Rephrasing visual questions by specifying the entropy of the answer distribution

Kento Terao, Toru Tamaki, Bisser Raytchev, Kazufumi Kaneda
Hiroshima University, Japan

Shun’ichi Satoh
National Institute of Informatics, Japan

Abstract

Visual question answering (VQA) is a task of answering a visual question that is a pair of question and image. Some visual questions are ambiguous and some are clear, and it may be appropriate to change the ambiguity of questions from situation to situation. However, this issue has not been addressed by any prior work. We propose a novel task, rephrasing the questions by controlling the ambiguity of the questions. The ambiguity of a visual question is defined by the use of the entropy of the answer distribution predicted by a VQA model. The proposed model rephrases a source question given with an image so that the rephrased question has the ambiguity (or entropy) specified by users. We propose two learning strategies to train the proposed model with the VQA v2 dataset, which has no ambiguity information. We demonstrate the advantage of our approach that can control the ambiguity of the rephrased questions, and an interesting observation that it is harder to increase than to reduce ambiguity.

1. Introduction

Visual Question Answering (VQA) is one of the most challenging tasks in computer vision [1, 2]: given a pair of question text and image (a visual question), a system is asked to answer the question. There is also a related task called Visual Question Generation (VQG), which generates questions about the given image (Figure 1). These tasks have been attracting a lot of attention in recent years because it has a large potential to impact many applications such as smart support for the visually impaired [3], providing instructions to autonomous robots [4], and intelligent interaction between humans and machines [5].

There may be two different scenarios when asking a visual question: one needs a clear question and the other requires an ambiguous question (Figure 2). The former scenario involves the case of a smart support system for visually impaired people [3] who are asking the system about the image. The context of such a visual question should be as clear as possible so that the system can provide a specific answer to the question. But sometimes users’ questions may be vague, and the system returns multiple answers that are equally possible. This can be a source of confusion, because the user is not certain whether the answers are correct. The latter scenario happens when you want to ask a visual question that might start a conversation or ask opinions. It is good for such visual questions to be unclear, so that different responses can be expected [6]. This kind of ambiguity is due to various factors [7], such as the subjectivity of questions (i.e., asking about opinions on the image) or the difficulty of the task (i.e., counting a number of objects). Such scenarios lead to the need to modify (rephrase, or control) the ambiguity of visual questions. Prior work on VQG either generates questions from answers [8] or categories of answers [9], or learns VQA and VQG simultaneously [8, 10]. However, to the best of our knowledge, no work has handled the ambiguity when generating or rephrasing questions.

In this paper, we propose a model that modifies a given visual question in such a way that the rephrased visual question has the ambiguity specified by the user (bottom of Figure 2). We define the ambiguity of a visual question with...
the entropy value of the answer distribution predicted by a VQA model. Let $A$ be the set of answers, and the entropy $H(A)$ is defined by

$$H(A) = - \sum_{a \in A} P(a) \ln P(a). \quad (1)$$

In general, entropy is large when the distribution is broad, and small when it has a narrow peak. If the entropy is large, we think that the visual question is ambiguous because it results in various answers, and if the entropy is small, it is a clear visual question that provides a unique answer (or fewer possible answers). Our proposed model is closely related to a common VQG model that takes an input image $I$ and generates a question text $Q_G$. The difference is that our model also takes a source question $Q_S$ and a target entropy value $E_T$ as input. The question $Q_G$ generated by our model is then fed into a VQA model to predict an answer distribution, from which the entropy $E_{QG}$ is computed. The proposed model learns to minimize the error between the specified target entropy $E_T$ and the entropy $E_{QG}$ (as shown in Figure 3). In addition, the loss between the generated question $Q_G$ and a target question $Q_T$, which is used as a reference text, is also minimized in order to impose restrictions on the generated question to be similar to questions in the dataset. We also propose two learning strategies to train our model. Our task is novel and hence there are no datasets for this task so far, therefore we utilize an existing VQA dataset, the VQA v2 [11], to train our rephrasing model.

2. Related Work

2.1. VQG

Our work is closely related to VQG, which is proposed by Mostafazadeh et al. [5]. VQG can be done by using image captioning models, however there is a gap from questions generated by humans even if the evaluation score (such as BLEU) is high. Therefore by focusing on the fact that humans use words related to abstract concepts when asking questions, they proposed the VQG task and created a VQG dataset containing more abstract words than image captioning datasets. Their objective is to generate a natural question to help start a conversation about a given image. This is similar to generating a more ambiguous question with higher entropy, but their model is not able to control the question ambiguity.

Other works tried to generate multiple different questions. Zhang et al. [12] proposed a model that can generate visually grounded questions with diverse types for a single image. In this model, not only images but also captions generated by a dense captioning model are used as input. This model can generate multiple types of questions that refer to different image regions described by dense captioning. Fan et al. [13] used a conditional variational autoencoder to generate multiple questions from each question type and a given image. They use two score functions that take into account probabilities of generating the target question and the question type, and generate top $k$ questions. Krishna et al. [9] built a model that maximizes mutual information between the image, the expected answer, and the generated question, based on the observation that VQG should be a task aimed at extracting concepts of a particular category from images. They used a variational autoencoder with two latent spaces; one is between the image and the expected answer, and the other is between the image and the answer category. Questions can be generated by specifying the answer category. These works can generate diverse questions about the image, but the ambiguity of the generated questions has not been taken into account.

2.2. Joint learning of VQG and VQA

Some works use VQG for improving VQA performance. Li et al. [10] proposed a model that simultaneously learns VQA and VQG, and showed that the joint learning based on the MUTAN [14] model improves the performance of VQA. Shah et al. [8] proposed a framework that uses the cycle consistency between VQA and VQG models to make VQA more robust to linguistic variations in visual questions. In this framework, VQG generates a visual question based on the answer predicted by VQA, and VQA predicts the answer to the generated question again.

Our work also uses VQA and VQG as components of the proposed model, however our goal is not to improve the...
2.3. Paraphrasing

Our question rephrasing is also similar to paraphrasing, which expresses the same meaning in different phrases. Liu et al. [15] proposed to modify (or paraphrase) a generated caption text so that modified captions have more diverse expressions like in human-provided captions. They used a score function to evaluate the syntactic complexity to select more descriptive caption candidates. However the syntax of question texts does not reflect the ambiguity of visual questions.

3. Proposed model

The proposed model is shown in Figure 3 (left). Given source visual question \( Q_S \), image \( I \), and target entropy \( E_T \), the rephrase model outputs generated (or rephrased) question \( Q_G \) whose entropy \( E_G \) is close to \( E_T \). Entropy \( E_G \) is calculated from the predicted answer distribution of a pre-trained VQA model, which is frozen during the training of the rephrasing model.

3.1. Model Details

The rephrase model has an encoder-decoder structure (Figure 3 (right)).

3.1.1 Encoder

Unlike VQG models which only take an image as input, our rephrase encoder takes image \( I \), source question \( Q_S \), and target entropy \( E_T \) as input. A pre-trained ResNet152 [16] is used to extract image features from \( I \), and an LSTM is used for extracting text information from the source question \( Q_S \). Target entropy \( E_T \), which is a scalar, is also used as input to specify the desired ambiguity of the rephrased question. These are concatenated and passed to a linear layer.

Furthermore, we use a pre-trained VQA model inside the encoder to modify the image features. This is because we want the model to rephrase the source question, rather than generating an unrelated new text. To this end, we utilize the attention that the VQA model is used. VQA with attention mechanisms is known to work well [17] because it is an important key to correctly answer the question by knowing which part of the image is relevant. Therefore the VQA attention is expected to encode the part used in the source question. By attending to that part and emphasizing the image features with it, the hidden feature passed from the encoder to the decoder is expected to retain information about where is the important part in the image relevant for the source question.

3.1.2 Decoder

The decoder receives the output of the encoder with the attention information, and is expected to generate a question that is similar to the source question in terms of the attention, which is what we want to achieve, that is, rephrasing. The decoder also takes target entropy \( E_T \), as in the encoder, to enforce again the model to take the target entropy into account. These are concatenated and passed to a linear layer, followed by an LSTM to generate a rephrased question, which is represented by a word sequence \( Q_G1, \ldots, Q_{Gn} \).

3.1.3 Entropy computation by VQA

The rephrase model with the encoder-decoder architecture is enough for rephrasing source questions, but we need to check if the rephrased question has the ambiguity specified by target entropy. Therefore we add the VQA model, which is the same as the one used in the encoder.

The rephrased (generated by the decoder) question \( Q_G \) and the same image \( I \) are then passed to the VQA model for predicting the answer distribution. Entropy \( E_G \) is computed from this answer distribution, and expected to express the ambiguity of \( Q_G \).

3.2. Training

There is no dataset for this task, and it is difficult to collect such datasets because the ambiguity of questions is subjective and hard to define. Instead, we propose to use an existing VQA dataset with the following protocol.

3.2.1 Losses

We use two losses. One is an entropy loss \( L_{\text{Ent}} \), which is the error between the entropy \( E_G \) of the generated question and the target entropy \( E_T \):

\[
L_{\text{Ent}} = (E_T - E_G)^2. \tag{2}
\]

This forces the model to generate questions with the specified target entropy value.

The other loss is a VQG loss \( L_{\text{VQG}} \) [8,9], which is the sum of the negative log likelihood over the generated word sequence \( Q_{G0}, \ldots, Q_{Gn} \), given the reference (target) word sequence \( Q_{T0}, \ldots, Q_{Tn} \), as follows:

\[
L_{\text{VQG}} = -\frac{1}{n} \sum_{i=1}^{n} \log p(Q_{Gi} | Q_{Ti}). \tag{3}
\]

Putting them together, the final loss for our model is

\[
L = L_{\text{VQG}} + \lambda L_{\text{Ent}},
\]

where \( \lambda \) is a hyper parameter.

3.2.2 Learning strategies

We introduce two different learning strategies: (1) noise and (2) sampling.
Setting 1: Noise  We use the source question $Q_S$ as the target question $Q_T$. Let $E_S$ be the source entropy, that is, the entropy value computed from the answer distribution of the VQA model with the image $I$ and the source question $Q_S$. The target entropy $E_T$ is defined by

$$E_T = E_S + \epsilon, \epsilon \sim U(-1, 1) \tag{4}$$

where noise $\epsilon$ is subject to a uniform noise $U(-1, 1)$ between $-1$ and $+1$.

The motivation for this setting is to enforce the model to generate a question $Q_G$ similar to the source question $Q_S$ (via the VQG loss), but having different entropy value (via the entropy loss). Therefore the model is expected to behave like a rephraser.

Setting 2: Sampling  Many VQA datasets have a set of multiple questions $Q_I$ for the same image $I$. Therefore, we use one of the different questions that belong to the same image $I$ to which the source question $Q_S$ belongs. The target question $Q_T$ is randomly sampled from $Q_I$, excluding $Q_S$, that is,

$$Q_T \sim Q_I/Q_S. \tag{5}$$

Target entropy $E_T$ is computed by using $Q_T$ with the VQA model.

In this setting the generated question $Q_G$ may not be a rephrased version of the source question $Q_S$ because the sampled target question $Q_T$ might not be asking similar concept with $Q_S$. However the entropy $E_G$ is expected to be similar to the specified target entropy $E_T$, because $E_T$ is computed from the actual question sentence $Q_T$. In contrast, the noise setting uses the same question for $Q_S$ and $Q_T$ but asks the model to change the entropy, which might confuse the model.

We compare these two settings in the experiments.

4. Experiments

4.1. Datasets and setting

We use VQA v2 [11], a standard benchmark dataset for VQA and used in annual challenges since 2016\(^1\). It consists of training, validation, and test sets. To train the proposed model, we use the training set (82,783 images, 443,757 questions, and 4,437,570 answers). We use the validation set (40,504 images, 214,354 questions, and 2,143,540 answers) for rephrasing visual questions and analysis.

We use Pythia v0.1 [18,19] as a pre-trained VQA model for the encoder and for entropy computation. As in prior work [17,20–22], 3129 answers in the training set that occur at least 8 times are chosen as candidates, which results in a multi-class problem predicting answer distributions of 3129 dimension.

4.1.1 Implementation Details

The LSTMs used in the encoder and decoder have one layer with hidden size of 512. The batch size was 64 and the maximum iterations were set to 44000. Adam with a learning rate of 0.0005 was used for the optimization. To back-propagate the losses over the generated questions, we used the Gumbel-Softmax trick [23] with a temperature parameter $\tau = 0.01$.

4.1.2 Evaluation metrics

We use the error between the target entropy $E_T$ and the generated entropy $E_G$ in order to evaluate whether the proposed model generates a question with the specified ambiguity.

In addition, we use standard similarity metrics (BLEU-4, CIDEr, METEOR, ROUGE-L [24]), commonly used for captioning and VQG, to evaluate the quality of the generated questions.

We also use the number of unique questions to evaluate the diversity of the generated questions.

\(^1\)https://visualqa.org/
4.1.3 Comparisons

We compare the following three different settings for each of the two learning strategies.

- **Noise/Sampling Pretrain**: The model learns only with the VQG loss without using the entropy loss. This is equivalent to setting \( \lambda = 0 \).

- **Noise/Sampling**: The model learns from scratch using both the VQG loss and entropy loss with an equal weight \( \lambda = 1 \).

- **Noise/Sampling-FT**: The model learns via fine-tuning (FT) using both the VQG loss and entropy loss with an equal weight \( \lambda = 1 \). The pre-training with the VQG loss only (i.e., Pretrain setting) is used for a warm start.

4.2. Rephrasing

When rephrasing source questions, our model requires the target entropy \( E_T \). For evaluating rephrasing results, we vary \( E_T \) with respect to \( E_S \). More specifically, given a source question \( Q_S \), we compute source entropy \( E_S \), and then add a scalar \( \Delta \) to \( E_S \): \( E_T = E_S + \Delta \). Here \( \Delta \) is varied from \(-2.0 \) to \(+2.0 \) by the step of \( 0.5 \); i.e., \( \Delta \in \{-2.0, -1.5, -1.0, -0.5, 0.0, +0.5, +1.0, +1.5, +2.0 \} \).

Note that entropy should be positive, hence we exclude questions when \( E_S + \Delta < 0 \). This means that there are less valid questions for large negative \( \Delta \). The number of questions and unique questions (diversity, see below) for each value of \( \Delta \) is shown in Table 1.

Table 2 shows quantitative results for each value of \( \Delta \) for six different situations. The column \(|E_T - E_G|\) shows averages and standard deviations (stds) of errors between the specified target entropy \( E_T \) and the entropy \( E_G \) of the rephrased question. This should be small so that rephrased questions \( Q_G \) have the desired ambiguity with the specified entropy.

Columns BLUE-4, CIDEr, METEOR, and ROUGE-L are similarity scores between source and rephrased questions. Unlike captioning or VQG tasks where generated texts should be close to those in the training set, higher scores do not mean better results in this task. This is because we rephrase (modify) the source question and generate a different question with different ambiguity, therefore these scores are not necessary high. In other words, if these scores are too high, then it means that the model doesn’t rephrase (or modify) the source question.

The column Diversity shows the number of unique questions generated by the model. The original VQA v2 validation set has at most 81565 different unique questions (the column Diversity of Table 1). The diversity score often decreases when the variety of questions generated by the model is limited (i.e., the same questions are often generated).

In Noise Pretrain and Sampling Pretrain settings, the model was trained without the entropy loss and can not control the generated entropy \( E_G \). Therefore their errors \( E_T - E_G \) are consistently high. The model is trained so that the rephrased questions should be identical to the source questions, then similarity scores are quite high, particularly for Noise Pretrain.

In the Noise and Sampling setting, the model was trained from scratch without pretraining, therefore the training was not stable. Entropy errors are smaller than those of Pretrain settings, however similarity scores and diversity are extremely low. A possible reason might be that reasonable questions cannot be generated at the initial stage of learning, while the model tries to reduce the entropy loss.

In Noise-FT and Sampling-FT settings, the model was fine-tuned after the pretraining. This achieves often the smallest entropy error for each of the \( \Delta \) values, and similarity and diversity scores are higher than those without pretraining.

The Noise-FT model can generate a variety of questions (high diversity), but has problems generating questions when increasing entropy (positive \( \Delta \) values). On the other hand, the Sampling-FT model can control the entropy, but the variety of questions that can be generated is small (low diversity).

We further investigate how the model controls entropy errors by using box plots shown in Figure 4. The horizontal axis \( (E_T - E_S) \) corresponds to each \( \Delta \) value in Table 2 from \(-2.0 \) to \(+2.0 \). The vertical axis is signed entropy difference \( E_G - E_S \) while the entropy error column in Table 2 shows absolute errors \( |E_T - E_G| \). Ideally, this plot should be linear with a unit slope because \( E_G \) and \( E_T \) are the same in the ideal case. We can see that Noise Pretrain is almost flat, which means the model has no control on the entropy. In contrast, Sampling Pretrain has too much large slope and it fails to control entropy too. The best-looking result is Sampling-FT; it has almost a unit slop, and \( E_G - E_S \) is almost the same with \( E_T - E_S \). For the other three cases, it looks difficult to make \( E_G - E_S \) larger when \( E_T - E_S \).
is positive. Thus it turns out that increasing entropy (more ambiguous) is harder than decreasing entropy (less ambiguous).

Figure 5 shows rephrasing results for three visual questions with $\Delta \in \{-1, 0, 1\}$ in four situations. In Noise Pretrain the model tends to generate the same question even if the target entropy is changed, indicating that the model cannot control the entropy. In Noise-FT and Sampling Pretrain, the model generates different questions but the entropy increases when $\Delta$ is increased. In Sampling-FT entropy errors increased with absolute $\Delta$.

### 4.3. Impact of VQA attention

For two situations, Sampling-Pretrain and Sampling-FT, we investigate the effect of the presence or absence of the VQA attention on rephrasing questions. When the model does not use the VQA attention, image features of ResNet152 are directly used. Results are shown in Table 3.
Table 3. Entropy errors, similarity and diversity scores of rephrasing results with and without VQA attention. Rows with “w/o A” stands for without attention, otherwise with attention.

\[ E_A \leq E_T \leq E_G \]

| Sampling | $\Delta$ | $E_T$ | $E_G$ | $E_T - E_G$ | Diversity |
|----------|----------|-------|-------|-------------|-----------|
| Pretrain | -2.0 | 0.89 | 0.85 | 0.04 | 0.82 |
| FT w/o A | -0.5 | 0.10 | 0.08 | 0.02 | 0.66 |
| FT | 1+0.5 | 1.02 | 1.00 | 0.02 | 0.95 |
| Pretrain | 1+0.5 | 1.02 | 1.00 | 0.02 | 0.95 |

Figure 7. Effects on rephrasing with VQA attention. Rows with “w/o A” stand for without attention, otherwise with attention. There are three visual questions (images $I$ are on the left, and source questions $Q_S$ are at the top of each table. Top-1 answers predicted by VQA are shown for reference.

4.4. Weight $\lambda$ for entropy loss

For all experiments above the weight of the entropy loss was $\lambda = 1$, but this weight should have a large impact on the result because it trades off between a simple VQA model and our entropy-aware rephrasing model. Table 4 shows results for Sampling-FT with different values of $\lambda$. Results are shown in Table 4, Figure 8, and Figure 9.

The entropy errors become small when $\lambda = 10$ or $\lambda = 100$ is too large because the entropy error increases. However similarity scores for $\lambda = 10$ and 100 are quite low because grammatically correct question sentences are not generated (as shown in Fig. 9) and training becomes unstable. The case of $\lambda = 0.01$ has higher similarity scores because it is close to the Sampling Pretrain situation, and questions similar to source questions are generated.

5. Conclusion

In this paper, we have proposed a task of rephrasing visual questions so that the rephrased question has the specified ambiguity in terms of entropy. Our proposed model has an encoder-decoder architecture, followed by a pre-train...
VQA model for entropy computation. Experimental results when using the proposed learning strategies with the VQA v2 dataset have shown that the model can minimize the error to the specified entropy value.

One of the limitations of our work is that rephrased questions are often irrelevant to the given image while entropy is close to the specified value. To tackle this limitation, it may be necessary to explore architectures for extracting the concept of the image in addition to VQA attention and applying it to the rephrased question. Another limitation is the small diversity of the rephrased questions. To avoid the generation of similar questions, the variational approach [9] might be one of the promising solutions. And last but not least, the construction of datasets specific to this task is the most challenging part of this work and we will continue to

---

**Table 4. Entropy errors, similarity and diversity scores of rephrasing results for different values of weight $\lambda$. Rows with “w/o A” stands for without attention, otherwise with attention.**

| $\Delta = E_T - E_S$ | Diversity |
|-----------------------|-----------|
| 0.01                  | 0.10      |
| 0.1                  | 0.14      |
| 1.0                  | 0.24      |

---

**Figure 8.** Effect of weight $\lambda$ on entropy errors for each value of $\Delta$ over all questions. The horizontal axis is the specified value of $\Delta = E_T - E_S$, and the vertical axis is the entropy error $E_G - E_T$ (without taking absolute).

**Figure 9.** Effects on rephrasing for different values of the weight $\lambda$. There are three visual questions (images 1, 2, and 3 shown above). There are three visual questions (images 4, 5, and 6 shown above). There are three visual questions (images 7, 8, and 9 shown above).
explore how to define the ambiguity of visual questions for dataset annotations.

**Acknowledgement**

We would like to thank Yoshitaka Ushiku for his comments as a mentor of the PRMU mentorship program. This work was supported by JSPS KAKENHI grant number JP16H06540.

**References**

[1] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, “Visual question answering: A survey of methods and datasets,” Computer Vision and Image Understanding, vol.163, pp.21 – 40, 2017. Language in Vision. 1

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, and D. Parikh, “VQA: Visual Question Answering.” International Conference on Computer Vision (ICCV), 2015. 1

[3] D. Gurari, Q. Li, A.J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J.P. Bigham, “Vizwiz grand challenge: Answering visual questions from blind people,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 1

[4] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, “Embodyed Question Answering,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 1

[5] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J.M. Moura, D. Parikh, and D. Batra, “Visual Dialog.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1

[6] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende, “Generating natural questions about an image,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, pp.1802–1813, Association for Computational Linguistics, Aug. 2016. 1, 2

[7] N. Bhattacharya, Q. Li, and D. Gurari, “Why does a visual question have different answers?” The IEEE International Conference on Computer Vision (ICCV), October 2019. 1

[8] M. Shah, X. Chen, M. Rohrbach, and D. Parikh, “Cycle-consistency for robust visual question answering.” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 1, 2, 3

[9] R. Krishna, M. Bernstein, and L. Fei-Fei, “Information maximizing visual question generation.” IEEE Conference on Computer Vision and Pattern Recognition, 2019. 1, 2, 3, 8

[10] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, X. Wang, and M. Zhou, “Visual question generation as dual task of visual question answering,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 1, 2

[11] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering,” Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 2, 4

[12] S. Zhang, L. Qu, S. You, Z. Yang, and J. Zhang, “Automatic generation of grounded visual questions,” Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pp.4235–4243, 2017. 2

[13] Z. Fan, Z. Wei, P. Li, Y. Lan, and X. Huang, “A question type driven framework to diversify visual question generation,” Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, pp.4048–4054, 2018. 2

[14] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, “Mutan: Multimodal Tucker fusion for visual question answering.” The IEEE International Conference on Computer Vision (ICCV), Oct 2017. 2

[15] L. Liu, J. Tang, X. Wan, and Z. Guo, “Generating diverse and descriptive image captions using visual paraphrases,” The IEEE International Conference on Computer Vision (ICCV), October 2019. 3

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. 3

[17] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 3, 4

[18] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, and D. Parikh, “Pythia v0.1: the winning entry to the VQA challenge 2018.” CoRR, vol.abs/1807.09956, 2018. 4
[19] A. Singh, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, and D. Parikh, “Pythia—a platform for vision & language research,” SysML Workshop, NeurIPS, 2018. 4

[20] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” arXiv preprint arXiv:1908.02265, 2019. 4

[21] D. Teney, P. Anderson, X. He, and A. van den Hengel, “Tips and tricks for visual question answering: Learnings from the 2017 challenge,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 4

[22] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, “Deep modular co-attention networks for visual question answering,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 4

[23] E. Jang, S. Gu, and B. Poole, “Categorical reparameterization with gumbel-softmax,” 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, 2017. 4

[24] S. Sharma, L. El Asri, H. Schulz, and J. Zumer, “Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation,” CoRR, vol.abs/1706.09799, 2017. 4