Multimodal Learning and Reasoning for Visual Question Answering

Vasin Srisupavanich | MSc Artificial Intelligence | Supervisor: Dr Srinandan Dasmahapatra

Visual Question Answering

Visual Question Answering (VQA) is the task of answering question about an image.

Visual Reasoning

A VQA task designed explicitly to test reasoning skills, such as spatial understanding, counting, comparing, logical reasoning, and storing information in memory.

"What is the girl eating?"

Hamburger

"Is there a drink in the image?"

Yes

"How many plastic containers are there?"

Two

"What is located to the left of the fries?"

Salad

"What colour is the thing under the food left of the little girl with the yellow shirt?"

Red

Motivation

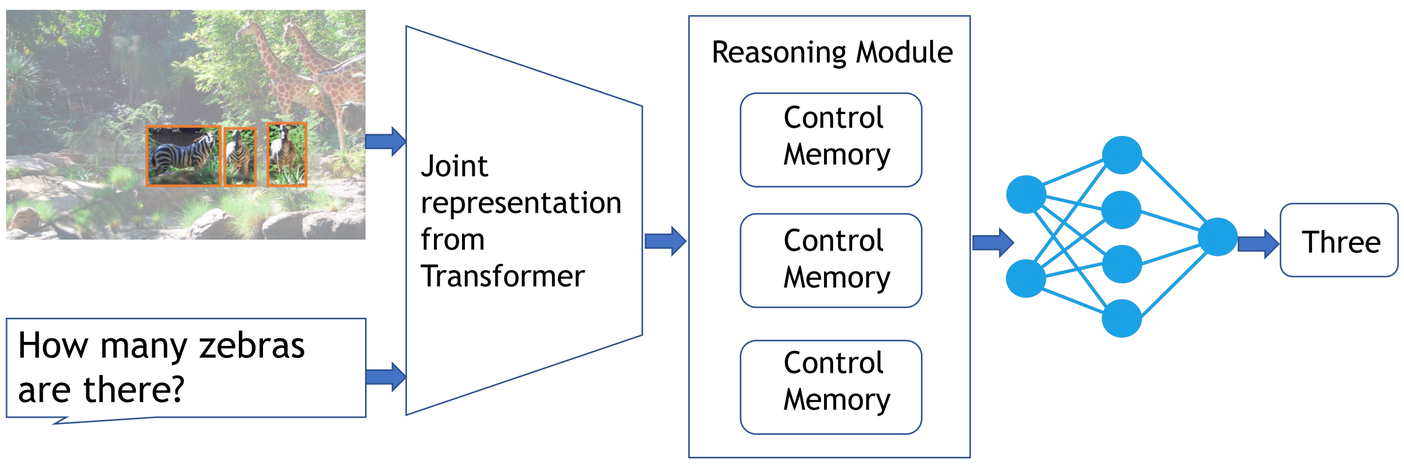
Over the last decade, deep learning systems have enjoyed tremendous success in the area of computer vision and natural language understanding. However, they still struggle in tasks which require deliberate thinking and reasoning process. Creating a machine that can learn to reason has been a long-standing goal in the field of Artificial Intelligence (AI).

VQA is a challenging multi-modal AI research problem, which requires both visual and language understanding. Making a progress in this domain would mean a significant step toward a more general AI. In addition, VQA is a great domain to measure the intelligence and reasoning capabilities of a machine, due to the variety of skills required to tackle different types of question.

Current VQA models tend to exploit the dataset biases without capturing the underlying perception and reasoning process to respond to the given question. Moreover, representations of text and visual features are usually independent and not properly aligned, which often make the model bias towards language modality.

Project Aims

The aim of this project is to develop a VQA model and evaluate how it performs on the visual reasoning dataset (GQA). The model should be capable of responding to relatively complex questions that require multi-steps reasoning and inference over real-world images.



Why the project matters?

Effective use of vast amount of visual data

Enhancing human machine interaction

Aiding visually impaired individuals

GQA Dataset

- A dataset for real-world visual reasoning and question answering [1]
- The dataset contains 113k real-world images, and 22M compositional questions involving a diverse set of reasoning skills

Methods

Model Development

The first step is to implement the baseline VQA model (LSTM, CNN) to set the benchmark to improve upon. Then the new model architecture will be designed to incorporate the joint representation of image and text [2] with the reasoning module [3].

Qualitative Analysis

Model's interpretability and explainablity will be explored and visualised (e.g. visualisation of attention mechanism). It will also be useful to study whether the model can generalize to new dataset with different type of questions or images.

Quantitative Analysis

The accuracy from each question type will be analysed and compared with existing models to identify the strength and weakness of the model. Moreover, analysis on data efficiency should give useful insights on how fast the model learns.

References

[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.

[4] Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B. and Wu, J., 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. arXiv preprint arXiv:1904.12584.

* Images in this poster are taken from [1], [4]

UNIVERSITY OF Southampton