Visual Question Generation from Radiology Images

Mourad Sarrouti **Asma Ben Abacha** **Dina Demner-Fushmen**
National Library of Medicine, National Institutes of Health
Bethesda, MD
{mourad.sarrouti, asma.benabacha}@nih.gov, ddemner@mail.nih.gov

Abstract

Visual Question Generation (VQG), the task of generating a question based on image contents, is an increasingly important area that combines natural language processing and computer vision. Although there are some recent works that have attempted to generate questions from images in the open domain, the task of VQG in the medical domain has not been explored so far. In this paper, we introduce an approach to generation of visual questions about radiology images called VQGR, i.e. an algorithm that is able to ask a question when shown an image. VQGR first generates new training data from the existing examples, based on contextual word embeddings and image augmentation techniques. It then uses the variational auto-encoders model to encode images into a latent space and decode natural language questions. Experimental automatic evaluations performed on the VQA-RAD dataset of clinical visual questions show that VQGR achieves good performances compared with the baseline system. The source code is available at https://github.com/sarrouti/vqgr.

1 Introduction

VQG refers to generating natural language questions based on the images contents. It is a new and exciting problem that combines both natural language processing (Sarrouti and Alaoui, 2017, 2020) and computer vision techniques (Mostafazadeh et al., 2016; Zhang et al., 2016). The motivation for the VQG task is two-fold: (1) generating large scale Visual Question Answering (VQA) pairs to produce more training data at little cost (Ben Abacha et al., 2019) and (2) improving efficiency of human annotation for VQA datasets construction (Li et al., 2018). In addition to the aforementioned motivations, medical VQG could also benefit both doctors and patients. For example, patients could use questions provided by VQG systems to better understand medical images and start a conversation with their doctors. Moreover, such systems could support medical education, medical decision, and patient education (Lau et al., 2018).

A few recent works have attempted to generate questions from images in the open domain. However, the task of VQG in the medical domain has not been studied or explored. One major problem with medical VQG is the lack of large scale labeled training data which usually requires huge efforts to build.

In this paper, we introduce VQGR, a VQG system that is able to generate natural language questions when shown radiology images. In summary, this paper makes the following contributions:

1. To the best of our knowledge, generating questions based on images contents has not been explored in the medical domain. This work is the first attempt to generate questions about radiology images.
2. In the medical domain, the lack of large sets of labeled data makes training supervised learning approaches inefficient. To overcome the data limitation of medical VQG, we present data augmentation on both the images and the questions.
3. VQGR is based on the variational auto-encoders architecture and designed so that it can take a radiology image as input and generate a natural question as output.
4. Experimental evaluations performed on the VQA-RAD dataset of clinical questions and radiology images show that VQGR is effective.

Proceedings of the First Workshop on Advances in Language and Vision Research, pages 12–18

The paper is organized as follows: Section 2 surveys related work. Section 3 describes the proposed VQG approach. Section 4 presents experimental results and discussion.

2 Related Work

Question generation, an increasingly important area, is the task of automatically creating natural language questions from a range of inputs, such as natural language text (Kalady et al., 2010; Kim et al., 2019; Li et al., 2019), structured data (Serban et al., 2016) and images (Mostafazadeh et al., 2016). In this work, we are interested in generating questions from medical images. VQG in the open-domain benefited from the available large annotated datasets (Agrawal et al., 2015; Goyal et al., 2019; Johnson et al., 2017). There is a variety of work studying generative models for generating visual questions in the open domain (Masuda-Mora et al., 2016; Zhang et al., 2016). Recent VQG approaches have used autoencoders architecture for the purposes of VQG (Jain et al., 2017; Li et al., 2018; Krishna et al., 2019). The successes of these systems have primarily been a result of variational autoencoders (VAEs) (Kingma and Welling, 2013). Conversely, VQG in the medical domain is still a challenging and under-explored task (Hasan et al., 2018; Ben Abacha et al., 2018, 2019).

Although a high-quality manually created medical VQA dataset exists, VQA-RAD (Lau et al., 2018), this dataset is too small for training and there is a need for VQG approaches to create training datasets of sufficient size. Generating new training data from the existing examples through data augmentation is an effective approach that has been widely used to handle the data insufficiency problem in the open domain (Şahin and Steedman, 2018; Kobayashi, 2018). Due to the problem of data scarcity in medical VQG, we automatically generate new training data. In this paper, we present VQGR, a VQG system capable of generating questions about radiology images. The system is based on the VAE architecture and data augmentation.

3 Methods

The goal of this study is to generate natural language questions based on radiology image contents. The overview of VQGR is shown in Figure 1.

3.1 Data Augmentation

Questions. We generated new training examples based on question augmentation. For a given medical question q, we generate a set of new questions. During the augmenting process, we use all the VQA-RAD training data $D = \{q_i\}_{i=1}^{n}$ where n is the number of training questions. We expand each training question q_i into a set of instances q_i^k where k is the number of derived pairs for each training question. To do so, we first select nouns and verbs as candidate words, using the part-of-speech tags *NN, NNS, NNPS, NNP, VBD, VBP, VBN, VBG, VBZ, VB*[1]. Each candidate word is then replaced by contextually similar words using Wiki-PubMed-PMC embedding[2] which was trained using four million English Wikipedia, PubMed, and PMC articles. Similar words k for a given word are retrieved from the word embeddings space using cosine similarity. We compute cosine similarity between a weight vector of the given word w_i in the question and the vectors for each word w_j in the pre-trained word embeddings. We carried out several experiments with $k = \{5, 10, 15, 20, 30\}$ and found that the best result in terms of evaluations metrics (described in Subsection 4.2) can be achieved with $k = 20$. For instance, for a given question “Are the kidneys normal?”, we generate the followings questions: “Were the kidneys normal?”, “Are the pancreas normal?, “Are the intestines normal?”, “Are the isografted normal?”, Are the livers normal?, “Are the lungs normal?”, “Are the organs normal?”, etc.

Images. We also generated new training instances based on image augmentation techniques. To do so, we applied flipping, rotation, shifting, and blurring techniques to all VQA-RAD training images.

3.2 Visual Question Generation

The proposed VQGR system is based on the variational autoencoders architecture (Kingma and Welling, 2013). It first encodes the image before generating the question. VAEs consist of two neural network modules, encoder, and decoder, for learning the probability distributions of data $p(x)$. The encoder creates a latent variable z from raw data x and transforms it into latent space $z - space$. The decoder plays the role of recovering x using z extracted from the latent space. Let $q(z|x)$

[1] We used NLTK to perform part-of-speech tagging.

[2] http://evexdb.org/pmresources/vec-space-models/

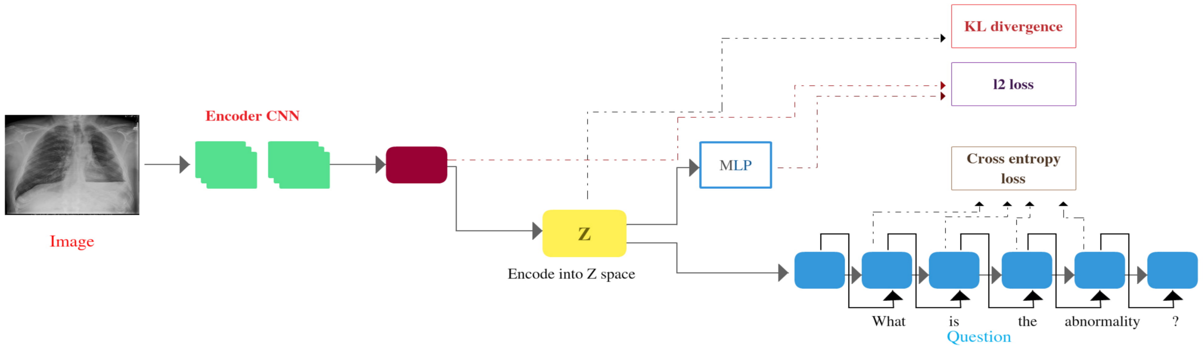


Figure 1: Overview of VQGR: a VQG model from radiology images.

and $p(x|z)$ be the probability distributions of the encoder and the decoder, respectively. Training of the encoder and decoder proceeds by maximizing marginal likelihood $\log p(x)$. Expanding the equation and finding the evidence lower bound (ELBO) yields:

$$\begin{aligned}\log p(x) \geq E_{z\sim q_\theta(z|x)}[\log p_\phi(x|z)] - \\ KL(q_\theta(z|x)||p(z)) \\ = ELBO\end{aligned} \tag{1}$$

The loss function of VAEs is the negative log-likelihood with a regularizer. The loss function l_i for datapoint x_i is:

$$\begin{aligned}l_i(\phi,\theta) = -E_{z\sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + \\ KL(q_\theta(z|x_i)||p(z))\end{aligned} \tag{2}$$

where $E_{z\sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)]$ is the reconstruction error and $KL(q_\theta(z|x)||p(z))$ is the Kullback-Leibler divergence regularization term. ϕ and θ, the parameters for the decoder distribution $p_\phi(x|z)$ and the encoder distribution $q_\theta(z|x)$ respectively.

Given an image v, a CNN is used for obtaining a feature map and encoding the dense vectors h_v into a latent (hidden) representation z-space. It then reconstructs the inputs from the z-space using a simple Multi Layer Perceptron (MLP) which is a neural network with fully connected layers. It generates the reconstructed image features $\hat{h_v}$ and optimizes the model by minimizing the following l_2 loss:

$$L_v = ||h_v - \hat{h_v}||_2 \tag{3}$$

We used the reparameterization trick (Kingma and Welling, 2013), to generate means μ_z and standard deviations σ_z, combine it with a sampled unit Gaussian noise ϵ to generate:

$$z = \mu_z + \epsilon\sigma_z \tag{4}$$

We assumed that z follows a multivariate Gaussian distribution with diagonal covariance.

Finally, it uses a decoder LSTM to generate the question $\hat{q}$ from the z-space. The decoder takes a sample from the latent dimension z-space, and uses that as an input to output the question $\hat{q}$. It receives a "start" symbol and proceeds to output a question word by word until it produces an "end" symbol. We used the Cross Entropy loss function to evaluate the quality of the neural network and to minimize the error L_g between the generated question $\hat{q}$ and the ground truth question q. The generation of each word of the question can be written:

$$\hat{w_t} = \arg\max_{w\in\mathbb{W}} p(w|v, w_0, ..., w_{t-1}) \tag{5}$$

where $\hat{w_t}$ is the predicted word at t step, $\mathbb{W}$ denotes the word vocabulary, and $\hat{w_i}$ represents the i-th ground-truth word.

The final loss of VQGR is as follows:

$$L_{VQGR} = \lambda_1 L_g + \lambda_2 KL + \lambda_3 L_v \tag{6}$$

where KL is Kullback-Leibler divergence which allows to know how well our variational posterior $q(z|v)$ approximates the true posterior $p(z|v)$. $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters that control the variational loss, the question generation loss, and the reconstruction loss, respectively.

4 Experimental Results

4.1 Dataset

In this study, we used the VQA-RAD dataset (Lau et al., 2018) of clinical visual questions and images. It contains 315 images and 3,515 corresponding questions. Each image is associated with more than one questions. In this work, we are particularly interested in questions about 'Modality", "Abnormality", "Organ", and "Plane".

The training set consists of 69,598 questions and 1,673 images after applying data augmentation, and 1,269 questions and 239 images before data augmentation. Table 1 presents the number of questions and images associated to each of the selected categories. The test set contains 100 reference questions with associated categories and images.

Category	#Questions	#Images
Abnormality	397/18642	112/784
Modality	288/5534	54/378
Organ	73/16408	135/945
Plane	163/9216	99/693
Other	348/19798	81/567
Total	1269/69598	239/1673

Table 1: The number of questions and images associated to each category. The values after "/" represent the number of questions and images obtained by data augmentation techniques.

4.2 Evaluation Metrics

To investigate the performance of the proposed VQGR system, we perform both automatic and manual evaluations.

4.2.1 Automatic evaluation

VQG is a sequence generation problem. Therefore, we used a variety of language modeling evaluation metrics such as BLEU, ROUGE, METEOR, and CIDEr to measure the similarity between the system-generated questions and the ground-truth questions of the test set. We use the evaluation package published by (Chen et al., 2015).

4.2.2 Human evaluation

For human evaluation, we follow the standard approach in evaluating text generation systems (Koehn and Monz, 2006), as used for question generation by (Du and Cardie, 2018; Hosking and Riedel, 2019). We manually checked the generated questions and rated them in terms of relevancy, grammaticality, and fluency. The relevancy of a question is determined by the relationship between the question, image and category. Grammaticality refers to the conformity of a question to the grammar rules. Fluency refers to the way individual words sound together within a question. The rating process has been done by two experts at the U.S. National Institutes of Health. For each rating scheme, the human raters are required to give a rating ranging from 1 to 3 scale (1 = completely not satisfying the rating, 3 = fully satisfying the rating scheme).

4.3 Implementation Details

We implemented the VQGR and the baseline models using PyTorch. We used ImageNet-pretrained ResNet-50 (He et al., 2016) provided by PyTorch as the image encoder and do not fine-tune its weights. LSTM decoder is used for generating questions. All images are resized to 224*224. Adam optimiser with a learning rate of 0.0001 and a batch size of 32 is used. All models are trained for 40 epochs and the best validation results are used as final results. The source code is publicly available at https://github.com/sarrouti/vqgr.

4.4 Results and Discussion

Table 2 presents a comparison between the VQGR and the baseline systems in terms of multiple language modeling metrics. The baseline system is trained on the original VQA-RAD dataset without data augmentation. VQGR is trained on the data generated by our data augmentation techniques. We can see that VQGR performs significantly better across all metrics in comparison to the baseline model. The results demonstrate that our data augmentation techniques helped considerably, producing a significant improvement. As we discussed above, one major challenge in medical VQG is the lack of large training datasets. To avoid overfitting the model, small data might require models that have low complexity. Whereas the proposed VAE requires a large amount of training data as it tries to learn deeply the underlying data distribution of the input to output new sequences.

Model	B1	B2	B3	B4	M	RL	C
Baseline	31.4	14.6	7.8	3.2	10.4	38.8	21.1
VQGR	55.0	43.3	37.9	34.5	29.3	56.3	31.1

Table 2: Automatic evaluation results of the VQGR and the baseline models in terms of BLEU-1 (B1), BLEU-2 (B2), BLEU-3 (B3), BLEU-4 (B4), METEOR (M), ROUGE-L (RL) and CIDEr (C).

Table 3 shows the results of the human evaluation. We randomly selected 20 (image, question) pairs from the test set for a manual evaluation by two experts. Detailed guidelines for the raters are listed in subsection 4.2.2. Inter-rater reliability was calculated on each of the 3 measures.

F1-score for each measure is presented in Table 4. Most of the reliability scores are close to 0.50, which is considered satisfactory reliable. The human evaluation showed that VQGR achieves close to human performance in terms of relevancy, grammaticality, and fluency. We have not reported the human evaluation results of the baseline system since it returns the same trivial question "what is the abnormality in this image?" for all given images. This question could be asked about any radiology image, even a normal image, without even looking at it. Our goal is to develop approaches capable of asking non-trivial questions, which is not possible without understanding the image contents, at least to some extent.

Model	**R**	**G**	**F**	**Score**
VQGR	78.3	93.3	80.0	83.8

Table 3: Human evaluation results in terms of relevancy (R), grammaticality (G) and fluency (F). The score is the average of R, G and F. These numbers are the average of annotators scores and divided by 60 to have them between 0 and 1. The perfect score is 100.

Model	**R**	**G**	**F**
VQGR	0.42	0.27	0.51

Table 4: Inter-rater Reliability based on F1-score (Hripcsak, 2005). R, G and F indicate relevancy, grammaticality and fluency, respectively.

Overall, VQG in the medical domain is a very challenging task, and VQGR provides a practical alternative to generate visual questions about radiology images. Figure 2 provides example questions generated by Lau et al. (2018) (ground truth questions) and the VQGR system. From these samples, we can see that the generated questions are consistent with the ground truth.

5 Conclusion and Future Work

We presented the first attempt to generate visual questions in the medical domain. We first presented a data augmentation method to generate new training questions and images from the VQA-RAD dataset. We then introduced the VQGR model that generates questions from radiology images. The results of the automatic and manual evaluations showed that VQGR outperforms the baseline model by generating fluent and relevant questions.

In the future, we will investigate the use of the generated questions to advance VQA in the medical domain.

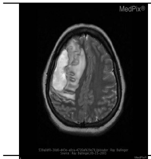
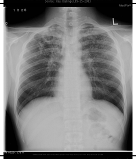
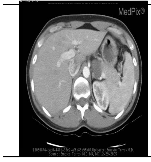
Image	Generated questions vs. ground truth
	what type of mri is used to acquire this image ? mri imaging modality used for this image?
	what is seen in the lung apices ? what abnormalities are in the lung apices ?
	is a ring enhancing lesion present in the right lobe of the liver? is the liver normal ?

Figure 2: Examples of test images with the generated questions (shown in blue) and the ground truth.

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Asma Ben Abacha, Soumya Gayen, Jason J. Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman. 2018. NLM at imageclef 2018 visual question answering in the medical domain. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *Int. J. Comput. Vision*, 127(4):398–414.

Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew P. Lungren. 2018. Overview of imageclef 2018 medical domain visual question answering task. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Tom Hosking and Sebastian Riedel. 2019. Evaluating rewards for question generation models. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics.

G. Hripcsak. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Unnat Jain, Ziyu Zhang, and Alexander Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 5415–5424, United States. Institute of Electrical and Electronics Engineers Inc.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.

Saidalavi Kalady, Ajeesh Elikkottil, and Rajarshi Das. 2010. Natural language question generation using syntax and keywords. In *Proceedings of QG2010: The Third Workshop on Question Generation*, volume 2, pages 5–14.

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. Improving neural question generation using answer separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6602–6609.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation - StatMT 06*. Association for Computational Linguistics.

Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2008–2018.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1).

Jingjing Li, Yifan Gao, Lidong Bing, Irwin King, and Michael R. Lyu. 2019. Improving question generation with to the point context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, volume 33, page 3216–3226.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, and Xiaogang Wang. 2018. Visual Question Generation as Dual Task of Visual Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6116–6124.

Issey Masuda-Mora, Santiago Pascual-deLaPuente, and Xavier Giró i Nieto. 2016. Towards automatic generation of question answer pairs from images. In *Visual Question Answering Challenge Workshop, CVPR 2016*, Las Vegas, NV, USA.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, Berlin, Germany. Association for Computational Linguistics.

Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mourad Sarrouti and Said Ouatik El Alaoui. 2017. A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering. *Journal of Biomedical Informatics*, 68:96–103.

Mourad Sarrouti and Said Ouatik El Alaoui. 2020. Sembionlqa: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. *Artificial Intelligence in Medicine*, 102:101767.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, page 588–598, Berlin, Germany. Association for Computational Linguistics.

Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2016. Automatic generation of grounded visual questions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4235–4243.