]. The LSTM unit first maps the image feature linearly into its  $H$  dimensional latent space using a matrix  $W_I \in \mathbb{R}^{H \times F}$ . For simplicity we neglect bias terms here and in the following.

Moreover, each  $V$ -dimensional 1-hot encoding  $x_i \in x = (x_1, \dots, x_T)$  selects an  $E$ -dimensional word embedding vector from the matrix  $W_e \in \mathbb{R}^{E \times V}$ , which is learned. The LSTM unit employs another linear transformation using the matrix  $W_{e,2} \in \mathbb{R}^{H \times E}$  to project the word embedding into the  $H$  dimensional space used inside the LSTM cells. We leave usage of more complex embeddings such as [65, 24] to future work.

Given the  $F$ -dimensional image feature  $f(I)$  and the  $E$ -dimensional word embeddings, the LSTM internally main-

Sampling	BLEU		METEOR	
	Average	Oracle	Average	Oracle
N1, 100	<b>0.356</b>	0.393	<b>0.199</b>	0.219
N1, 500	0.352	0.401	0.198	0.222
U10, 100	0.328	0.488	0.190	0.275
U10, 500	0.326	0.511	0.186	0.291
U20, 100	0.316	0.544	0.183	0.312
U20, 500	0.311	<b>0.579</b>	0.177	<b>0.342</b>

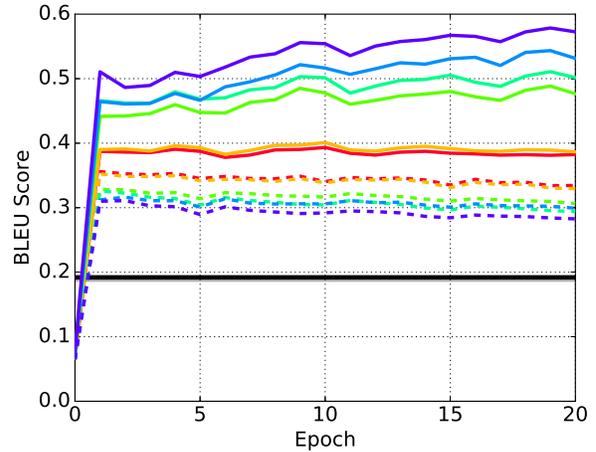
Table I: **Accuracy metrics:** Maximum (over the epochs) of average and oracle values of BLEU and METEOR metrics. Sampling the latent space by uniform distribution leads to better oracle scores. Sampling the latent space by a normal distribution leads to better average metrics. Interpretation in Sec. 4.3. Table for VQG-Flickr and VQG-Bing are similar and are included in the supplementary material.

tains an  $H$ -dimensional representation. We found that providing the image embedding in the first step and each word embedding in subsequent steps to perform best. After having parsed the image embedding and the word embeddings, we extract the final hidden representation  $h_T \in \mathbb{R}^H$  from the last LSTM step. We subsequently apply two linear transformations to the final hidden representation in order to obtain the mean  $\mu = W_\mu h_T$  and the log variance  $\log(\sigma^2) = W_\sigma h_T$  of an  $M$ -variate Gaussian distribution, *i.e.*,  $W_\mu \in \mathbb{R}^{M \times H}$  and  $W_\sigma \in \mathbb{R}^{M \times H}$ . During training a zero mean and unit variance is encouraged, *i.e.*, we use the prior  $p_\theta(z) = \mathcal{N}(0, 1)$  in Eq. (3).

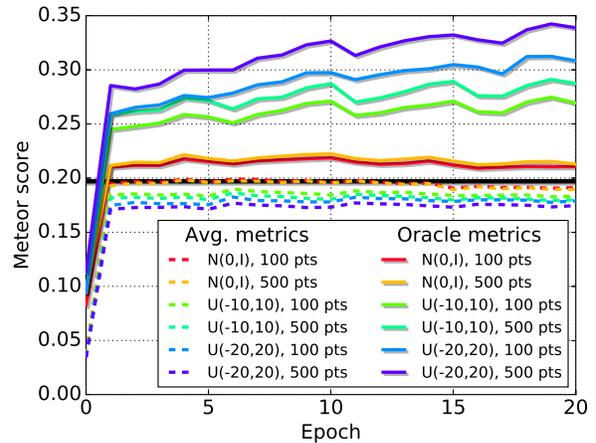
**P-distribution:** The  $p$ -distribution is used to reconstruct a question  $\hat{x}$  given, in our case, the image representation  $f(I) \in \mathbb{R}^F$ , and an  $M$ -variate random sample  $z$ . During inference the sample is drawn from a standard normal  $\mathcal{N}(0, 1)$ . During training, this sample is shifted and scaled by the mean  $\mu$  and the variance  $\sigma^2$  obtained as output from the encoder (the reparameterization trick). For the  $p$ -distribution and the  $q$ -distribution, we use the same image features  $f(I)$ , but learn a different word embedding matrix, *i.e.*, for the decoder  $W_d \in \mathbb{R}^{E \times V}$ . We observe different embedding matrices for the encoder and decoder to yield better empirical results. Again we omit the bias terms.

Analogously to the encoder we use an LSTM network for decoding, which is visualized in Fig. 5. Again we provide the  $F$ -dimensional image representation  $f(I)$  as the first input signal. Different from the encoder we then provide as the input to the second LSTM unit a randomly drawn  $M$ -variate sample  $z \sim \mathcal{N}(0, 1)$ , which is shifted and scaled by the mean  $\mu$  and the variance  $\sigma^2$  during training. Input to the third and all subsequent LSTM units is an  $E$ -dimensional embedding of the start symbol and subsequently the word embeddings  $W_d x_i$ . As for the encoder, those inputs are transformed by the LSTM units into its  $H$ -dimensional operating space.

To compute the output we use the  $H$ -dimensional hidden representation  $h_i$  which we linearly transform via a  $V \times H$ -



(a) Average-BLEU and oracle-BLEU score (Same legend as below)



(b) Average-METEOR and oracle-METEOR scores

Figure 6: **Accuracy metrics:** BLEU and METEOR scores for VQG-COCO. Experiments with various sampling procedures and results compared to the performance of the baseline model [50] (line in **black** color). VQG-Flickr and VQG-Bing results are similar and have been included in the supplementary material.

dimensional matrix into the  $V$ -dimensional vocabulary vector of logits, on top of which a softmax function is applied. This results in a probability distribution  $p_0$  over the vocabulary at the third LSTM unit. During training, we maximize the predicted log-probability of the next word in the sentence, *i.e.*,  $x_1$ . Similarly for all subsequent LSTM units.

In our framework, we jointly learn the word-embedding  $W_e \in \mathbb{R}^{E \times V}$  together with the  $V \times H$ -dimensional output embedding, the  $M \times H$ -dimensional encoding, and the LSTM projections to the  $H$ -dimensional operating space. The number of parameters (including the bias terms) in our case are  $2VE$  from the word embeddings matrix, one for the encoder and another for the decoder;  $HV + V$  as well as  $2(HM + M)$  from the output embedding of the decoder and the encoder respectively;  $(FH + H) + 2(EH + H) + (MH + H) + (HH + H)$  internal LSTM unit variables.

Sampling	Generative Strength (%)	Inventiveness (%)
N1, 100	1.98	10.76
N1, 500	2.32	12.19
U10, 100	9.82	18.78
U10, 500	16.14	24.32
U20, 100	22.01	19.75
U20, 500	<b>46.10</b>	<b>27.88</b>

Table 2: **Diversity metrics:** Maximum (over the epochs) value of generative strength and inventiveness on the VQG-COCO test set. Sampling the latent space by a uniform distribution leads to more unique questions as well as more unseen questions. Table for VQG-Flickr and VQG-Bing are similar and are included in the supplementary material.

**Inference:** After having learned the parameters of our model on a dataset consisting of pairs of images and questions we obtain a decoder that is able to generate questions given an embedding  $f(I) \in \mathbb{R}^F$  of an image  $I$  and a randomly drawn  $M$ -dimensional sample  $z$  either from a standard normal or a uniform distribution. Importantly for every different choice of input vector  $z$  we generate a new question  $x = (x_1, \dots, x_T)$ .

Since no groundtruth  $V$ -dimensional embedding is available, during inference, we use the prediction from the previous timestep as the input to predict the word for the current timestep.

### 3.3. Implementation details

Throughout, we used the 4096-dimensional  $fc6$  layer of the 16-layer VGG model [58] as our image feature  $f(I)$ , *i.e.*,  $F = 4096$ . We also fixed the 1-hot encoding of the vocabulary, *i.e.*,  $V = 10849$ , to be the number of words we collect from our datasets (VQA+VQG, detailed in the next section). We investigated different dimensions for the word embedding ( $E$ ), the hidden representation ( $H$ ), and the encoding space ( $M$ ). We found  $M = 20$ ,  $H = 512$ , and  $E = 512$  to provide enough representational power for training on roughly 400,000 questions obtained from roughly 126,000 images.

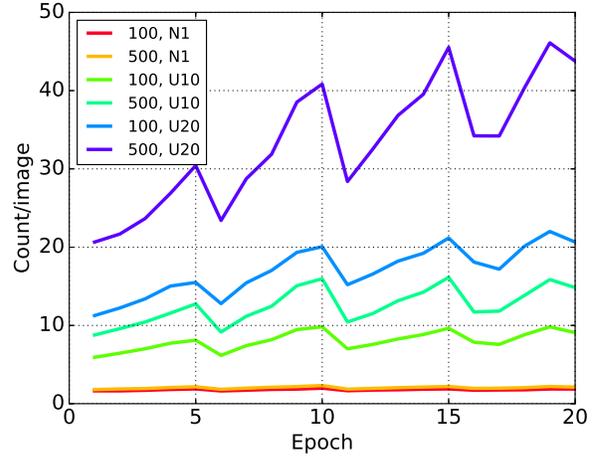
We found an initial learning rate of 0.01 for the first 5 epochs to reduce the loss quickly and to give good results. We reduce this learning rate by half every 5 epochs.

## 4. Experiments

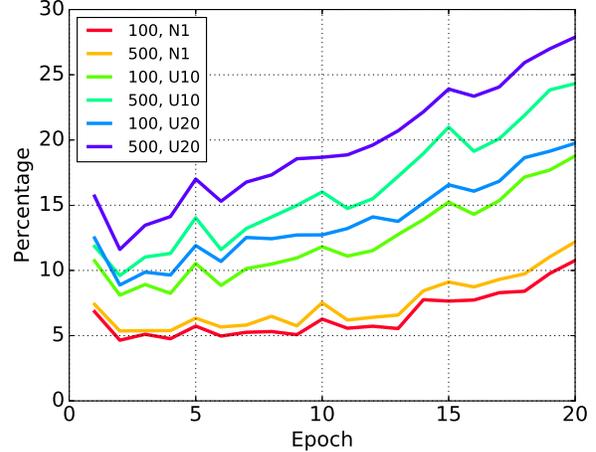
In the following we evaluate our proposed technique on the VQG dataset [50] and present a variety of different metrics to demonstrate the performance. We first describe the datasets and metrics, before providing our results.

### 4.1. Datasets:

**VQA dataset:** The images of the VQA dataset [2] are obtained from the MS COCO dataset [42], and divided into 82,783 training images, 40,504 validation images and



(a) **Generative strength:** Number of unique questions averaged over the number of images.



(b) **Inventiveness:**  $\frac{\text{Unique questions which were never seen in training set}}{\text{Total unique questions for that image}}$

Figure 7: **Diversity metrics:** Generative strength and Inventiveness, averaged over all the images in the VQG-COCO test set. VQG-Flickr and VQG-Bing results are similar and are included in the supplementary material.

40,775 testing images. Each image in the training and validation sets is annotated with 3 questions. The answers provided in the VQA dataset are not important for the problem we address.

**VQG datasets:** The Visual Question Generation [50] dataset consist of images from MS COCO, Flickr and Bing. Each of these sets consists of roughly 5,000 images and 5 questions per image (with some exceptions). Each set is split into 50% training, 25% validation and 25% test. VQG is a dataset of natural and engaging questions, which goes beyond simple literal description based questions.

The VQG dataset targets the ambitious problem of ‘natural question generation.’ However, due to its very small size, training of larger scale generative models that fit the high-dimensional nature of the problem is a challenge. Throughout our endeavor we found a question dataset size similar to

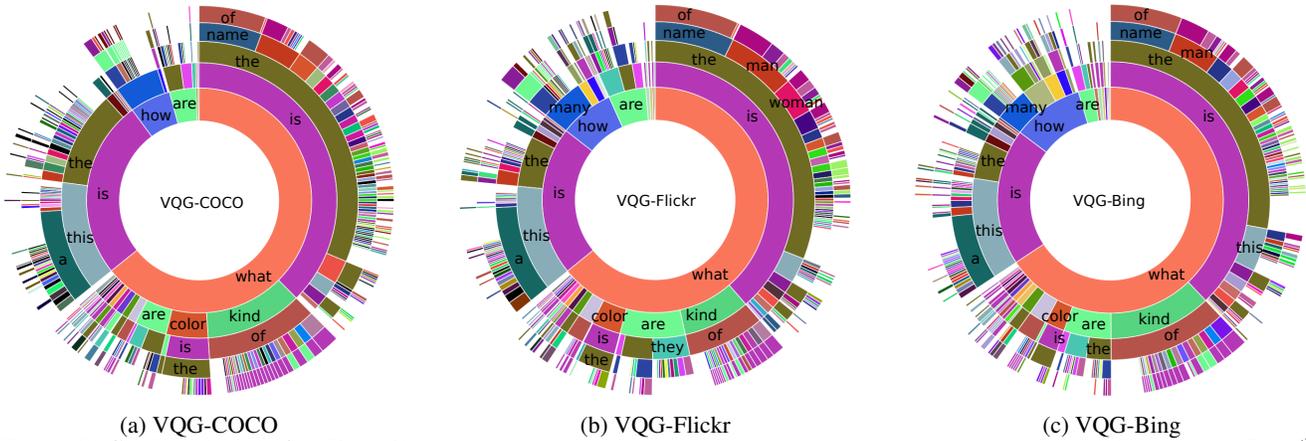


Figure 8: **Sunburst plots for diversity:** Visualizing the diversity of questions generated for each of VQG datasets. The  $i^{\text{th}}$  ring captures the frequency distribution over words for the  $i^{\text{th}}$  word of the generated question. The angle subtended at the center is proportional to the frequency of the word. While some words have high frequency, the outer rings illustrate a fine blend of words similar to the released dataset [50]. We restrict the plot to 5 rings for easy readability.

the size of the VQA dataset to be extremely beneficial.

**VQA+VQG dataset** To address this issue, we combined the VQA and VQG datasets. VQA’s sheer size provides enough data to learn the parameters of our LSTM based VAE model. Moreover, VQG adds additional diversity due to the fact that questions are more engaging and natural. The combined training set has 125,697 images (VQA training + VQA validation + VQG-COCO training - VQG-COCO validation - VQG-COCO test + VQG-Flickr training + VQG-Bing training) and a total of 399,418 questions. We ensured that there is absolutely no overlap between the images we train on and the images we evaluate. Since different images may have the same question, the number of unique questions out of *all* training question is 238,699.

## 4.2. Metrics

**BLEU:** BLEU, originally designed for evaluating the task of machine translation, was one of the first metrics that achieved good correlation with human judgment. It calculates ‘modified’  $n$ -gram precision and combines them to output a score between 0 to 1. BLEU-4 considers up to 4-grams and has been used widely for evaluation of existing works on machine translation, generating captions and questions.

**METEOR:** The METEOR score is another machine translation metric which correlates well with human judgment. An F-measure is computed based on word matches. The best among the scores obtained by comparing the candidate question to each reference question is returned. In our case there are five reference questions for each image in VQG test sets. Despite BLEU and METEOR having considerable shortcomings (details in [62]), both are popular metrics of comparison.

**Oracle-metrics:** There is a major issue in directly using machine translation metrics such as BLEU and METEOR for evaluating generative approaches for caption and ques-

tion generation. Unlike other approaches which aim to create a caption or question which is similar to the ‘reference,’ generative methods like [64, 63] and ours produce multiple diverse and creative results which might not be present in the dataset. Generating a dataset which contains all possible questions is desirable but illusive. Importantly, our algorithm may not necessarily generate questions which are only simple variations of a groundtruth question as sampling of the latent space provides the ability to produce a wide variety of questions. [64, 63] highlight this very issue, and combat it by stating their results using what [63] calls *oracle-metrics*. Oracle-BLEU, for example, is the maximum value of the BLEU score over a list of  $k$  potential candidate questions. Using these metrics we compare our results to approaches such as [50] which infer one question per image aimed to be similar to the reference question.

**Diversity score:** Popular machine translation metrics such as BLEU and METEOR provide an insight into the accuracy of the generated questions. In addition to showing that we perform well on these metrics, we felt a void for a metric which captures the *diversity*. This metric is particularly important when being interested in an engaging system. To demonstrate diversity, we evaluate our model on two intuitive metrics which could serve as relevant scores for future work attempting to generate diverse questions. The two metrics we use are average number of unique questions generated per image, and the percentage among these questions which have never been seen at training time. The first metric assesses what we call the *generative strength* and the latter represents the *inventiveness* of models such as ours.

## 4.3. Evaluation

In the following we first evaluate our proposed approach quantitatively using the aforementioned metrics, *i.e.*, BLEU score, METEOR score and the proposed diversity score. Subsequently, we provide additional qualitative results il-

illustrating the diversity of our approach. We show results for two sampling techniques, *i.e.*, sampling  $z$  uniformly and sampling  $z$  using a normal distribution.

**BLEU:** BLEU score approximates human judgment at a *corpus level* and does not necessarily correlate well if used to evaluate sentences individually. Hence we state our results for the corpus-BLEU score (similar to [50]). The best performing models presented in [50] have corpus-BLEU of 0.192, 0.117 and 0.123 for VQG-COCO, VQG-Flickr and VQG-Bing datasets respectively. To illustrate this baseline, we highlight these numbers using black lines on our plots in Fig. 6 (a).

**METEOR:** In Fig. 6 (b) we illustrate the METEOR score for our model on the VQG-COCO dataset. Similar to BLEU, we compute corpus-level scores as they have much higher correlation with human judgment. The best performing models presented in [50] have corpus-METEOR of 0.197, 0.149 and 0.162 for VQG-COCO, VQG-Flickr and VQG-Bing datasets respectively. To illustrate this baseline, we highlight these numbers using black lines on our plots in Fig. 6 (b). In Tab. 1 we compile the corpus and oracle metrics for six different sampling schemes. The sampling for results listed towards the bottom of the table is less confined. The closer the sampling scheme is to the  $\mathcal{N}(0, 1)$ , the closer is our generated corpus of questions to the reference question of the dataset. On the other hand, the more exploratory the sampling scheme, the better is the best candidate (hence, increasing oracle metrics).

**Diversity:** Fig. 7 illustrates the *generative strength* and *inventiveness* of our model with different sampling schemes for  $z$ . For the best  $z$  sampling mechanism of  $U(-20, 20)$  using 500 points, we obtained on average 46.10 unique questions per image (of which 26.99% unseen in the training set) for COCO after epoch 19; For Flickr, 59.57 unique questions on average (32.80% unseen) after epoch 19; For Bing, 63.83 unique questions on average (36.92% unseen) after epoch 15. In Tab. 2, even though the training prior over the latent space is a  $\mathcal{N}(0, 1)$  distribution, sampling from the exploratory  $U(-20, 20)$  distribution leads to better diversity of the generated questions.

To further illustrate the diversity of the generated questions we use the **sunburst** plots shown in Fig. 8 for the COCO, Flickr and Bing datasets. Despite the fact that a large number of questions start with “what” and “is,” we still observe a quite reasonable amount of diversity.

**Qualitative results:** In Fig. 2 we show success cases of our model. A range of literal to inferential questions are generated by our model, some requiring strong prior (human-like) understanding of objects and their interaction. In previous subsections we showed that our model does well on metrics of accuracy and diversity. In Fig. 9 we illustrate two categories of failure cases. *Recognition failures*, where the pre-learned visual features are incapable of capturing correctly

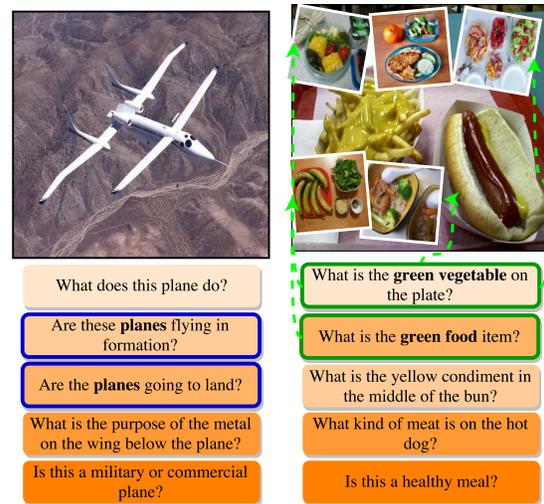


Figure 9: **Recognition and co-occurrence based failure cases:** *Left:* A special aircraft is recognized as multiple ‘airplanes’ (two sets of wings instead of one may cause the confusion), therefore, erroneous questions (marked in blue) arise. *Right:* Due to very frequent co-occurrence of green vegetable/food/fruit in food images, our VQG model generates questions (marked in green) about green vegetables even when they are missing. The five small images are few examples of how training set food images almost always contain greens.

the information required to formulate diverse questions. As illustrated by the image of a complex aircraft which appears similar to two airplanes. Hence, our system generates questions coherent to such a perception.

Second are *co-occurrence based failures*. This is illustrated using the image of fries and a hot dog. In addition to some correct questions, some questions on green food/fruit/vegetables inevitably pop up in food images (even for images without any greens). Similarly, questions about birds are generated in some non-bird images of trees. This could be accounted to very frequent co-occurrence of reference questions on greens or birds whenever an image contains food or trees, respectively.

## 5. Conclusion

In this paper we propose to combine the advantages of variational autoencoders with long-short-term-memory (LSTM) cells to obtain a “creative” framework that is able to generate a diverse set of questions given a single input image. We demonstrated the applicability of our framework on a diverse set of images and envision it being applicable in domains such as computational education, entertainment and for driving assistants & chatbots. In the future we plan to use more structured reasoning [12, 56, 44, 49, 55, 54].

**Acknowledgements:** We thank NVIDIA for providing the GPUs used for this research.

## References

- [1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Deep compositional question answering with neural module networks. In *Proc. CVPR*, 2016. 2
- [2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. In *Proc. ICCV*, 2015. 2, 6
- [3] E. Archer, I. M. Park, L. Buesing, J. Cunningham, and L. Paninski. Black box variational inference for state space models. In *ICLR Workshop*, 2016. 3
- [4] K. Barnard, P. Duygulu, D. Forsyth, N. D. Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 2003. 2
- [5] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse M-Best Solutions in Markov Random Fields. In *Proc. ECCV*, 2012. 3
- [6] L. E. Baum and T. Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Annals of Mathematical Statistics*, 1966. 3
- [7] Y. Bengio, E. Thibodeau-Laufer, G. Alain, and J. Yosinski. Deep Generative Stochastic Networks trainable by Backprop. In *JMLR*, 2014. 1
- [8] D. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *J. of Bayesian Analysis*, 2006. 1
- [9] D. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *JMLR*, 2003. 3
- [10] Y. Burda, R. Grosse, and R. R. Salakhutdinov. Importance Weighted Autoencoders. In *Proc. ICLR*, 2016. 3
- [11] X. Chen and C. L. Zitnick. Mind’s eye: A recurrent visual representation for image caption generation. In *Proc. CVPR*, 2015. 2
- [12] L.-C. Chen\*, A. G. Schwing\*, A. L. Yuille, and R. Urtasun. Learning Deep Structured Models. In *Proc. ICML*, 2015. \* equal contribution. 8
- [13] K. Cho, A. C. Courville, and Y. Bengio. Describing multimedia content using attention-based encoder-decoder networks. In *IEEE Transactions on Multimedia*, 2015. 2
- [14] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Proc. NIPS*, 2015. 3
- [15] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *EMNLP*, 2016. 2
- [16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society*, 1977. 3
- [17] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Proc. NIPS*, 2015. 3
- [18] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. CVPR*, 2015. 2
- [19] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *Proc. CVPR*, 2015. 2
- [20] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proc. ECCV*, 2010. 2
- [21] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016. 2
- [22] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? Dataset and Methods for Multilingual Image Question Answering. In *Proc. NIPS*, 2015. 2
- [23] K. Gimpel, D. Batra, G. Shakhnarovich, and C. Dyer. A Systematic Exploration of Diversity in Machine Translation. In *EMNLP*, 2013. 3
- [24] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. In *Proc. ECCV*, 2014. 4
- [25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. In *Proc. NIPS*, 2014. 1, 3
- [26] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. DRAW: A recurrent neural network for image generation. In *Proc. ICML*, 2015. 3
- [27] G. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 2006. 1, 3
- [28] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 2012. 1
- [29] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 4
- [30] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proc. SIGIR*, 1999. 1, 3
- [31] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting Visual Question Answering Baselines. In *Proc. ECCV*, 2016. 2
- [32] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *Proc. CVPR*, 2016. 2
- [33] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, 2015. 2
- [34] J.-H. Kim, S.-W. L. D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In *Proc. NIPS*, 2016. 2
- [35] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. Semi-Supervised Learning with Deep Generative Models. In *Proc. NIPS*, 2014. 1, 3
- [36] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014. 1, 3, 4
- [37] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *TACL*, 2015. 2

- [38] R. G. Krishnan, U. Shalit, and D. Sontag. Deep Kalman Filters. In *NIPS Workshop*, 2015. 3
- [39] A. Krizhevsky, I. Sutskever, , and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012. 1
- [40] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011. 2
- [41] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 2015. 1
- [42] 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. In *Proc. ECCV*, 2014. 6
- [43] S. P. Lloyd. Least squares quantization in PCM. *Trans. Information Theory*, 1982. 3
- [44] B. London\* and A. G. Schwing\*. Generative Adversarial Structured Networks. In *Proc. NIPS Workshop on Adversarial Training*, 2016. \* equal contribution. 8
- [45] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Proc. NIPS*, 2016. 2
- [46] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. In *Proc. AAAI*, 2016. 2
- [47] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proc. ICCV*, 2015. 2
- [48] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-rnn). In *ICLR*, 2015. 2
- [49] O. Meshi, M. Mahdavi, and A. G. Schwing. Smooth and Strong: MAP Inference with Linear Convergence. In *Proc. NIPS*, 2015. 8
- [50] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Generating natural questions about an image. In *Proc. ACL*, 2016. 2, 3, 5, 6, 7, 8
- [51] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 3
- [52] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Proc. NIPS*, 2015. 2
- [53] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. In *Proc. NIPS*, 2016. 3
- [54] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Globally Convergent Dual MAP LP Relaxation Solvers using Fenchel-Young Margins. In *Proc. NIPS*, 2012. 8
- [55] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Globally Convergent Parallel MAP LP Relaxation Solver using the Frank-Wolfe Algorithm. In *Proc. ICML*, 2014. 8
- [56] A. G. Schwing and R. Urtasun. Fully Connected Deep Structured Networks. In <https://arxiv.org/abs/1503.02351>, 2015. 8
- [57] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *Proc. CVPR*, 2016. 2
- [58] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4, 6
- [59] P. Smolensky. Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. MIT Press, 1986. 3
- [60] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. In *Proc. ACL*, 2014. 2
- [61] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proc. NIPS*, 2014. 1
- [62] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proc. CVPR*, 2015. 7
- [63] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. In <https://arxiv.org/abs/1610.02424>, 2016. 2, 3, 7
- [64] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proc. CVPR*, 2015. 2, 3, 7
- [65] L. Wang, Y. Li, and S. Lazebnik. Learning Deep Structure-Preserving Image-Text Embeddings. In *Proc. CVPR*, 2016. 4
- [66] Q. Wu, C. Shen, A. van den Hengel, P. Wang, and A. Dick. Image captioning and visual question answering based on attributes and their related external knowledge. In *arXiv 1603.02814*, 2016. 2
- [67] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *Proc. ICML*, 2016. 2
- [68] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Proc. ECCV*, 2016. 2
- [69] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, 2015. 2
- [70] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2Image: Conditional Image Generation from Visual Attributes. In *Proc. ECCV*, 2016. 3
- [71] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proc. CVPR*, 2016. 2
- [72] L. Yu, E. Park, A. Berg, and T. Berg. Visual madlibs: Fill in the blank image generation and question answering. In *Proc. ICCV*, 2015. 2
- [73] B. Zhou, Y. Tian, S. Sukhataar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. In *arXiv:1512.02167*, 2015. 2
- [74] C. L. Zitnick, A. Agrawal, S. Antol, M. Mitchell, D. Batra, and D. Parikh. Measuring machine intelligence through visual question answering. *AI Magazine*, 2016. 2