Automatic Generation of Grounded Visual Questions

Shijie Zhang [#1], **Lizhen Qu*** [§¶2], **Shaodi You** [§¶3], **Zhenglu Yang** [†4], **Jiawan Zhang** [‡5]
[#]School of Computer Science and Technology, Tianjin University, Tianjin, China
[§]Data61-CSIRO, Canberra, Australia
[¶]Australian National University, Canberra, Australia
[†]College of Computer and Control Engineering, Nankai University, Tianjin, China
[‡]The School of Computer Software, Tianjin University, Tianjin, China
{shijiezhang,jwzhang}@tju.edu.cn, {lizhen.qu,shaodi.you}@data61.csiro.au, yangzl@nankai.edu.cn

Abstract

In this paper, we propose the first model to be able to generate visually grounded questions with *diverse* types for a single image. Visual question generation is an emerging topic which aims to ask questions in natural language based on visual input. To the best of our knowledge, it lacks automatic methods to generate meaningful questions with various types for the same visual input.

To circumvent the problem, we propose a model that automatically generates visually grounded questions with *varying* types. Our model takes as input both images and the captions generated by a dense caption model, samples the most probable question types, and generates the questions in sequel. The experimental results on two real world datasets show that our model outperforms the strongest baseline in terms of both correctness and diversity with a wide margin.

1 Introduction

Multi-modal learning of vision and language is an important task in artificial intelligence because it is the basis of many applications such as education, user query prediction, interactive navigation, and so forth. Apart from describing visual scenes by using declarative sentences [Chen and Zitnick, 2014; Gupta and Mannem, 2012; Karpathy and Fei-Fei, 2015; Hodosh *et al.*, 2013; Kulkarni *et al.*, 2011; Kuznetsova *et al.*, 2012; Li *et al.*, 2009; Vinyals *et al.*, 2015; Xu *et al.*, 2015], recently, automatic answering of visually related questions (VQA) has also attracted a lot of attention in computer vision communities [Antol *et al.*, 2015; Malinowski and Fritz, 2014; Gao *et al.*, 2015; Ren *et al.*, 2015; Yu *et al.*, 2015; Zhu *et al.*, 2015]. However, there is little work on automatic generation of questions for images.

"The art of proposing a question must be held higher value than solving it. -Georg Canton". An intelligent system should be able to ask meaningful questions given the environment. Beyond demonstrating a high-level of AI, in practice, multi-modal question-asking modules find their use in a wide range of AI systems such as child education and dialogue systems.

Figure 1: Automatically generated grounded visual questions.

To the best of our knowledge, almost all existing VQA systems rely on manually constructed questions [Antol *et al.*, 2015; Malinowski and Fritz, 2014; Gao *et al.*, 2015; Ren *et al.*, 2015; Yu *et al.*, 2015; Zhu *et al.*, 2015]. An common assumption of the existing VQA systems is that answers are visually grounded thus all relevant information can be found in the visual input. However, the construction of such data sets are labor-intensive and time consuming, thus limits the diversity and coverage of questions being asked. As a consequence, the data incompleteness imposes a special challenge for supervised-learning based VQA systems.

In light of the above analysis, we focus on automatic generation of visually grounded questions, coined VQG. The generated questions should be grammatically well-formed, reasonable for given images, and as diverse as possible. However, the existing systems are either rule-based such that they generate questions with few limited textual patterns [Ren *et al.*, 2015; Zhu *et al.*, 2015], or they are able to ask only one question per image and the generated questions are frequently not visually grounded [Simoncelli and Olshausen, 2001].

To tackle this task, we propose the first model capable of asking questions of various types for the same image. As illustrated in Fig. 2, we first apply DenseCap [Johnson *et al.*, 2015] to construct dense captions that provides a almost complete coverage of information for questions. Then we feed these captions into the question type selector to sample the most probable question types. Taking as input the questions types, the dense captions, as well as visual features generated by VGG-16 [Simonyan and Zisserman, 2014], the question

*Corresponding author.

generator decodes all these information into questions. We conduct extensive experiments to evaluate our model as well as the most competitive baseline with three kinds of measures adapted from the ones commonly used in the tasks of image caption generation and machine translation.

The contributions of our paper are three-fold:

- We propose the first model capable of asking visually grounded questions with diverse types for a single image.
- Our model outperforms the strongest baseline up to 216% in terms of the coverage of asked questions.
- The grammaticality of the questions generated by our model as well as their relatedness to visual input also outperform the strongest baseline with a wide margin.

The rest of the paper is organized as follows: we cover the related work in Section 2, followed by presenting our model in Section 3. After introducing the experimental setup in Section 4, we discuss the results in Section 5, and draw the conclusion in Section 6.

2 Related Work

The generation of textual description for visual information has gained popularity in recent years. The key challenge is to learn the alignment between text and visual information [Barnard *et al.*, 2003; Kong *et al.*, 2014; Zitnick *et al.*, 2013]. Herein, a popular task is to describe images with a few declarative sentences, which are often referred to as image captions [Barnard *et al.*, 2003; Chen and Zitnick, 2014; Gupta and Mannem, 2012; Karpathy and Fei-Fei, 2015; Hodosh *et al.*, 2013; Kulkarni *et al.*, 2011; Kuznetsova *et al.*, 2012; Li *et al.*, 2009; Vinyals *et al.*, 2015; Xu *et al.*, 2015].

Visual Question and Answering Automatic answering of questions based on visual input is one of the most popular tasks in computer vision [Geman *et al.*, 2015; Malinowski and Fritz, 2014; Malinowski *et al.*, 2015; Pirsiavash *et al.*, 2014; Ren *et al.*, 2015; Weston *et al.*, 2015; Yu *et al.*, 2015]. Most VQA models are evaluated on a few benchmark datasets [Antol *et al.*, 2015; Malinowski and Fritz, 2014; Gao *et al.*, 2015; Ren *et al.*, 2015; Yu *et al.*, 2015; Zhu *et al.*, 2015]. The images in those datasets are sampled from the MS-COCO dataset [Lin *et al.*, 2014], the questions-answer pairs are manually constructed [Antol *et al.*, 2015; Gao *et al.*, 2015; Yu *et al.*, 2015; Zhu *et al.*, 2015].

Visual Question Generation Automatic question generation from text is explored in-depth in NLP, however, it is rarely studied for visual questions, despite of the fact that such questions are highly desired for many applications. In order to generate multiple questions per image, the most common approach is to ask human to manually build the question-answer pairs, which is labor-intensive and time-consuming [Antol *et al.*, 2015; Gao *et al.*, 2015; Malinowski and Fritz, 2014]. As one of the most recent examples, Zhu *et al.*[Zhu *et al.*, 2015] manually create questions of seven wh-question types such as what, where, when and *etc*. People also explored automatic generation of visual questions by using rules. Yu *et al.*[Yu *et al.*, 2015] consider question generation as a task of selectively removing content words that serve as answers from a caption and reformulate the resulted sentences as questions. In a similar manner, Ren *et al.*[Ren *et al.*, 2015] carefully designed a handful of rules to transform image captions into questions with limited types. However, those rule-based methods are limited by the types of questions they can generate. Apart from that, model-based methods are also studied to overcome the diversity issue, the most closed work is [Simoncelli and Olshausen, 2001], which trains an image caption model on a dataset of visual questions. However, their model cannot generate more than one question per image.

Knowledge Base (KB) based Question Answering (KB-QA) KB-QA has attracted considerable attention due to the ubiquity of the World Wide Web and the rapid development of the artificial intelligence (AI) technology. Large-scale structured KBs, such as DBpedia [Auer *et al.*, 2007], Freebase [Bollacker *et al.*, 2008], and YAGO [Suchanek *et al.*, 2007], provide abundant resources and rich general human knowledge, which can be used to respond to users' queries in open-domain question answering (QA). However, how to bridge the gap between visual questions and structured data in KBs remains a huge challenge.

The existing KB-QA methods can be broadly classified into two main categories, namely, semantic parsing based methods[Kwiatkowski *et al.*, 2013; Reddy *et al.*, 2016] and information retrieval based methods [Yao and Durme, 2014; Bordes *et al.*, 2014] methods. Most semantic parsing based methods transform a question into its meaning representation (i.e., logical form), which will be then translated to a KB query to retrieve the correct answer(s). Information retrieval based methods initially roughly retrieve a set of candidate answers, and subsequently perform an in-depth analysis to re-rank the candidate answers and select the correct ones. These methods focus on modeling the correlation of question-answer pairs from the perspective of question topic, relation mapping, answer type, and so forth.

3 Question Generation

Our goal is to generate visually grounded questions directly from images with diverse question types. We start with randomly picking a caption from a set of automatically generated captions, which describes a certain region of image with natural language. Then we sample a reasonable question type and varying the caption. In the last step, our question generator learns the correlation between the caption and the image, generates a question of the chosen type.

Formally, for each raw image $\mathbf{x}$, our model generates a set of captions $\{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_M\}$, samples a set of question types $\{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_{\hat{M}}\}$, followed by yielding a set of grounded questions $\{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_{\hat{M}}\}$. Herein, a caption or a question is a sequence of words.

$$\mathbf{w} = \{\mathbf{w}_1, ..., \mathbf{w}_L\} \tag{1}$$

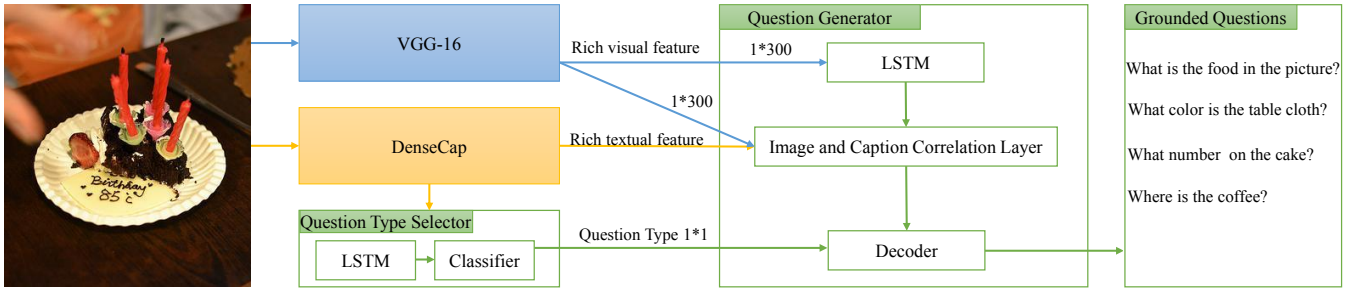


Figure 2: The proposed framework.

Where L is the length of the word sequence. Each word $\mathbf{w}_i$ employs 1-of-K encoding, where K is the size of the vocabulary. A question type is represented by the first word of a question, adopting 1-of-T encoding where T is the number of question types. The same as [Zhu *et al.*, 2015], we consider six question types in our experiments: *what*, *when*, *where*, *who*, *why* and *how*.

For each image $\mathbf{x}_i$, we apply a dense caption model (*DenseCap*) [Johnson *et al.*, 2015] trained on the Visual Genome dataset [Krishna *et al.*, 2016] to produce a set of captions $\mathcal{C}_i$. Then the generative process is described as follows:

1. Choose a caption $\mathbf{c}_n$ from $\mathcal{C}_i$.
2. Choose a question type $\mathbf{t}_n$ given $\mathbf{c}_n$.
3. Generate a question $\mathbf{q}_n$ conditioned on $\mathbf{c}_n$ and $\mathbf{t}_n$.

Denoted by $\boldsymbol{\theta}$ all model parameters, for each image $\mathbf{x}_i$, the joint distribution of $\mathbf{c}_n$, $\mathbf{t}_n$ and $\mathbf{q}_n$ is factorized as follows:

$$\begin{aligned} P(\mathbf{q}_n, \mathbf{t}_n, \mathbf{c}_n | \mathbf{x}_i, \mathcal{C}_i; \boldsymbol{\theta}) = & P(\mathbf{q}_n | \mathbf{c}_n, \mathbf{x}_i, \mathbf{t}_n; \boldsymbol{\theta}_q) \\ & P(\mathbf{t}_n | \mathbf{c}_n; \boldsymbol{\theta}_t) P(\mathbf{c}_n | \mathcal{C}_i) \end{aligned} \quad (2)$$

where $\boldsymbol{\theta} = \boldsymbol{\theta}_q \cup \boldsymbol{\theta}_t$, $P(\mathbf{q}_n | \mathbf{c}_n, \mathbf{x}_i, \mathbf{t}_n; \boldsymbol{\theta}_q)$ is the distribution of generating question, $P(\mathbf{t}_n | \mathbf{c}_n; \boldsymbol{\theta}_t)$ and $P(\mathbf{c}_n | \mathcal{C}_i)$ are the distributions for sampling question type and caption respectively. More details are given in the following sections.

Since we do not observe the alignment between captions and questions, $\mathbf{c}_n$ is latent. Sum over $\mathbf{c}$, we obtain:

$$P(\mathbf{q}_n, \mathbf{t}_n | \mathbf{x}_i, \mathcal{C}_i; \boldsymbol{\theta}) = \sum_{\mathbf{c}_n \in \mathcal{C}_i} P(\mathbf{q}_n, \mathbf{t}_n, \mathbf{c}_n | \mathbf{x}_i, \mathcal{C}_i; \boldsymbol{\theta})$$

Let $\mathcal{Q}_i$ denote the question set of the image $\mathbf{x}_i$, the probability of the training dataset $\mathcal{D}$ is given by taking the product of the above probabilities over all images and their questions.

$$P(\mathcal{D} | \boldsymbol{\theta}) = \prod_i \prod_{n \in \mathcal{Q}_i} P(\mathbf{q}_n, \mathbf{t}_n | \mathbf{x}_i, \mathcal{C}_i; \boldsymbol{\theta}) \quad (3)$$

For word representations, we initialize a word embedding matrix $\mathbf{E} \in \mathcal{R}^{300 \times K}$ by using Glove [Pennington *et al.*, 2014], which are trained on 840 billions of words. For the image representations, we apply a VGG-16 model [Szegedy *et al.*, 2015] trained on ImageNet [Deng *et al.*, 2009] without fine-tuning to produce 300-dimensional feature vectors. The dimension is chosen to match the size of the pre-trained word embeddings.

Compared to the question generation model [Simoncelli and Olshausen, 2001], which generates only one question per image, the probabilistic nature of this model allows generating questions of multiple types which refer to different regions of interests, because each caption predicted by *DenseCap* is associated with a different region.

3.1 Sample Captions and Question Types

The caption model *DenseCap* generates a set of captions for a given image. Each caption c is associated with a region and a confidence o_c of the proposed region. Intuitively, we should give a higher probability to the caption with higher confidence than the lower one. Thus, given a caption set $\mathcal{C}_i$ of an image $\mathbf{x}_i$, we define the prior distribution as:

$$P(\mathbf{c}_k | \mathcal{C}_i) = \frac{\exp(o_k)}{\sum_{j \in \mathcal{C}_i} \exp(o_j)}$$

A caption is either a declarative sentence, a word, or a phrase. We are able to ask many different types of questions but not all of them for a chosen caption. For example, for a caption "floor is brown" we can ask "what color is the floor" but it would be awkward to ask a *who* question. Thus, our model draws a question type given a caption with the probability $P(\mathbf{t}_n | \mathbf{c}_n)$ by assuming it suffices to infer question types given a caption.

Our key idea is to learn the association between question types and key words/phrases in captions. The model $P(\mathbf{t}_n | \mathbf{c}_n)$ consists of two components. The first one is a Long Short Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] that maps a caption into a hidden representation. LSTM is a recurrent neural network taking the following form:

$$\mathbf{h}_t, \mathbf{m}_t = \text{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{m}_{t-1}) \quad (4)$$

where $\mathbf{x}_t$ is the input and the hidden state of LSTM at time step t, and $\mathbf{h}_t$ and $\mathbf{m}_t$ are the hidden states and memory states of LSTM at time step t, respectively. As the representation of the whole sequence, we take the last state $\mathbf{h}_L$ generated at the end of the sequence. This representation is further fed into a softmax layer to compute a probability vector $\mathbf{p}_t$ for all question types. The probability vector characterizes a multinomial distribution of all question types.

3.2 Generate Questions

At the core of our model is the question generation module, which models $P(\mathbf{q}_n | \mathbf{c}_n, \mathbf{x}_i, \mathbf{t}_n; \boldsymbol{\theta}_q)$, given a chosen caption $\mathbf{c}_n$ and a question type $\mathbf{t}_n$. It is composed of three modules: i) an LSTM encoder to generate caption embeddings; ii) a

correlation module to learn the association between images and captions; iii) a decoder consisting of an LSTM decoder and an ngram language model.

A grounded question is deeply anchored in both the sampled caption and the associated image. In our preliminary experiments, we found it useful to let the LSTM encoder $\text{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{m}_{t-1})$ to read the image features prior to reading captions. In particular, at time step $t = 0$, we initialize the state vector $\mathbf{m}_0$ to zero and feed the image features as $\mathbf{x}_0$. At the 1st time step, the encoder reads in a special token S_0 indicating the start of a sentence, which is a good practice adopted by many caption generation models [Vinyals *et al.*, 2015]. After reading the whole caption of length L, the encoder yields the last state vector $\mathbf{m}_L$ as the embedding of caption.

The correlation module takes as input the caption embeddings from the encoder and the image features from VGG-16, produces a 300-dimensional joint feature map. We apply a linear layer of size 300×600 and a PReLU [He *et al.*, 2015] layer in sequel to learn the associations between captions and images. Since an image gives an overall context and the chosen caption provides the focus in the image, the joint representation provides sufficient context to generate grounded questions. Although the LSTM encoder incorporates image features before reading captions, this correlation module enhances the correlation between images and text by building more abstract representations.

Our decoder extends the LSTM decoder of [Vinyals *et al.*, 2015] with a ngram language model. The LSTM decoder consists of an LSTM layer and a softmax layer. The LSTM layer starts with reading the joint feature map and the start token S_0 in the same fashion as the caption encoder. From time step $t = 0$, the softmax layer predicts the most likely word given the state vector at time t yielded by the LSTM layer. A word sequence ends when the end of sequence token is produced.

Joint decoding Although the LSTM decoder alone can generate questions, we found that it would frequently produce repeated words and phrases such as "*the the*". The problem didn't disappear even the beam search [Koehn *et al.*, 2003] was applied. It is due to the fact that the state vectors produced at adjunct time steps tend to be similar. Since repeated words and phrases are rarely observed in text corpora, we discount such occurrence by joint decoding with a ngram language model. Given a word sequence $\mathbf{w} = \{w_1, ..., w_N\}$, a bigram language model is defined as:

$$P(\mathbf{w}) = \prod_{i=2}^{N} P(w_i|w_{i-1})P(w_0)$$

Instead of using neural models, we adopt the word count based estimation of model parameters. In particular, we apply the KneserNey smoothing [Kneser and Ney, 1995] to estimate $P(w_i|w_{i-1})$, which is given by:

$$\frac{\max(\text{count}(w_{i-1}, w_i) - d, 0)}{\text{count}(w_i)} + \lambda(w_{t-1})P_{KN}(w_i)$$

where $\text{count}(x)$ denotes the corpus frequency of term x, $P_{KN}(w_i)$ is a back-off statistic of unigram w_i in case the bigram (w_i, w_{i-1}) does not appear in the training corpus. The parameter d is usually fixed to 0.75 to avoid overfitting for low frequency bigrams. And $\lambda(w_{t-1})$ is a normalizing constant conditioned on w_i.

We incorporate bigram statistics with the LSTM decoder from the time step $t = 1$ because the LSTM decoder can well predict the first words of questions. The LSTM decoder essentially captures the conditional probability $P_l(\mathbf{q}_t|\mathbf{q}_{<t})$, while the bigram model considers only the previous word $P_b(\mathbf{q}_t|\mathbf{q}_{t-1})$ by using word counts. By interpolating these two, we obtain the final probability as:

$$P(\mathbf{q}_t|\mathbf{q}_{<t}) = (1 - \beta)P_l(\mathbf{q}_t|\mathbf{q}_{<t}) + \beta P_b(\mathbf{q}_t|\mathbf{q}_{t-1})$$

where $\beta \in [0, 1]$ is an interpolation weight. In addition, we fix the first k words of questions during decoding according to the chosen question types.

3.3 Training

The key challenge of training is the involvement of the latent variables indicating the alignment between captions and gold standard questions for a deep neural network. We estimate the latent variables in a similar fashion as EM but computationally more efficient.

Suppose we are given the training set $\{(\mathbf{x}_1, \mathbf{q}_1), ..., (\mathbf{x}_N, \mathbf{q}_N)\}$, the loss is given by:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{n \in \mathcal{Q}_i} -\log P(\mathbf{q}_n, \mathbf{t}_n|\mathbf{x}_i, \mathcal{C}_i; \boldsymbol{\theta})$$

Suppose $Q(\mathbf{c}_n)$ denote some proposed distribution such that $\sum_n Q(\mathbf{c}_n) = 1$ and $Q(\mathbf{c}_n) \geq 0$. Consider the following:

$$\begin{aligned} \log P(\mathbf{q}_n, \mathbf{t}_n|\mathbf{x}_i, \mathcal{C}_i; \boldsymbol{\theta}) = & \log \sum_{\mathbf{c}_k in \mathcal{C}_i} Q(\mathbf{c}_k) \frac{P(\mathbf{q}_n, \mathbf{t}_n, \mathbf{c}_k|\mathbf{x}_i, \mathcal{C}_i; \boldsymbol{\theta})}{Q(\mathbf{c}_k)} \\ \geq & \sum_{\mathbf{c}_k in \mathcal{C}_i} Q(\mathbf{c}_k) \log \frac{P(\mathbf{q}_n, \mathbf{t}_n, \mathbf{c}_k|\mathbf{x}_i, \mathcal{C}_i; \boldsymbol{\theta})}{Q(\mathbf{c}_k)} \end{aligned} \tag{5}$$

The last step used Jensens inequality. The Equation (5) gives a upper bound of the loss $l(\boldsymbol{\theta})$. When the bound is tight, we have $Q(\mathbf{c}_k) = P(\mathbf{c}_k|\mathbf{q}_n, \mathbf{t}_n; \boldsymbol{\theta})$.

To save the EM loop, we propose a non-parametric estimation of $P(\mathbf{c}_k|\mathbf{q}_n, \mathbf{t}_n; \boldsymbol{\theta})$. As a result, for each question-image pair $(\mathbf{x}_n, \mathbf{q}_n)$, we maximize the lower bound by optimizing:

$$\arg\min_{\boldsymbol{\theta}} -P(\mathbf{c}_n|\mathbf{q}_n, \mathbf{t}_n; \boldsymbol{\theta}_c) \log[P(\mathbf{q}_n|\mathbf{c}_n, \mathbf{x}_n, \mathbf{t}_n; \boldsymbol{\theta}_q)P(\mathbf{t}_n|\mathbf{c}_n; \boldsymbol{\theta}_t)] + \text{const} \tag{6}$$

This in fact assigns a weight $P(\mathbf{c}_n|\mathbf{q}_n, \mathbf{t}_n; \boldsymbol{\theta})$ to each instance. By using a non-parametric estimation, we are still able to apply BackProp and the SGD style optimizing algorithms by just augmenting each instance with an estimated weight.

Given a question $\mathbf{q}$ and a caption set $\mathcal{C}$ from the train set, we estimate $P(\mathbf{c}_k|\mathbf{q}, \mathbf{t}; \boldsymbol{\theta})$ by using the kernel density estimator [Scott, 2008]:

$$P(\mathbf{c}_k|\mathbf{q}, \mathbf{t}; \boldsymbol{\theta}) = P(\mathbf{c}_k|\mathbf{q}; \boldsymbol{\theta}) = \frac{s(\mathbf{q}, \mathbf{c}_k)}{\sum_{\mathbf{c}_j \in \mathcal{C}} s(\mathbf{q}, \mathbf{c}_j)} \tag{7}$$

where $s(\mathbf{q}, \mathbf{c})$ is a similarity function between a question and a caption. We assume $\mathbf{c}_k$ are conditionally independent of $\mathbf{t}$ because we can directly extract the question type from the question $\mathbf{q}$ by looking at the first few words.

For a given question, there are usually very few matched captions generated by *DenseCap* , hence the distribution of captions given a question is highly skewed. It is sufficient to randomly draw a caption each time to compute the probability based on Equation (7).

We formulate the similarity between a question and a caption by using both string similarity and embedding based similarity measures.

The surface string of a caption could be an exact or partial match of a given question. Thus we employ the Jaccard Index as string similarity measure between the surface string of a caption and that of a question.

$$\text{sim}_s(q, c) = \frac{q \cap c}{q \cup c}$$

where c and q denote their surface string respectively. Both strings are broken down to a set of char-based trigrams during the computation so that this measure still gives a high similarity if two strings differ only in some small variations such as singular and plural forms of nouns.

In case of synonyms or words of similar meanings come with different form such as "car" and "automobile", we adopt the pre-trained word embeddings to calculate their similarity by using the weighted averaged of word embeddings:

$$\text{sim}_e(\mathbf{q}, \mathbf{c}) = \cos\Big(\sum_{\mathbf{w}_i \in \mathbf{q}} \frac{\text{IDF}(\mathbf{w}_i)}{\sum_j \text{IDF}(\mathbf{w}_j)} \mathbf{E}\mathbf{w}_i, \sum_{\mathbf{w}_k \in \mathbf{c}} \frac{\text{IDF}(\mathbf{w}_k)}{\sum_j \text{IDF}(\mathbf{w}_j)} \mathbf{E}\mathbf{w}_k\Big)$$

where cos denotes the cosine similarity, $\text{IDF}(x)$ is the inverse document frequency of word x defined by $\frac{|\mathcal{V}|}{|\{d \in \mathcal{D} : x \in d\}|}$, and $\mathcal{D}$ is the corpus containing all questions, answers, and captions.

The final similarity measure is computed as the interpolation of the two measures:

$$s(\mathbf{q}, \mathbf{c}) = \alpha \text{sim}_s(\mathbf{q}, \mathbf{c}) + (1 - \alpha)\text{sim}_e(\mathbf{q}, \mathbf{c})$$

where the hyperparameter $\alpha \in (0, 1)$.

4 Experimental Setup

4.1 Datasets

We conduct our experiments on two datasets: VQA-Dataset [Antol *et al.*, 2015] and Visual7W [Zhu *et al.*, 2015]. The former is the most popular benchmark for VQA and the latter is a recently created dataset with more visually grounded questions per image than VQA.

VQA: a sample from the MS-COCO dataset [Lin *et al.*, 2014], which contains 254,721 images and 764,163 manually compiled questions respectively. Each image is associated with three questions on average.

Visual7W: a dataset composed of 327,939 QA pairs on 47,300 COCO images, collected from the MS-COCO dataset [Lin *et al.*, 2014] as well. In addition, it includes 1,311,756 human-generated answers in form of multiple-choice and 561,459 object groundings from 36,579 categories. Each image is associated with five questions on average.

4.2 Baseline

In this paper, we consider a baseline by training the image caption generation model NeuralTalk2 [Vinyals *et al.*, 2015] on image-question pairs. The baseline is almost the same as [Simoncelli and Olshausen, 2001], which is the only work generating questions from visual input. The model of neuraltalk2 differs from [Simoncelli and Olshausen, 2001] only in the RNNs used in the decoder. NeuralTalk2 adopts LSTM while [Simoncelli and Olshausen, 2001] chooses GRU [Cho *et al.*, 2014]. The two RNN models achieve almost identical performance in language modeling [Chung *et al.*, 2015].

4.3 Evaluation Measures

As a common practice for evaluating generated word sequences we employ three different evaluation metrics: BLEU [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005] and ROUGE-L [Lin, 2004].

BLEU is a modified n-gram precision. We varied the size of ngram from one to four, computed the corresponding measures respectively for each image and averaged the results across all images. Both METEOR and ROUGE-L[1] are F-Measures favoring precision, computed against the reference question with the highest score among all reference questions in the same image. The measures are averaged in sequel across all images. Therefore, all three of them are precision-oriented measures.

To measure the diversity of our generated questions, we also compute the the same set of evaluation measures by comparing each reference sentence with the best matching generated sentence of the same images. This provides an estimate of coverage in analogy of recall.

4.4 Implementation Details

We optimize all models with Adam [Kingma and Ba, 2014]. We fix the batch size to 64. We set the maximal epochs to 64 for Visual7W and the maximal epochs to 128 for VQA. The corresponding model hyperparameters were tuned on the validation sets. Herein, we set $\alpha = 0.75$.

5 Results and Discussions

Figure 3 illustrates all three precision-oriented measures evaluated on Visual7W and VQA datasets respectively. Our baseline is able to generate only one question per image. When we compare its results with the highest scored question per image generated by our model, our model outperforms the baseline with a wide margin. On the VQA test set, in the case of BLEU measures, the improvement over the baseline grows from 24% with unigram to 97% with four-gram. It is evident that our model is capable to generate many more higher-order n-grams co-occurred in reference questions. This improvement is also consistent with ROUGE-L because it is based on the longest common subsequence between generated questions and reference questions. Our model performs better than the baseline also not just because it generates more exact higher order n-grams than reference questions. METEOR

[1]We take the same β of F-Measure as the implementation in https://github.com/tylin/coco-caption.

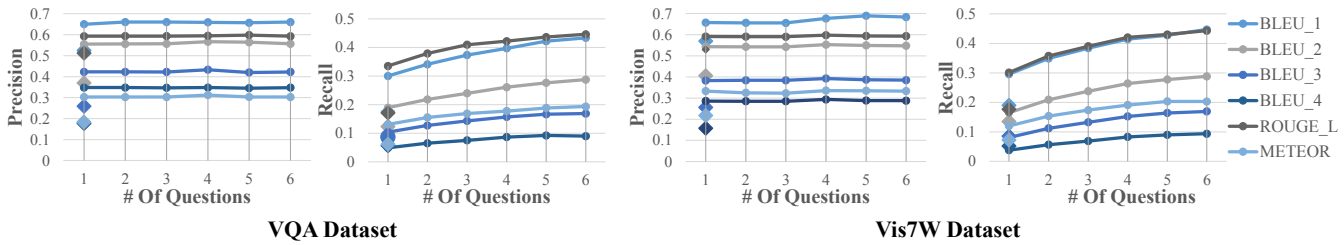


Figure 3: Precision and Recall of our method and baseline method(Neural Talk 2) on VQA Dataset and Visual7W Dataset. Results of our method are depicted with line with circles, we show the results by varying number of questions from one to six. By generating more questions,

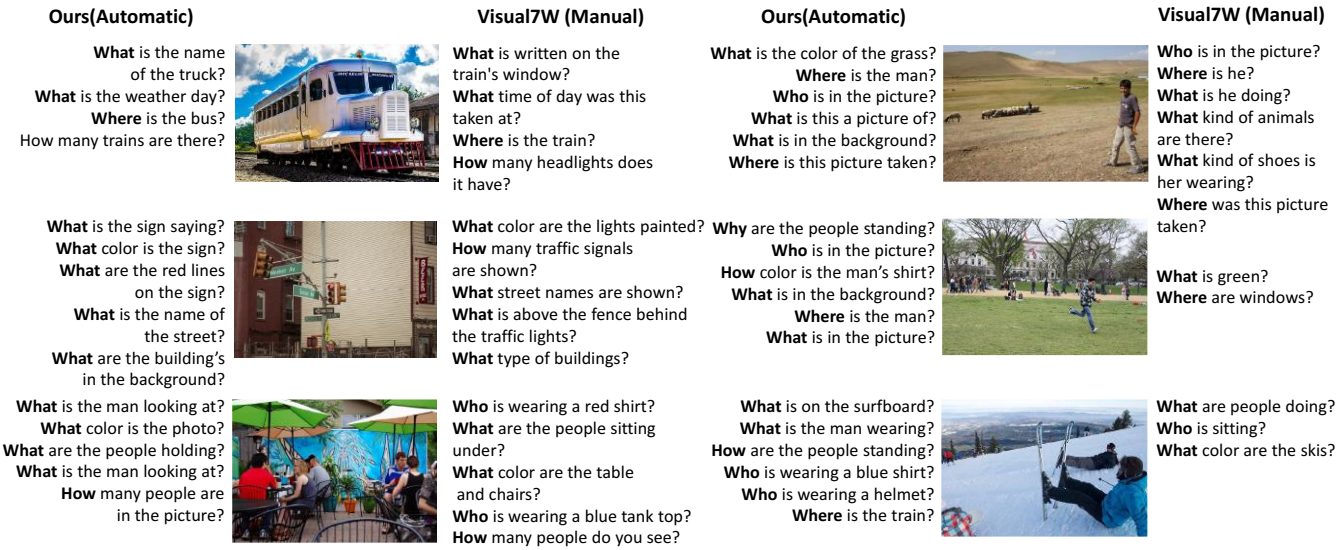


Figure 4: Comparison between manually composed questions and automatically generated questions by our method. Our method generates reasonable grounded questions and more versatile the manually provided ones.

considers unigram alignment by allowing multiple matching modules to consider synonymy and alternating word forms. With this measure, our model is still 65% higher than the baseline on the VQA test set. We also observe similar level of improvement over baseline on Visual7W dataset.

On both datasets, when the number of generated questions per image grows, the precison-oriented measures of our model are either similar or slightly declining because our model often generates meaningful questions that are not included in the ground-truth. The more questions we generate the more likely that the questions are not covered by manually constructed ones.

To measure the coverage of generated questions, we computed each reference question against all generated questions per image with all evaluation measures. As shown by Figure 3, all measures improves as the number of questions grows. Herein, both ROUGE-L and METEOR are way better than the baseline regardless of the number of generated questions on both datasets. When all six questions are generated, our model is 130% better than the baseline across all measures. In particular, with METEOR, our model shows an improvement of 216% and 179% over the baseline on VQA and Visual7W respectively. When the number of manually constructed questions is small, our model provides even more types questions than manual ones, as shown with the examples in Figure 4.

The distribution of question types generated by our model is more balanced than that of the ground-truth, while almost 55% of questions in Visual7W and 89% in VQA start with "*what*", as illustrated by Figure 6. Our model has also no tendency of generating too long or too short questions because the length distribution of the generated questions are very similar to that of the manually constructed datasets.

We also evaluate the effectiveness of the integration of bigram language model on both datasets. Herein, we compare two variants of our model, with and without the bigram model during decoding. As shown in Figure 5, regardless of precision or recall, decoding with the bigram model consistently outperforms the one without it. The inclusion of the bigram model effectively eliminates almost all repeated terms such as "the the" because the statistics collected by the bigram model favors grammatically well-formed sentences. This observation is also reflected in BLEU with higher-order ngrams by showing larger gaps.

6 Conclusion

In this paper, we propose the first model to automatically generate visually grounded questions with *varying* types. Our

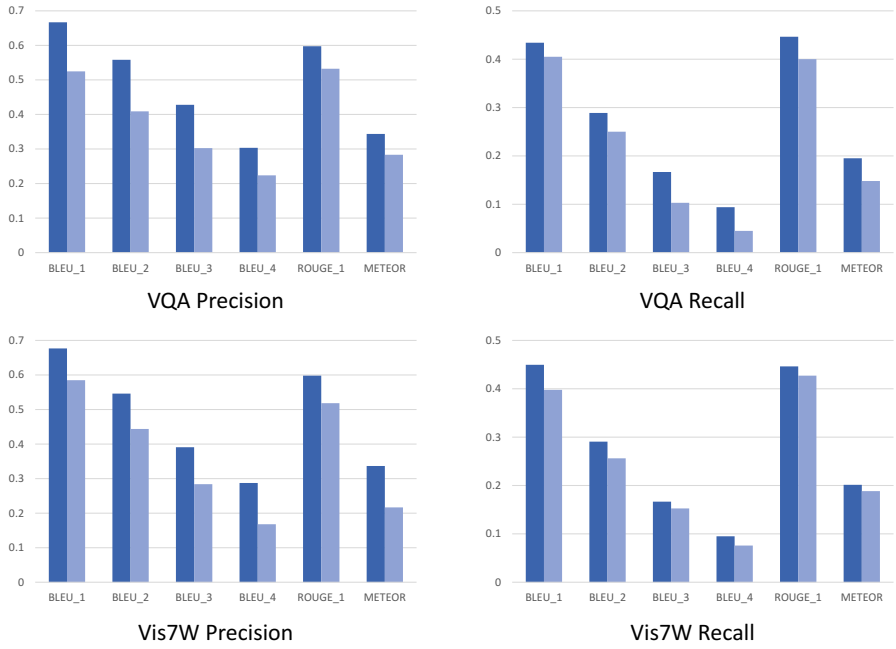


Figure 5: Comparing decoding between with the bigram model (dark blue) and without the bigram model (light blue).

model is capable of automatically selecting most likely question types and generating corresponding questions based on images and captions constructed by *DenseCap*. Experiments on VQA and Visual7W dataset demonstrates that the proposed model is able to generate reasonable and grammatically well-formed questions with high diversity. For future work, we consider automatically generation of visual question-answer pairs, which will likely enhance training of VQA systems.

Acknowledgments

This work is supported by National Key Technology R&D Program of China: 2014BAK09B04, National Natural Science Foundation of China: U1636116, 11431006, Research Fund for International Young Scientists: 61650110510, Ministry of Education of Humanities and Social Science: 16YJC790123.

References

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[Auer *et al.*, 2007] Sóren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC*, 2007.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

[Barnard *et al.*, 2003] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *Journal of machine learning research*, 2003.

[Bollacker *et al.*, 2008] Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 2008.

[Bordes *et al.*, 2014] Antoine Bordes, Jason Weston, and Nicolas Usunier. Open question answering with weakly supervised embedding models. In *ECML/PKDD*, pages 165–180, 2014.

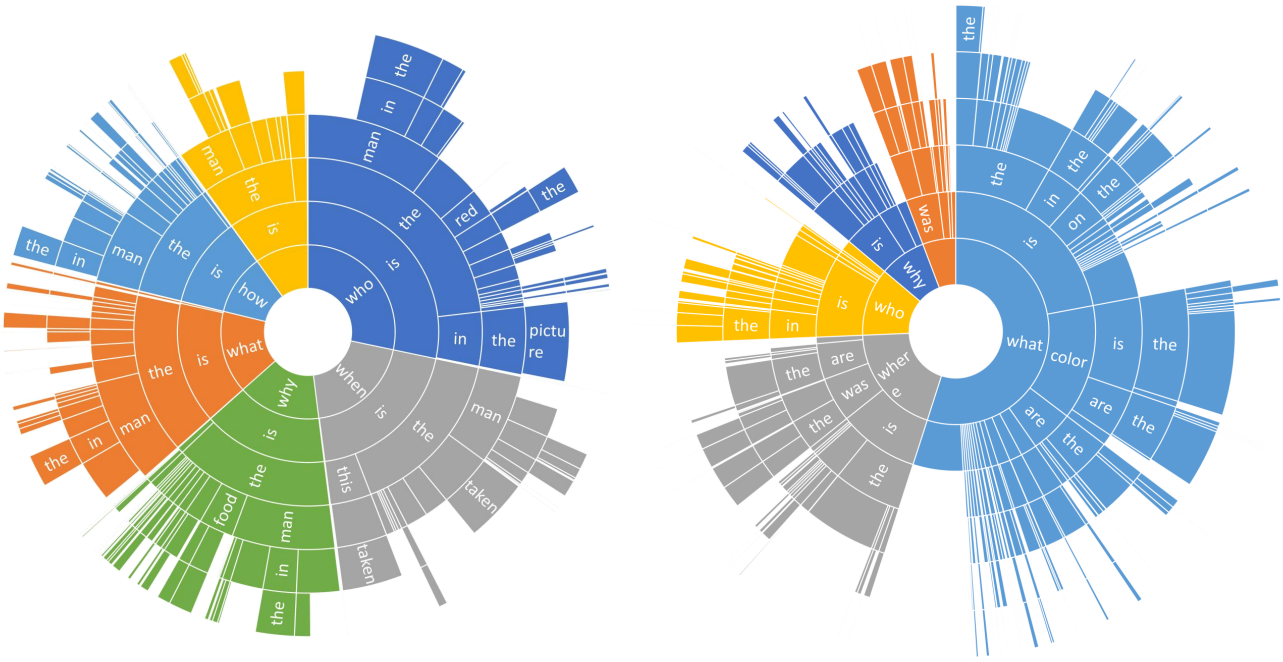
[Chen and Zitnick, 2014] Xinlei Chen and C Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*, 2014.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[Chung *et al.*, 2015] Junyoung Chung, Caglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Gated feedback recurrent neural networks. *CoRR, abs/1502.02367*, 2015.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Gao *et al.*, 2015] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image



(a) Automatically generated questions by our model.

(b) Manually generated questions in Visual7W dataset.

Figure 6: Distribution of question types.

question. In *Advances in Neural Information Processing Systems*, 2015.

[Geman *et al.*, 2015] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 2015.

[Gupta and Mannem, 2012] Ankush Gupta and Prashanth Mannem. From image annotation to image description. In *ICNIP*. Springer, 2012.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[Hodosh *et al.*, 2013] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 2013.

[Johnson *et al.*, 2015] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *arXiv preprint arXiv:1511.07571*, 2015.

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

[Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kneser and Ney, 1995] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *ICASSP-95*, 1995.

[Koehn *et al.*, 2003] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *NAACL*, pages 48–54, 2003.

[Kong *et al.*, 2014] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014.

[Krishna *et al.*, 2016] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[Kulkarni *et al.*, 2011] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating image descriptions. In *CVPR*, 2011.

[Kuznetsova *et al.*, 2012] Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. Collective generation of natural image descriptions. In *50th Annual Meeting of the Association for Computational Linguistics*, 2012.

[Kwiatkowski *et al.*, 2013] Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. Scaling semantic

parsers with on-the-fly ontology matching. In *EMNLP*, 2013.

[Li *et al.*, 2009] Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009.

[Lin *et al.*, 2014] 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 *ECCV*, 2014.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: ACL-04 workshop*, 2004.

[Malinowski and Fritz, 2014] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, NIPS'14, 2014.

[Malinowski *et al.*, 2015] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. Bleu: a method for automatic evaluation of machine translation. 2002.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[Pirsiavash *et al.*, 2014] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Inferring the why in images. *arXiv preprint arXiv:1406.5472*, 2014.

[Reddy *et al.*, 2016] Siva Reddy, Oscar Täckström, Michael Collins, Tom Kwiatkowski, Dipanjan Das, Mark Steedman, and Mirella Lapata. Transforming dependency structures to logical forms for semantic parsing. In *TACL*, 2016.

[Ren *et al.*, 2015] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, 2015.

[Scott, 2008] David W Scott. Kernel density estimators. *Multivariate Density Estimation: Theory, Practice, and Visualization*, pages 125–193, 2008.

[Simoncelli and Olshausen, 2001] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 2001.

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

[Suchanek *et al.*, 2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *WWW*, 2007.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

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

[Weston *et al.*, 2015] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

[Yao and Durme, 2014] Xuchen Yao and Benjamin Van Durme. Information extraction over structured data: Question answering with freebase. In *ACL*, 2014.

[Yu *et al.*, 2015] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank image generation and question answering. *arXiv preprint arXiv:1506.00278*, 2015.

[Zhu *et al.*, 2015] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. *arXiv preprint arXiv:1511.03416*, 2015.

[Zitnick *et al.*, 2013] C Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. Learning the visual interpretation of sentences. In *ICCV*, pages 1681–1688, 2013.