6.869 Project Proposal **Yifan Yang, Zhen Guo**
Visual Question Answering in Toy World Due Date: 10/24/2019

1 Introduction & Motivation

Artificial Intelligence (AI) communities believe that multi-discipline research which combines Computer Vision (CV), Natural Language Processing (NLP) and Knowledge Representation & Reasoning (KR) could be a big leap towards the next generation AI algorithms. Such belief is supported by the argument that capturing multi-modal knowledge beyond a single sub-domain may further enable the ability of a machine to perform general intelligent action, achieving AI-complete task with a well-defined evaluation metric. To this end, we propose to work on Visual Question Answering (VQA), a reduced version of the multi-discipline AI task, performed in a toy world where the questions and answer refers to elements in the toy world. Numbers of novel models have been proposed and implemented, such as incorporation of scene text to answer questions [1], probabilistic neural-symbolic model [2], and multimodal relational network which is learned end-to-end [3].
To simplify the project, we constrain the questions to be multiple-choice which only requires our model to pick from a predefined list of possible answers, rather than giving open-ended and free-form response as it was originally proposed [4]. With a subset of the dataset in [4], we aim to build and train a machine learning/deep learning model for the VQA task, evaluating its accuracy and efficiency by the number of questions it answer correctly.

2 Our Approach

We plan to use the state of art deep CNN (convolutional neural network) and LSTM (Long short-term memory) to carry out the VQA task. Specifically, we will be looking into those efficient and high-performance networks such as ResNet [5], MobileNet [6] for image classification and Fast R-CNN [7], YOLO [8] for object detection. Meanwhile, we will be using simple Multilayer Perceptron (MLP) or LSTM to understand the language/questions. Our model will be built using pytorch, then be trained and tested on the abstract scene of VQA dataset. Top-1 and top-k accuracy will be evaluated and reported on the benchmark. Ablation study and comparison with the baseline implementation in [4] will be deployed to demonstrate the effectiveness of our design.

References

[1] Ali Furkan Biten, Rubèn Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. *arXiv preprint arXiv:1907.00490*, 2019.

[2] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural-symbolic models for interpretable visual question answering. *arXiv preprint arXiv:1902.07864*, 2019.

[3] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1989–1998, 2019.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.