Category-Guided Visual Question Generation (Student Abstract)

Hongfei Liu[1,2], **Jiali Chen**[1,2], **Wenhao Fang**[1,2], **Jiayuan Xie**[1,2], **Yi Cai**[1,2*]

[1] School of Software Engineering, South China University of Technology, Guangzhou, China
[2] Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education
{201830340384, 201930390036, sewenhaofang, sexiejiayuan}@mail.scut.edu.cn, ycai@scut.edu.cn

Abstract

Visual question generation aims to generate high-quality questions related to images. Generating questions based only on images can better reduce labor costs and thus be easily applied. However, their methods tend to generate similar general questions that fail to ask questions about the specific content of each image scene. In this paper, we propose a category-guided visual question generation model that can generate questions with multiple categories that focus on different objects in an image. Specifically, our model first selects the appropriate question category based on the objects in the image and the relationships among objects. Then, we generate corresponding questions based on the selected question categories. Experiments conducted on the TDIUC dataset show that our proposed model outperforms existing models in terms of diversity and quality.

Introduction

Visual question generation (VQG) aims to generate questions based on given images (Xie et al. 2021), which has attracted increasing attention in recent years due to its vast potential applications in dialogue systems and intelligent education systems (He et al. 2017), etc.

Existing methods on VQG are mainly divided into constrained VQG and unconstrained VQG based on whether specific constraints (e.g., answers) are given. Unconstrained VQG with less labor-intensive can be more easily applied to various applications. However, ignoring the constraints may cause their models to tend to generate similar general questions for each image, e.g., "Where is the man?". These general questions are not relevant to the specific content in each image, which fails to apply to different scenarios in various applications. As shown in Figure 1, we are more inclined to generate a question "What is the player wearing the helmet doing?" when our conversation about the scenarios of "sport" or "activity". Therefore, it is necessary to generate questions with different question categories based on different objects and their relationships in an image to be more suitable for various scenarios.

To this end, we propose a category-guide visual question generation (CGVQG) model, which aims to generate diverse questions with different question categories, e.g., "Activity Recognition"and "Positional Reasoning". Our model consists of three components, i.e., visual encoder, category selector, and question generator. Given an image, we usually select the appropriate question category based on a specific object or the relationships among objects. For example, when a specific object "baseball" appears in the image of Figure 1, we are likely to ask questions about the category of sports recognition or activity recognition. Thus, the visual encoder first utilizes VisualBert to consider the relationships among multiple objects and then utilizes the TF-IDF method to calculate the contribution of each object to different question categories. Then, we design a category selector to extract multiple categories suitable for questioning based on the object-level features. Finally, the question generator generates corresponding questions based on each selected question category. Since different questions are generated by focusing on different image regions corresponding to their related question categories, the generated questions are diverse, which can be better applied in various scenarios.

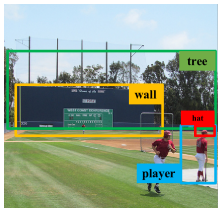

Figure 1: A sample from the TDIUC dataset.

Model

Visual Encoder

For each given image, the appropriate question category is usually determined by specific objects or the relationships among multiple objects contained in the image. Thus, the visual encoder extracts relationships among multiple objects and focuses on certain specific objects to obtain categories suitable for asking.

We first extract K object-level features $\{v_i\}_{i=0}^{K}$ of each image by using a pre-trained object detection model, i.e., Faster R-CNN. To extract the relationships among objects in an image for a question category, we employ VisualBert

*Corresponding author: Yi Cai, ycai@scut.edu.cn

to consider both the language (i.e., question category words, e.g., "Activity Recognition") and vision (i.e., object-level features). For the j-th question category, we obtain the updated object-level features $\{\bar{v}_{i,j}\}_{i=0}^{K}$ of the image, and the word representations h_j^w of the j-th category.

Considering that different objects in the image have different contributions to each question category, we utilize a statistical method (i.e., term frequency-inverse document frequency (TF-IDF)) to calculate the importance of each object in a predefined question category.

We compute $\mathrm{TF}_{i,j}$ (i.e., the occurrences number of i-th object in j-th category divided by the occurrences number of the object in all categories) and IDF_i (i.e., the number of all images divided by the number of images in which the i-th object appears and then take the logarithm) to obtain the TF-IDF weighting value $\alpha_{i,j}$ for the i-th object of j-th category,

$$\alpha_{i,j} = \mathrm{TF}_{i,j} \odot \mathrm{IDF}_i, \tag{1}$$

where $\odot$ denotes element-wise multiplication.

Category Selector

For the j-th question category, we perform an average weighted sum of object-level features $\bar{v}_j$ extracted from the VisualBERT to obtain the image representation I_j,

$$I_j = \sum_{i=1}^{K} \alpha_{i,j} \odot \bar{v}_{i,j}, \tag{2}$$

where the weight is calculated by the TF-IDF, i.e., $\alpha_{i,j}$.

Then, a fully connected layer and sigmoid layer are introduced to obtain the probability of the category p_j based on the image representation I_j, where $p_j > m$ represents the j-th category that is appropriate and m is selection threshold.

We consider a category to be labeled as "positive" if most of the category is contained in the ground truth.

Question Generator

We use an LSTM as the decoder to generate a question for each appropriate question category. For the decoder, we use the representation of the j-th question category h_j^w from VisualBert for initialization. In decoder step t of the j-th category, the decoder module utilizes the image representation I_j and the last generated word embedding w_{t-1} of step t to generate the current hidden state,

$$s_{t,j} = \mathrm{LSTM}(w_{t-1,j}, s_{t-1,j}, I_j). \tag{3}$$

Following the previous work (Xie et al. 2021), we learn a fully connected layer over $s_{t,j}$ and then utilize a softmax activation to get the word probability distribution.

Experiment

Dataset & Settings. We evaluate our model by using the TDIUC dataset, which contains 167,437 images, and 1,654,167 QA pairs, with 12 question categories. We remove the "absurd" category which is not suitable for this task. We set the object number K=36, selection threshold m=0.5.

Experiment Results. We use GRNN with VGGNet and Faster R-CNN as encoders respectively and QT (Fan et al. 2018) which is based on "wh-" question categories as baselines. We use the BLEU method to evaluate the quality of the generated questions. mBLEU is used to evaluate the diversity of the generated questions. We use the Beam Search method on the GRNN model to evaluate its diversity metric. The results of the experiments are shown in Table 1.

Model	BLEU4 ↑	mBLEU1↓	mBLEU3↓
GRNN-VGGNet	15.79	45.06	33.04
GRNN-RCNN	**16.37**	48.57	35.83
QT	8.47	39.43	25.38
CGVQG w/o TFIDF	8.78	**27.10**	14.32
CGVQG	9.81	27.44	**8.48**

Table 1: Experiments on different models.

The experiment shows that our model outperforms other models in terms of diversity metrics. This shows that categories can help models focus on different objects and generate questions in different scenarios. In quality evaluation, our model outperforms QT but is worse than GRNN. Considering that humans tend to choose salient objects to ask multiple questions about an image, these questions may be similar to the question generated by GRNN, which usually focuses on salient features when generating questions. However, when questions are generated by our model based on different categories, they focus on different objects but are not salient in the image, thus the generated questions are far from the ground truth.

Conclusion

In this paper, we focus on the VQG task that aims at generating diverse questions corresponding to different question categories. The category selector in our model aims to judge the appropriate question category based on the objects contained in the image. Then the question generator aims to generate the corresponding category questions. Experimental results show that our proposed model is able to generate diverse questions, which can be more effectively applied to various scenarios.

Acknowledgements

This work was supported by National Natural Science Foundation of China(62076100), and Fundamental Research Funds for the Central Universities, SCUT(x2rjD2220050), the Science and Technology Planning Project of Guangdong Province(2020B0101100002).

References

Fan, Z.; Wei, Z.; Li, P.; Lan, Y.; and Huang, X. 2018. A Question Type Driven Framework to Diversify Visual Question Generation. In *proc. of IJCAI*, 4048–4054.

He, B.; Xia, M.; Yu, X.; Jian, P.; Meng, H.; and Chen, Z. 2017. An educational robot system of visual question answering for preschoolers. In *proc. of ICRAE*, 441–445.

Xie, J.; Cai, Y.; Huang, Q.; and Wang, T. 2021. Multiple Objects-Aware Visual Question Generation. In *proc. of ACM MM*, 4546–4554.