A Question Type Driven Framework to Diversify Visual Question Generation

Zhihao Fan[1], **Zhongyu Wei**[*1], **Piji Li**[2], **Yanyan Lan**[3,1] and **Xuanjing Huang**[4]
[1]School of Data Science, Fudan University, China
[2]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong
[3]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[4]School of Computer Science, Fudan University, China
{14300180043,zywei}@fudan.edu.cn, pjli@se.cuhk.edu.hk, lanyanyan@ict.ac.cn, xjhuang@fudan.edu.cn

Abstract

Visual question generation aims at asking questions about an image automatically. Existing research works on this topic usually generate a single question for each given image without considering the issue of diversity. In this paper, we propose a question type driven framework to produce multiple questions for a given image with different focuses. In our framework, each question is constructed following the guidance of a sampled question type in a sequence-to-sequence fashion. To diversify the generated questions, a novel conditional variational auto-encoder is introduced to generate multiple questions with a specific question type. Moreover, we design a strategy to conduct the question type distribution learning for each image to select the final questions. Experimental results on three benchmark datasets show that our framework outperforms the state-of-the-art approaches in terms of both relevance and diversity.

1 Introduction

Recent years see the popularity of multi-modal research on vision and language. Popular tasks include visual caption generation (VCG) [Vinyals *et al.*, 2015] and visual question answering (VQA) [Antol *et al.*, 2015]. VCG aims at generating descriptions for a given image with a goal of scene understanding, while VQA provides a related question and requires an answer to it. Research for these two tasks are fueled by several manually generated corpora [Lin *et al.*, 2014; Zhu *et al.*, 2016]. Different from generating a statement (descriptions or answers), visual question generation (VQG) aims at asking questions about the given image. Teaching machine the skill of asking is important in a variety of areas, e.g., providing demonstrations in child education [Kunichika *et al.*, 2004], initializing a conversation for chatbots [Mostafazadeh *et al.*, 2017], etc. On the other hand, it can benefit the question answering task by constructing question sets automatically [Ren *et al.*, 2015] to reduce the labor of human annotation.

Figure 1: Two sample images with human generated questions. The left image contains five types of questions (i.e. *when*, *what*, *how*, *where* and *did*); the right image contains two types of questions (i.e. one question for *how* and four questions for *what*).

VQG is a rising research topic in both fields of computer vision and natural language processing [Ren *et al.*, 2015; Mostafazadeh *et al.*, 2016]. Mostafazadeh et al. [2016] explored different approaches for this task, and the experimental results showed a retrieval based approach which chose the question from the closest image achieved the best performance. In their experiment setting, only one question is generated for a given image. We argue that different people might have different questions about the same image, therefore, a visual question generation system should also be able to produce questions with various of focuses. Although there are some attempts to diversify the results for text generation [Vijayakumar *et al.*, 2018; Jain *et al.*, 2017], none of them considers the special characteristics of questions.

Question is a linguistic expression used to make requests for information. In terms of information needs, we can classify questions into different types, such as, *what*, *which*, *how*, etc. Figure 1 presents two sample images and each of them is

*Corresponding author

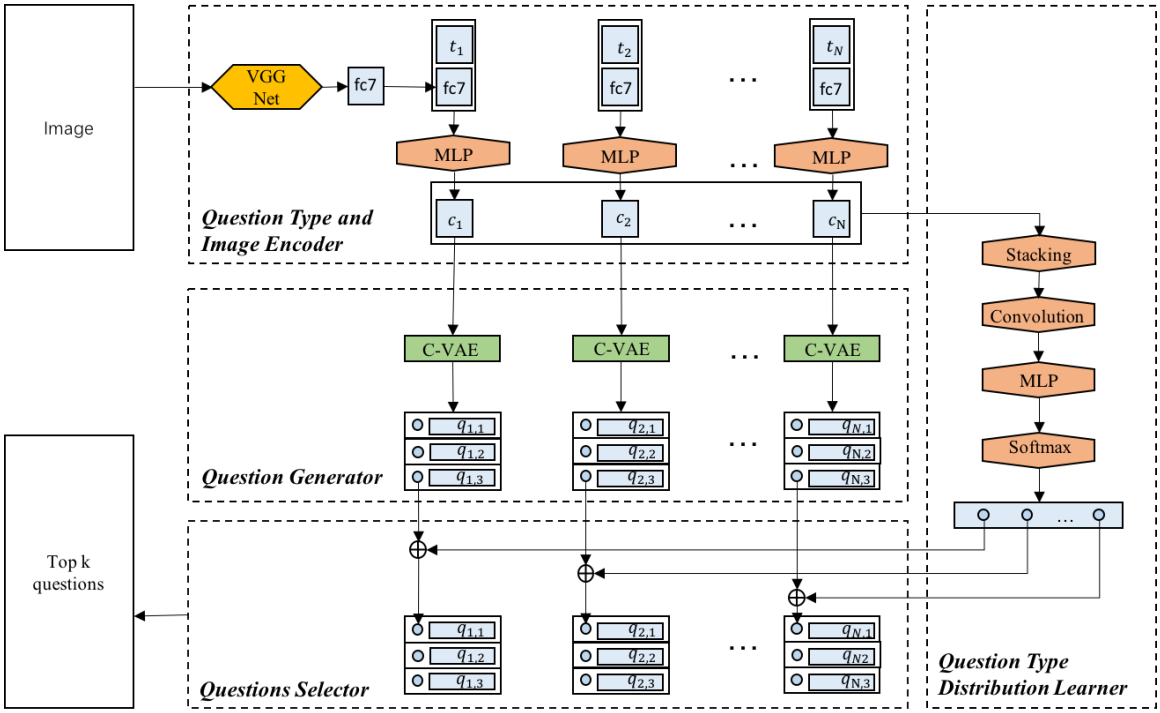


Figure 2: The overall framework of the question type driven diverse question generation model.

accompanied with five human generated questions. We have two observations: 1) Questions related to a given image have various types. 2) There can be several questions for each single type with different focuses. Further analysis on a human generated image-question paired dataset (*VQG-Flickr*) from [Mostafazadeh *et al.*, 2016] shows that around 52.8% of images are questioned by more than 2 types of questions. It is deemed that question type can be used to organize questions for different information needs. Thus we explore the problem that how question type can be used to enable diverse question generation.

In this paper, we propose a question type guided strategy for diverse question generation. In our framework, a question is constructed in two steps. First, a question type is sampled to determine what kind of information is requested. Second, the content of the question is generated conditioning on the sampled question type and the visual information of the image. Two components are proposed to enforce the question diversification.

- **Variational auto-encoder based question generator**. For each question type, multiple questions can be asked with different focuses. Instead of using a deterministic sequence-to-sequence model [Cho *et al.*, 2014], we propose a conditional variational auto-encoder (*C-VAE*) to produce multiple questions for a specific question type.
- **Question type distribution learner**. The probability of different question types should be different according to the content of images. For example, questions with the type *who* might be less possible for a landscape picture. To utilize this intrinsic characteristic, we design a model to estimate the probability distribution of the question types for the input image.

A neural network based framework is proposed to conduct the learning of these two components jointly. Questions are selected by considering both the generation probability from *C-VAE* and the probability of the corresponding question type. We evaluate our framework on three public benchmark datasets in terms of relevance and diversity. Experimental results show that our framework outperforms the state-of-the-art visual text generation models in a large margin.

2 Model

Our model takes an image I_i as input and generates diverse questions as output. The overall framework is shown in Figure 2. It consists of four components, namely question type and image encoder, question type distribution learner, question generator, and question selector. The question type and image encoder learns the mixed representations for question types and the given image as the input for the other components. The question type distribution learner computes the probability distribution of different question types. The question generator produces multiple questions for a specific question type based on *C-VAE*. Finally, the question selector outputs top-k questions considering both the generation probability from *C-VAE* and the probability of corresponding question type.

2.1 Question Type and Image Encoder

Interrogative words (e.g. *who*, *which*, etc.) in questions imply the type of the information that questioners want to acquire.

It is also widely used as a way to group questions [Zhu *et al.*, 2016]. Although there are many alternative taxonomies to organize questions [Graesser *et al.*, 2008], they also involve heavy human annotation. For simplicity, we directly use interrogative words to represent question types in our framework. Note that the model is compatible for other question classification schemes.

Based on interrogative words, we predefined *N* question types. Each question type is represented as an embedding vector with fixed length. Embedding vectors are initialized by the word embeddings (from word2vec) of their corresponding interrogative words and would be optimized in the training process. We use VGGNet [Simonyan and Zisserman, 2015] to process the image and borrow *fc7*[1] feature to represent the image. In order to learn the correlation between the question type and the image feature, we concatenate them and feed them into a three-layer MLP (multilayer perceptron) with batch normalization. The procedure can be described by $c_{i,j} = f_1([v_i, t_j])$. In which, f_1 stands for the corresponding MLP transformation, $[\cdot]$ stands for the concatenating operation and $c_{i,j}$ is the processed feature vector related to image *i* and the question type t_j.

2.2 Question Generation via Conditional Variational Auto-Encoder

Our question generator aims to produce questions for the given image I_i with a specific question type t_j. Given the mixed representation c_{ij} for I_i and t_j, a recurrent neural network (RNN) is used to decode the source information into a sequence of words to form a question. Considering that there can be multiple questions about an image with a specific question type, a deterministic decoder is not satisfactory for diverse question generation.

Recently, variational auto-encoders (VAEs) [Kingma and Welling, 2014; Rezende *et al.*, 2014] showed strong capability in modeling latent random variables and improved the performance for generation tasks on both text [Bowman *et al.*, 2016; Li *et al.*, 2017] and image [Gregor *et al.*, 2015]. Inspired by [Jain *et al.*, 2017], we employed a conditional variational auto-encoder (C-VAE) to generate multiple questions for a specific question type. In our case, questions are generated in condition on the given image and a question type. A specific question type can be interpreted as a pre-defined cluster for the VAE to generate questions from a fine-grained setting which can further enforce diversity. In other words, this modification provide a guidance for VAE generators.

A *C-VAE* consists of two components: variational encoder (inference) and variational decoder (generation). During the inference, we map the ground-truth questions into a latent space and learn to recover it in the generation process. A distribution representing the latent space is optimized in the training process, from which we sample a latent variable $\mathbf{z}$ for generation. Since the latent variable follows some distribution instead of being a deterministic value, the decoder is able to generate different questions via sampling. In inference and generation processes, visual and the question type information (denoted by c_{ij}) are used as condition to guide latent distribution learning and question generation respectively. The framework of our *C-VAE* is shown in Figure 3.

[1] *fc7* is the output of 7^{th} fully connected layer of VGGNet.

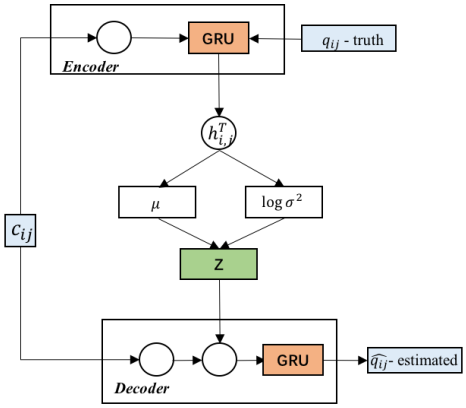


Figure 3: The framework of the conditional variational auto-encoder: $q_{ij} - truth$ stands for human generated questions in training dataset and $q_{ij} - estimated$ stands for the question generated by *C-VAE*.

Suppose that the target latent distribution for the image I_i and the question type t_j is $P_\theta(z_{i,j}|I_i, t_j)$. We aim to learn a distribution $P_\phi(z_{i,j}|I_i, t_j)$ to approximate the target distribution $P_\theta(z_{i,j}|I_i, t_j)$. We can have the following equations:

$$
\begin{aligned}
&log\big(P_\theta(q_{i,j}|I_i, t_j)\big) \\
&= KL\big(P_\phi(z_{i,j}|q_{i,j}, I_i, t_j)\|P_\theta(z_{i,j}|q_{i,j}, I_i, t_j)\big) \\
&+ E_{P_\phi(z_{i,j}|q_{i,j},I_i,t_j)}\left[log\left(\frac{P_\theta(q_{i,j}, z_{i,j}|I_i, t_j)}{P_\phi(z_{i,j}|q_{i,j}, I_i, t_j)}\right)\right] \\
&\geq E_{P_\phi(z_{i,j}|q_{i,j},I_i,t_j)}\left[log\left(\frac{P_\theta(q_{i,j}, z_{i,j}|I_i, t_j)}{P_\phi(z_{i,j}|q_{i,j}, I_i, t_j)}\right)\right] \\
&= -KL\big(P_\phi(z_{i,j}|q_{i,j}, I_i, t_j)\|P_\theta(z_{i,j}|I_i, t_j)\big) \\
&+ E_{P_\phi(z_{i,j}|q_{i,j},I_i,t_j)}\Big[log\Big(P_\theta\big(q_{i,j}|z_{i,j}, I_i, t_j\big)\Big)\Big] \qquad (1)
\end{aligned}
$$

Given the latent variable follows a normal distribution $N(0, I)$ in *VAE*, the first term of the last expression in Equation 1 can be marginalized. And the second term can be estimated by drawing samples from the distribution $P_\phi(z_{i,j}|q_{i,j}, I_i, t_j)$. The lower bound of $log\big(P_\theta(q_{i,j}|I_i, t_j)\big)$, which is also the loss of variational auto-encoder, can thus be written as Equation 2:

$$
\begin{aligned}
&\mathcal{L}_{C-VAE}(q_{i,j}, I_i, t_j|\theta, \phi) \\
&= -KL\big(P_\phi(z_{i,j}|q_{i,j}, I_i, t_j)\|N(0, I)\big)\big) \\
&+ \frac{1}{L}\sum_{l=1}^{L} log\Big(P_\theta\big(q_{i,j}|z^{(l)}_{i,j}, I_i, t_j\big)\Big) \qquad (2)
\end{aligned}
$$

where $z^{(l)}_{i,j} = g_\phi(q_{i,j}, I_i, t_j, \epsilon_{i,j}), \epsilon_{i,j} \sim N(0, I)$, and L is the number of samples drawn.

We make use of a RNN g_ϕ to draw samples from the distribution $P_\phi(z_{i,j}|q_{i,j}, I_i, t_j)$, from which we can get a final state $\mathbf{h}^T_{i,j} \in R^{d_f}$. Two fully connected layer f_2 and f_3 are

then used to transfer $\mathbf{h}_{i,j}^T$ into $u_{i,j}$ and $log\sigma_{i,j}^2$ which are the mean and logarithm variance of some Gaussian distribution. The process can be represented by Equation 3.

$$u_{i,j} = f_2(\mathbf{h}_{i,j}^T),\ log\sigma_{i,j}^2 = f_3(\mathbf{h}_{i,j}^T) \tag{3}$$

Different from the traditional variational auto-encoder that only uses the variables sampled from the latent space for decoding, we need to take the visual information into consideration. In the generation component, feature vector $c_{i,j}$ (information for the question type t_j and the image i) and the sampled variable $z_{i,j}$ are fed into the first state and second state of the RNN decoder respectively. In the training stage, $z_{i,j}$ is derived from function g_ϕ, through which the model encodes generated questions into the latent space and approximates the prior $N(0, I)$. In the testing stage, $z_{i,j}$ is directly sampled from the learned latent space. In order to get more diverse questions, we can sample from a probability distribution that has larger variance such as $N(0, 5I)$. Then the sampled value is regarded as the latent variable $z_{i,j}$.

2.3 Question Type Distribution Learner

Given an image, question type distribution learner produces a probability distribution to indicate how likely the image will be inquired by different question types. As this involves all question types, we stack all the $c_{i,j}$ across different question types together to form a matrix with a dimension of $N \times d_f$. The obtained matrix is then fed into the three-layer MLP transformation f_1 (shared with the one used in the representation component) and a three-layer convolution component in sequence. Two fully connected layers followed by a softmax layer are then used to estimate the final question type distribution for the given image.

A cross-entropy based loss function is utilized for parameter learning. The loss for image I_i in terms of question type probability prediction is shown as Equation 4:

$$\mathcal{L}_t(I_i) = -\sum_{j=1}^{N} p_{i,j} log\hat{p}_{i,j} \tag{4}$$

where $p_{i,j}$ and $\hat{p}_{i,j}$ denote the target and predicted probability for question type t_j respectively. We use the maximum likelihood estimation to compute the target probability of question type t_j for image I_i.

Note that both question type distribution learner and question generator are optimized jointly with the objective function:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{C-VAE}(q_{i,j}, I_i, t_j|\theta, \phi) + \lambda\mathcal{L}_t(I_i) \\ &= -KL(P_\phi(z_{i,j}|Q_{i,j}, I_i, t_j)\|N(0, I)) \\ &+ \frac{1}{L}\sum_{l=1}^{L} log\Big(P_\theta\big(Q_{i,j}|z_{i,j}^{(l)}, I_i, t_j\big)\Big) \\ &+ \lambda\sum_{j=1}^{N} p_{i,j} log\hat{p}_{i,j}\end{aligned} \tag{5}$$

where λ is introduced to balance the contribution of these two loss items.

2.4 Question Selection

For the given image, we assign a probability to each generated question and select top questions as the output. Taking question type t_j into consideration, the probability of question $q_{i,j}$ generated for image I_i can be expressed as Equation 6:

$$P(q_{i,j}|I_i) = P(q_{i,j}|I_i, t_j)P(t_j|I_i) \tag{6}$$

where $P(t_j|I_i)$ stands for the probability of generating questions of type t_j and it can be estimated by the result from Section 2.3 directly. $P(q_{i,j}|I_i, t_j)$ stands for the generation probability of the constructed question and it can be computed in the decoding process of the *C-VAE*.

In the process of question generation, our target is the following expression:

$$\underset{q_{i,j}\in\mathcal{Q}}{\operatorname{argmax}} \big\{P(q_{i,j}|I_i, t_j)\big\} \tag{7}$$

where $q_{i,j}$ is decoded word by word based on a RNN decoder. Suppose $w_{i,j,t}$ is the word in the t_{th} time step for for generating $q_{i,j}$, the target for each time step in decoder can be viewed as the following expression:

$$\underset{w_{i,j,t}\in\mathcal{W}}{\operatorname{argmax}} \big\{P(w_{i,j,t}|I_i, t_j, w_{i,j,1}, \cdots, w_{i,j,t-1})\big\} \tag{8}$$

And the likelihood of the generated question can be calculated by multiplying all the probability elements of each word:

$$\begin{aligned}P(q_{i,j}|I_i, t_j) = &P(w_{i,j,1}|I_i, t_j) \\ &P(w_{i,j,2}|I_i, t_j, w_{i,j,1}) \cdots \\ &P(w_{i,j,T}|I_i, t_j, w_{i,j,1}, \cdots, w_{i,j,T-1})\end{aligned} \tag{9}$$

Usually, log-likelihood is used in practice. We can compute every single probability P on the right side through every time step in decoder. Obviously, maximizing every single probability cannot guarantee maximizing the final target. Therefore, we use beam search [Wu *et al.*, 2017] to obtain $max\{P(q_{i,j}|I_i, t_j)\}$.

We use S_1 to represent the probability of generating target question based on $p(q_{i,j}|I_i, t_j)$:

$$S_1(I_i, t_j, q_{i,j}) = log(P(q_{i,j}|I_i, t_j))/lp(q_{i,j}) \tag{10}$$

In order to avoid the problem that such model favors shorter question, a length penalty $lp(q_{i,j})$ is introduced and it can be computed as Equation 11 following [Wu *et al.*, 2017]:

$$lp(q_{i,j}) = \frac{(5 + |q_{i,j}|)^\alpha}{(5+1)^\alpha},\ \alpha \in (0, 1) \tag{11}$$

where $|q_{i,j}|$ represents the number of words in the generated question.

In line with S_1, we simply take logarithmic on the probability of the corresponding question type given the image which is represented as S_2:

$$S_2(I_i, t_j) = log(P(t_j|I_i)) \tag{12}$$

During the testing, we consider scores from question generation (S_1) and question type probability (S_2) to compute the final score of generated questions. Based on such score, generated questions are able to be ranked, and we can select top k from them as output.

$$S(I_i, q_{i,j}) = S_1(I_i, t_j) + S_2(I_i, t_j, q_{i,j}) \tag{13}$$

Model	*VQG-Flickr*				*VQG-MS COCO*				*Visual7W*			
	corpus B-4	B4	M	R	corpus B-4	B-4	M	R	corpus B-4	B-4	M	R
(I)	16.67	36.48	18.76	51.17	22.21	39.22	23.55	55.30	19.70	38.21	23.82	57.62
(II)	18.27	35.96	20.59	51.97	24.92	39.45	24.80	57.32	25.20	41.18	25.75	59.74
(III)	19.26	36.42	20.15	52.07	25.68	39.83	25.22	57.60	27.28	42.74	26.26	62.42
(IV)	17.57	35.70	19.15	47.05	23.02	39.47	23.92	52.98	17.75	35.55	25.05	48.86
(V)	19.47	37.12	21.17	53.00	27.78	41.32	25.82	57.23	27.90	43.54	26.42	64.23
(VI)	**21.84**	**38.53**	**21.30**	**53.24**	**30.17**	**43.28**	**26.74**	**59.39**	**28.44**	**43.81**	**26.89**	**64.82**

Table 1: Results for all comparative models in terms of relevance scores (100%) based on Top - 1 question: **bolded** numbers are the best performance in each column; B for BLEU, R for ROUGE and M for METOR.

3 Experiment

3.1 Experiment Datasets

We use three public datasets to evaluate our models, namely *VQG-MS COCO*, *VQG-Flickr* [Mostafazadeh *et al.*, 2016] and *Visual7W-telling* [Zhu *et al.*, 2016]. First two datasets contain 5,000 images with 5 human generated questions for each image and the third one has 28,653 images with various number of questions (refer to [Zhu *et al.*, 2016] for details). In *Visual7W-telling*, questions are classified into 6 types based on interrogative words ('what', 'who', 'where', 'when', 'why', 'how'). With the same strategy, we categorize questions in *VQG-MS COCO* and *VQG-Flickr* into different types. By merging some interrogative words with different tenses into the same group (e.g., treat 'was' and 'is' as 'is'), we obtain 9 types ('what', 'who', 'where', 'when', 'why', 'how', 'is', 'do', 'can').

3.2 Evaluation Methods

Our model aims to generate multiple questions for a given image without sacrificing the quality of each single question. Therefore, we conduct two kinds of metrics to evaluate the performance of our model in terms of relevance and diversity respectively. For relevance evaluation, we report BLEU-4 [Papineni *et al.*, 2002], corpus-BLEU-4, METEOR [Denkowski and Lavie, 2014] and ROUGE [Lin, 2004]. For diversity evaluation, we utilize *mBLEU* [Wang *et al.*, 2016]. It assumes that the system is better if similarities among questions generated for an image are lower. Suppose that $B(h, \mathcal{R})$ denotes the function of BLEU, and $\mathcal{Q}_i = \{q_{i,k}, k = 1, 2, \cdots, K\}$ is the K questions generated for image I_i, *mBLEU* can be computed by the Equation 14.

$$mB(\mathcal{Q}_i, I_i) = \frac{1}{K} \sum_{k=1}^{K} B(q_{i,k}, \mathcal{Q}_i \setminus q_{i,k}) \tag{14}$$

3.3 Comparative Models

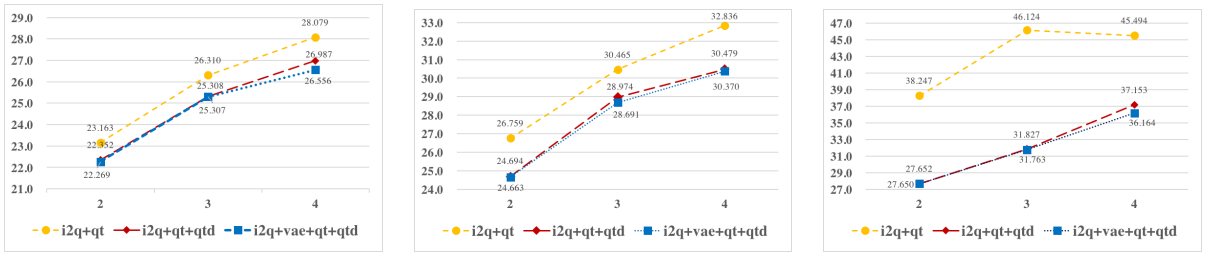
We compare our models with some baselines and some state-of-the-art methods.

- *NN-generator (I)*: Use the question from the most similar image as the question for a target image [Mostafazadeh *et al.*, 2016]. Cosine similarity based on *fc7* features is used to search for similar images. Note that only one question is generated by this model following its original setup.
- *i2q (II)*: This is the state-of-the-art approach for text generation that generates a question from image features based on a deterministic sequence-to-sequence model [Ren *et al.*, 2015]. Only one question is generated for each image.
- *i2q+C-VAE (III)*: This is the model proposed by Jain et al. [2017] that uses C-VAE for questions generation.
- *i2q+QT (IV)*: In addition to i2q, we introduce question type to guide question generation. Only generation probability S_1 (refer to equation 10) is used to select top questions. This model is similar to the one proposed in [Shijie *et al.*, 2017]. Note that only one question is generated with a specific type of question.
- *i2q+QT+QTD (V)*: On top of *i2q+QT*, question type probability distribution is learned to guide question selection.
- *i2q+QT+C-VAE+QTD (VI)*: This is the complete version of our model that uses both *C-VAE* for multiple question generation and selects top questions with the guidance of question type probability distribution.

3.4 Results and Analysis

Since some of the comparative models are unable to generate multiple questions, we use the top-1 question from multiple question generators for comparison. The overall results in terms of relevance can be seen in Table 1. We have several findings:

- The performance of *NN-generator* that uses information retrieval based approach is quite competitive in-line with the results reported in [Mostafazadeh *et al.*, 2016]. However, further analysis on questions generated by *NN-generator* reveals that a large percentage of generated questions are not relevant to the target image. Therefore, the strategy of reusing questions from similar images is not sufficient for generating question with high quality.
- The performance of *i2q* is better than *i2q+QT*. Without considering question type, *i2q* is capable to compare questions across question types to select the one with highest probability. However, the question generated by *i2q+QT* is guided by question type that makes the probability comparison of questions across question types indirectly.
- By adding the probability distribution of question types to guide the top question selection, both *i2q+QT+QTD* and *i2q+QT+C-VAE+QTD* outperform *i2q*. This proves that our question type distribution learner is able to learn



(a) VQG-Flickr (b) VQG-MS COCO (c) Visual7W

Figure 4: mBLEU score of generated diverse questions in terms of different number of questions generated.

Ground Truth Questions:
Q1: Why is everyone gathered outside?
Q2: What city is this temple located at?
Q3: What is the name of this landmark?
Q4: What does the sign say?
Q5: What is the name of the temple?

i2q+C-VAE+QT+QTD :
Q1: Why are there so many people ?
Q2: What event are these people attending ?
Q3: Is this a festival ?
Q4: Where was this taken ?
Q5: What building is that ?

i2q+QT:
Q1: When was this taken ?
Q2: Is this a temple ?
Q3: Where is this temple ?
Q4: How is the temple?
Q5: Who is that taking a picture of ?

i2q+QT+QTD:
Q1: What are they celebrating ?
Q2: When was the parade for ?
Q3: Where is this building located ?
Q4: Which city is that in ?
Q5: Were they there on vacation ?

Figure 5: Samples of generated questions from different models.

the intrinsic characteristic of the given image. Besides, it can evaluate the appropriateness of questions from different question types in a global way to select questions with higher relevance.

- Our proposed model *i2q+QT+C-VAE+QTD* that uses both components of question type distribution learner and *C-VAE* based question generator achieves the best performance in terms of all the four relevance metrics across three datasets. This confirms the effectiveness of our framework in terms of relevance.

We further evaluate the performance of our models in terms of diversity. We vary the number of questions generated by three multiple question generators to see how good they are for diverse question generation in terms of *mBLEU* (the less the better). Experimental results are shown in Figure 4. As the number of questions generated increases, *mBLUE* increases. By adding question type distribution learner, *i2q+QT+QTD* can generate more diverse questions than *i2q+QT*. With the help of *C-VAE*, *i2q+QT+C-VAE+QTD* can improve the diversity further.

An example image with questions generated by different models is shows in Figure 5. Although *i2q+QT* is able to generate questions with different question types, the focus of questions are largely concentrated on the word *temple*. This is because the model would receive higher score for generating such a key term. With the guidance of question type distribution learner, questions generated by *i2q+QT+QTD* are more diverse in terms of topics. By adding *C-VAE*, *i2q+C-VAE+QT+QTD* is able to generate two questions for the type of *what* and improve the quality of the questions further. It is also exciting that our model can generate some good questions that are not mentioned in the ground-truth.

4 Conclusion and Future Work

In this paper, we propose a neural-network based framework for visual question generation using interrogative words as question type to organize questions and enforce the diversification of the results. Experimental results on three publicly available question generation datasets showed the effectiveness of our framework in terms of both relevance and diversity. Further analysis on datasets shows that questions can be generated by considering information from multiple zones in the input image. Therefore, a feature extractor that treats all information equally is not able to manipulate input a such a fine-grained way. We thus will explore to use an attention mechanism to select and combine salient zones of a given image for question generation in future.

Acknowledgments

The work is partially supported by National Natural Science Foundation of China (Grant No. 61702106), Shanghai Science and Technology Commission (Grant No. 17JC1420200, Grant No.17YF1427600 and Grant No. 16JC1420401).

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 *Proceedings of the IEEE ICCV*, pages 2425–2433, 2015.

[Bowman *et al.*, 2016] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of CoNLL*, pages 10–21, 2016.

[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. In *Proceedings of EMNLP*, pages 1724–1734, 2014.

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

[Graesser *et al.*, 2008] Art Graesser, Vasile Rus, and Zhiqiang Cai. Question classification schemes. In *Proceedings of first Workshop on Question Generation*, 2008.

[Gregor *et al.*, 2015] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *Proceedings of ICML*, pages 1462–1471, 2015.

[Jain *et al.*, 2017] Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of IEEE CVPR*, pages 6485–6494, 2017.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of ICLR*, 2014.

[Kunichika *et al.*, 2004] Hidenobu Kunichika, Tomoki Katayama, Tsukasa Hirashima, and Akira Takeuchi. Automated question generation methods for intelligent english learning systems and its evaluation. In *Proceedings of ICCE*, 2004.

[Li *et al.*, 2017] Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. Deep recurrent generative decoder for abstractive text summarization. In *Proceedings of EMNLP*, pages 2091–2100, 2017.

[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 *Proceedings of ECCV*, pages 740–755. Springer, 2014.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.

[Mostafazadeh *et al.*, 2016] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *Proceedings of the 54th ACL*, pages 1802–1813. Association for Computational Linguistics, 2016.

[Mostafazadeh *et al.*, 2017] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of 8th IJCNLP*, 2017.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, pages 311–318. Association for Computational Linguistics, 2002.

[Ren *et al.*, 2015] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Proceedings of NIPS*, pages 2953–2961, 2015.

[Rezende *et al.*, 2014] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th ICML*, pages 1278–1286, 2014.

[Shijie *et al.*, 2017] Zhang Shijie, Qu Lizhen, You Shaodi, Yang Zhenglu, and Zhang Jiawan. Automatic generation of grounded visual questions. In *Proceedings of the 26th IJCAI*, pages 4235–4243, 2017.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR*, 2015.

[Vijayakumar *et al.*, 2018] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. In *Proceedings of AAAI*, 2018.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE CVPR*, pages 3156–3164, 2015.

[Wang *et al.*, 2016] Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. Diverse image captioning via grouptalk. In *IJCAI*, pages 2957–2964, 2016.

[Wu *et al.*, 2017] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *TACL*, 5:339–351, 2017.

[Zhu *et al.*, 2016] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE CVPR*, pages 4995–5004, 2016.