Beyond VQA: Generating Multi-word Answer and Rationale to Visual Questions

Radhika Dua
KAIST
radhikadua@kaist.ac.kr

Abstract

Visual Question Answering is a multi-modal task that aims to measure high-level visual understanding. VQA models are restrictive in the sense that answers are obtained via classification over a limited vocabulary (in the case of open-ended VQA), or via classification over a set of multiple-choice-type answers. In this work, we present a completely generative formulation where a multi-word answer is *generated* for a visual query. To take this a step forward, we introduce a new task: ViQAR (Visual Question Answering and Reasoning), wherein a model must generate the complete answer and a rationale that seeks to justify the generated answer. We propose an end-to-end model to solve this task and show that it generates strong answers and rationales through qualitative and quantitative evaluation.[1]

1 Introduction

Visual Question Answering(VQA) (Antol et al., 2015) is a vision-language task that has seen a lot of attention in recent years. In general, the VQA task consists of either open-ended or multiple choice answers to a question about the image. However, answers in existing VQA datasets and models are largely one-word answers (average length is 1.1 words), which gives existing models the freedom to treat answer generation as a classification task. For the open-ended VQA task, the top-K answers are chosen, and models perform classification over this vocabulary. However, many questions which require commonsense reasoning cannot be answered in a single word. A textual answer for a sufficiently complicated question may need to be a sentence. For example, a question of the type "What will happen.." usually cannot be answered completely using a single word. Current VQA systems are not well-suited for questions of this type. To reduce this gap, more recently, the Visual Commonsense Reasoning (VCR) task (Zellers et al., 2018) was proposed, which is still a classification task. When multi-word answers are required for a visual question, options are not sufficient, since the same 'correct' answer can be paraphrased in a multitude of ways, each having the same semantic meaning but differing in grammar. Going beyond contemporary efforts in VQA, we propose a task, ViQAR, that involves automatically generating both multi-word answers and an accompanying rationale, that serves as a justification for the answer. We also propose an end-to-end methodology to address this task.

[1]This is a fake paper being submitted to KAIST AI605 class as an assignment.

2 Related Work

VQA. A lot of work in VQA is based on attention-based models that aim to 'look' at the relevant regions of the image in order to answer the question (Anderson et al., 2018; Lu et al., 2016; Yu et al., 2017a; Yi et al., 2018). Other recent work has focused on better multimodal fusion methods (Kim et al., 2018, 2017; Fukui et al., 2016; Yu et al., 2017b), the incorporation of relations (Norcliffe-Brown et al., 2018; Li et al., 2019; Santoro et al., 2017), the use of multi-step reasoning (Cadene et al., 2019), and neural module networks for compositional reasoning (Johnson et al., 2017b; Chen et al., 2021; Hu et al., 2017). Visual Dialog (Das et al., 2017; Zheng et al., 2019) extends VQA but requires an agent to hold a meaningful conversation with humans in natural language based on visual questions.

The efforts closest to ours are those that provide justifications along with answers (Li et al., 2018b; Hendricks et al., 2016; Li et al., 2018a; Park et al., 2018; Wu et al., 2019; Park et al., 2018), each of which however also answers a question as a classification task (and not in a generative manner). Datasets have also been proposed for VQA to test visual understanding (Zhu et al., 2016; Goyal et al., 2017; Johnson et al., 2017a); for e.g., the Visual7W dataset (Zhu et al., 2016) contains a richer class of

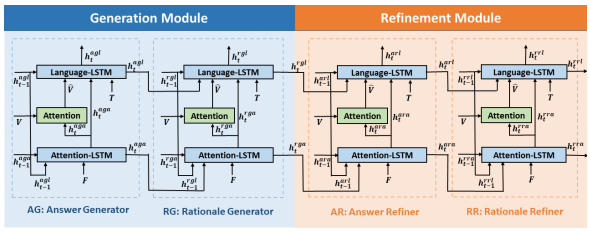

Figure 1: The decoder of our proposed architecture.

questions about an image with textual and visual answers. However, all these aforementioned efforts continue to focus on answering a question as a classification task (often in one word, such as Yes/No), followed by simple explanations.

Visual Commonsense Reasoning (VCR). VCR (Zellers et al., 2018) is a vision-language dataset, which involves choosing a correct answer (among four provided options) for a given question about the image, and then choosing a rationale to justify the answer. The task associated with the dataset aims to test for visual commonsense understanding and provides images, questions and answers of a higher complexity than other datasets such as CLEVR (Johnson et al., 2017a). The dataset has attracted various methods (Zellers et al., 2018; Lu et al., 2019; Dua et al., 2019; Zheng et al., 2019; Talmor et al., 2018; Lin et al., 2019; Brad, 2019), each of which however follow the dataset's task and treat this as a classification problem.

In contrast to all the aforementioned efforts, our work focuses on automatic complete *generation* of the answer, and of a rationale, given a visual query.

3 Proposed methodology

Figure 1 presents an end-to-end, attention-based decoder for answer and rationale generation which is based on an iterative refinement procedure. The encoder part of the architecture generates the spatial image features (Anderson et al., 2018), and use BERT (Devlin et al., 2019) representations to encode question and caption. These features are used by the decoder to generate the answer and rationale for a question. Our decoder has two modules: *generation* (GM) and *refinement* (RM). The GM and RM each consist of two sequential, stacked LSTMs, henceforth referred to as answer generator (AG), rationale generator (RG) and answer refiner (AR), rationale refiner (RR) respectively. Each sub-module (presented inside dashed lines in the figure) is a complete LSTM (Hochreiter and Schmidhuber, 1997). Given an image, question, and caption, the AG sub-module unrolls for l_a time steps to generate an answer. Using the representation of the generated answer from AG, RG sub-module unrolls for l_r time steps to generate a rationale and obtain its representation. Then the AR sub-module uses the features from RG to generate a refined answer. Lastly, the RR sub-module uses the answer features from AR to generate a refined rationale.

Table 1: Quantitative evaluation on VCR dataset; we compare against a basic two-stage LSTM model and a VQA model as baselines; remaining columns are proposed model variants.[CS = cosine similarity]

Metrics	VQA-Baseline	Baseline	Q+I+C (Ours)	Q+I (Ours)	Q+C (Ours)
Univ Sent Encoder CS (Cer et al., 2018)	0.419	0.410	**0.455**	0.454	0.440
Infersent CS (Conneau et al., 2017)	0.370	0.400	0.438	**0.442**	0.426
Embedding Avg CS	0.838	0.840	0.846	**0.853**	0.845
Vector Extrema CS(Forgues and Pineau, 2014)	0.474	0.444	**0.493**	0.483	0.475
Greedy Matching Score(Rus and Lintean, 2012)	0.662	0.633	**0.672**	0.661	0.657
METEOR(Lavie and Agarwal, 2007)	0.107	0.095	**0.116**	0.104	0.103
Skipthought CS (Kiros et al., 2015)	0.430	0.359	**0.436**	0.387	0.385
RougeL(Lin, 2004)	0.259	0.206	**0.262**	0.232	0.236
CIDEr(Vedantam et al., 2015)	0.364	0.158	**0.455**	0.310	0.298
F-BERTScore (Zhang* et al., 2020)	0.877	0.860	**0.879**	0.867	0.868

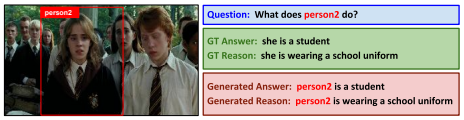

Figure 2: *(Best viewed in color)* Example output from our proposed Generation-Refinement architecture.

4 Performance evaluation of ViQAR

Quantitative results. Quantitative results on the suite of evaluation metrics are shown in Table 1. Since this is a new task, there are no known methods to compare against. We compare our model against a baseline (called *Baseline*) composed of two separate two-stage LSTMs, one for answer and one for the rationale, and a VQA-based method (Anderson et al., 2018)(called *VQA-Baseline*). We show results on three variants of proposed Generation-Refinement model: Q+I+C (question, image and caption as inputs), Q+I (question and image as inputs), and Q+C (question and caption as inputs). Evidently, our model performed the most consistently across all the metrics.

Qualitative results. Figure 2 shows an example where the proposed model generates a meaningful answer with a supporting rationale. Qualitative results indicate that our model is capable of generating answer-rationale pairs to complex questions.

5 Conclusion

In this paper, we propose ViQAR, a novel task for generating a multi-word answer and a rationale given an image and a question. Our work aims to go beyond classical VQA by moving to a completely generative paradigm. We also present an end-to-end generation-refinement architecture. We showed the promise of our model on VCR dataset through qualitative and quantitative evaluation.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

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

Florin Brad. 2019. Scene graph contextualization in visual commonsense reasoning. In *ICCV*.

Remi Cadene, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. 2019. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. 2021. Meta module network for compositional visual reasoning. In *WACV*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL-HLT*.

Gabriel Forgues and Joelle Pineau. 2014. Bootstrapping dialog systems with word embeddings.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *ECCV*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017a. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017b. Inferring and executing programs for visual reasoning supplementary material.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *NeurIPS*.

Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard Product for Low-rank Bilinear Pooling. In *ICLR*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *NIPS*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *WMT@ACL*.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. In *ICCV*.

Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. 2018a. Tell-and-answer: Towards explainable visual question answering using attributes and captions. In *EMNLP*.

Qing Li, Qingyi Tao, Shafiq R. Joty, Jianfei Cai, and Jiebo Luo. 2018b. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *ECCV*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL*.

Jingxiang Lin, Unnat Jain, and Alexander G. Schwing. 2019. Tab-vcr: Tags and attributes based vcr baselines. In *NeurIPS*.

Jiasen Lu, Dhruv Batra, D. Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*.

Will Norcliffe-Brown, Efstathios Vafeias, and Sarah Parisot. 2018. Learning conditioned graph structures for interpretable visual question answering. In *NeurIPS*.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*.

Vasile Rus and Mihai C. Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *BEA@NAACL-HLT*.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Timothy P. Lillicrap. 2017. A simple neural network module for relational reasoning. In *NIPS*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT*.

R. Vedantam, C. L. Zitnick, and D. Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Jialin Wu, Zeyuan Hu, and Raymond J. Mooney. 2019. Generating question relevant captions to aid visual question answering. In *ACL*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *NeurIPS*.

Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. 2017a. Multi-level attention networks for visual question answering. *CVPR*.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017b. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *ICCV*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. 2019. Reasoning visual dialogs with structural and partial observations. In *CVPR*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. In *CVPR*.