Project Brief

Project Title: Multimodal Learning and Reasoning for Visual Question Answering
Name: Vasin Srisupavanich
Supervisor: Dr Srinandan Dasmahapatra

Project Description

Over the last decade, deep learning systems have enjoyed tremendous success in the area of computer vision and natural language processing, such as object recognition and machine translation. However, they still struggle in tasks which require deliberate thinking and multi-step reasoning, as they have relied mostly on the correlation of the dataset to make predictions rather than capturing the true underlying reasoning processes. In order to develop a more general and intelligent learning machine, we would need to seek for new techniques that could enhance deep learning models with the ability to reason.

This project will explore the reasoning capabilities of artificial neural networks in the context of a visual question answering (VQA) problem and investigate how existing deep learning models designed for visual reasoning task can be further improved. The new VQA model will be developed and evaluated on GQA[1], a real-world visual reasoning dataset. This dataset contains over 113K real-world images and 22M compositional questions, and is designed explicitly to test various reasoning skills, including spatial understanding, counting, comparing, logical reasoning, and storing information in memory. In addition, VQA is a challenging multimodal AI research problem, requiring both visual and language understanding, and making a progress in this domain would mean immense real-world impacts, such as aiding visually impaired individuals and enhancing human machine interaction.

Project Goal

The goal of this project is to develop a new VQA model capable of responding to relatively complex questions that require multi-steps reasoning and inference over real-world images. The performance evaluated on the GQA dataset should be comparable to the current state of the art deep learning models.

Scope of the Project

1. **Literature Review**
   The first step is to investigate recent methods which use transformer for joint representation of visual content and natural language, such as VL-BERT[2], and understand how they can be used in a VQA system. The next step is to explore the use of various deep learning techniques, which has been shown to perform well on visual reasoning tasks, such as attention mechanism, memory augmented neural networks and MAC networks[3], and investigate how further improvements can be made.
2. **Model Development**
   First, the baseline VQA model (LSTM, CNN) will be implemented to set the benchmark to improve upon. Then the new model architecture will be designed to incorporate the pre-trained joint representation of image and text with the reasoning module.
3. **Model Evaluation**
   The model will be evaluated using several metrics which are introduced along with the GQA dataset, including accuracy, consistency, validity and plausibility. For instance, the validity metric evaluates whether the model gives valid answer (respond with colour for colour questions or yes/no for binary questions).
4. **Qualitative Analysis**
   Model's interpretability and explainablity will be explored and visualised (e.g. visualisation of attention mechanism) to understand the underlying reasoning process of the model. Also, the generalization ability of the model will be studied using another dataset with different type of questions and images.
5. **Quantitative Analysis**
   The accuracy of each question type will be analysed and compared with existing models to identify the strength and weakness of the model. Moreover, the analysis on data efficiency will be performed to understand how fast the model learns.

[1] Hudson, D.A. and Manning, C.D., 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6700-6709).

[2] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F. and Dai, J., 2019. Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530.

[3] Hudson, D.A. and Manning, C.D., 2018. Compositional attention networks for machine reasoning. arXiv preprint arXiv:1803.03067.