Generating Rationale in Visual Question Answering Task

Hammad Abdullah Ayyubi, Ranak Roy Chowdhury
Department of Computer Science and Engineering
University of California San Diego
Email: {hayyubi,rrchowdh}@eng.ucsd.edu

January 21, 2020

1 Team Information

- **Names of Team Members:** Hammad Abdullah Ayyubi, Ranak Roy Chowdhury
- **PIDs of Team Members:** A53283495, A53317421

2 Problem Definition

We propose to work on the task of Visual Commonsense Reasoning (VCR) [1]. This is a recently proposed task at CVPR, 2019. The objective of the task is to predict answer given a question and image. This part is similar to Visual Qustion Answering [2]. The novelty of task lies in the fact that it further asks the model to provide a rational to justify the answer it predicted. This makes the task more interesting and concurrently more difficult.

The baseline provided [3] with the task and leaders of this task [4], [5], [6], [7], [8] approach this task by independently predicting the answer (from question and image) and rationale (from question appended by correct answer and image). Such an approach doesn't require the model to consider rationale or reasons while predicting answer, which defeats the purpose of the task. We propose to generate rationales from the predicted answer as a means to investigate the understanding of the model while predicting answer. Further, by training the model jointly for answer prediction and rationale generation, we seek to force the model to consider rationales while predicting answers.

2.1 Importance of the problem

In recent times, NLP has seen the emergence of impressive Language models [9], [10], [11]. However, studies have shown that they lack commonsense reasoning [12]. The ability to reason, especially from visual data, can power intelligent models to understand abstract and latent information like state of mind of actors, their motivation, context and background. This has huge applications in social robots, healthcare and is in general a step towards more human like abilities.

2.2 Non trivial nature of the problem

The problem of commonsense reasoning is a difficult task in itself. To generate reason/rationale makes it even more difficult. Very limited work have addressed this task in text only data

[13]. And as far as we know, no one has tried to generate rationales from visual data. This is especially non-trivial because the model not only has to understand visual data, but also express rationales in text mode.

3 Related Paper Summary

In this section, we discuss the paper we chose to work with, its key contributions and provide a thorough analysis of its various aspects.

3.1 Paper

The details of the paper are as follows:

- **Paper Name:** From Recognition to Cognition: Visual Commonsense Reasoning [1]
- **Authors:** Rowan Zellers, Yonatan Bisk, Ali Farhadi, Yejin Choi
- **Publication Venue:** Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019

3.2 Contributions

The major contributions of the paper are fourfold:

1. Formalization of a new task, Visual Commonsense Reasoning. Given a challenging question about an image, a machine must answer correctly and then provide a rationale justifying its answer.
2. Presentation of a large-scale multiple-choice QA dataset, named VCR. It consists of 290k multiple choice QA problems derived from 110k movie scenes.
3. Adversarial Matching, a new algorithm for robust multiple-choice dataset creation. It transforms rich annotations into multiple choice questions with minimal bias.
4. Proposal of a new model, Recognition to Cognition Networks **(R2C)**, that aims to mimic the layered inferences from recognition to cognition. R2C models the necessary layered inferences for grounding, contextualization, and reasoning. It also establishes the baseline performance on the new challenge.

3.3 Critical Analysis

The biggest limitation of the paper is that it decouples the reasoning from the question answering step. This inhibits the model from considering any rationales while predicting answers. Another strong assumption is made in the paper in the rationale prediction phase. While predicting rationale, they append question with the correct answer. This follows from the assumption that the answer prediction module can successfully predict answer 100% correctly, which is over-ambitious.

A simple solution to force model to consider rationale could be to jointly train the model for answer prediction and rational prediction. Secondly, while predicting rationale the question could be appended by the predicted answer rather than the correct answer. This will not only force model to consider rationale while predicting answer, but also get rid of the assumption that answer prediction module needs to predict the correct answer with 100% accuracy to feed to rationale prediction module.

References

[1] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6720–6731, 2019.

[2] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "Vqa: Visual question answering," *International Journal of Computer Vision*, vol. 123, p. 4–31, Nov 2016.

[3] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[4] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019.

[5] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," 2019.

[6] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Learning universal image-text representations," 2019.

[7] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," 2019.

[8] C. Alberti, J. Ling, M. Collins, and D. Reitter, "Fusion of detected objects in text for visual question answering," 2019.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[10] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," 2019.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[12] A. Ettinger, "What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models," 2019.

[13] N. F. Rajani, B. McCann, C. Xiong, and R. Socher, "Explain yourself! leveraging language models for commonsense reasoning," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.