Multimodal Learning for Visual Question Answering

1 INTRODUCTION

Over the last decade, deep learning systems have enjoyed tremendous success in the area of computer vision and natural language processing (NLP). From image recognition, object detection, to machine translation, deep learning has completely changed the state of Artificial Intelligence (AI). However, in recent years, it has become apparent that there are several limitations in the current deep learning models [1, 2]. In particular, they struggle in tasks with a compositional and structured nature which require more deliberate thinking and multi-step reasoning [3, 4]. Many argued that current deep neural networks are just a large correlation engine, which only captures statistical patterns between input and output to make prediction rather than the true underlying reasoning processes.

To improve the capabilities of the current deep learning systems, many datasets and benchmarks have been developed. One such task which has become very popular is Visual Question Answering (VQA). It is one of the main testbeds for pushing state of the art in vision and language research. Specifically, VQA is the task of answering question about an image. It is considered to be an "AI-complete" task, as it demands fine-grained understanding of images as well as the ability to understand questions and provide answers in natural language. The task requires a rich set of abilities, such as object recognition, activity recognition, commonsense understanding and relation extraction, spanning both the visual and linguistic domains.

Another line of research which is closely related to VQA is Visual Reasoning. Visual reasoning can be considered a sub-task of VQA, with a focus on evaluating the reasoning capabilities of deep learning models. Several datasets such as CLEVR [5] and GQA [6] have emerged to specifically test various reasoning skills, including spatial understanding and higher-level skills such as counting, performing logical inference and making comparisons. The questions in these datasets tend to be highly compositional and requires multi-step inferences. This is in contrast with the typical VQA datasets where questions can be quite straightforward and may contain real-world bias.

2 PROBLEM

The aim of this project is to study several VQA models utilising Transformer's self-attention mechanism [7] and study how to adapt the Transformer architecture for a multimodal task testing on new datasets, such as visual reasoning. This is inspired by the recent success of deep learning models based on Transformer architecture, such as BERT [8] and GPT [9]. These models have shown astonishing results in various NLP tasks, such as question answering, machine translation and text generation, and seem to be capable of reasoning to a certain extent. In addition, they have been shown to learn the underlying linguistic structures in natural language [10]. The hypothesis of using a Transformer architecture in a visual reasoning task is that by stacking multiple Transformer's self-attention layers, the model would learn to focus on different parts of the image, identify most important words in the question, and then come up with an answer.

The initial objective of this project is to study various types of deep learning models, and how they are applied in VQA and visual reasoning task. Afterwards, the next objective is to explore recent approaches which extend the reasoning capabilities of artificial neural networks, and investigate how they can be effective on a visual reasoning task. In particular, this project will focus on examining attention mechanism and its various types on own customized dataset, including visual attention and self-attention. Additionally, several deep learning models based on Transformer architecture will be analysed to understand its inner working.

3 RELATED WORK

There are a variety of models proposed for VQA and visual reasoning based on alternative approaches. These models build on other recent advances in deep learning, such as the use of external memory (Memory Augmented Neural Networks) [7, 8] and graph structure (Graph Neural Networks) [10]. For example, the model proposed in [12] employs external memory to accurately predict answer which rarely occurs in the training set. Another recent model called Neural State Machine [11] decomposed image into probabilistic graph that captures the semantic relation of the visual scene, then sequential reasoning process is performed over the graph structure to derive the answer. Another common approach for solving visual reasoning is based on modular neural network [3,4,5]. This type of network composes of a collection of pre-defined neural modules, where each module is responsible for an elementary reasoning operation, such as identifying object's colour and counting. In modular neural network, the questions are first translated into an action plan, then the network is constructed dynamically based on the plan using neural modules to obtain the answer. Citing Related Work
This section cites a variety of journal [5, 15], conference [1, 6, 8, 12, 13], and magazine [3] articles to illustrate how they appear in the references section. It also cites books [9, 10], a technical report [7], a PhD dissertation [4], an online reference [14], a software artifact [11], and a dataset [2].

4 APPROACH

As shown in Figure 1.1, the architecture of our model consists of input feature extractors, a stack of multimodal Transformer layer (Nx), multimodal fusion layer, and a classifier. The main configurable parameters of our model are the hidden dimension of the model (d_{model}), the number of Transformer layer (N), the number of attention heads in the Transformer layer (A), and the hidden dimension of feed-forward networks in the Transformer layer (df f). These hyperparameters can be adjusted based on the difficulty of the dataset. The input feature extractors (CNN, LSTM) and the classifier components are similar to most existing VQA models. However, the main differences between our proposed model and existing VQA models are the use of Co-Attention Transformer layer and how they are stacked together, as well as the pooling strategy, which we utilise the [IMG] and [TXT] token as a pooled representation of the image and question. The details of each component are explained as follows.

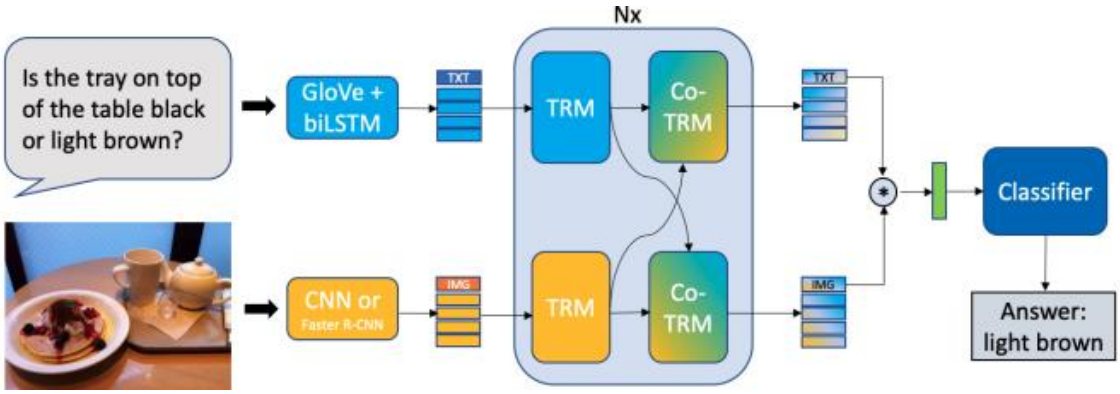


Figure 1 Proposed model architectur

4.1 Input Representation

4.1.1 Image Representation

The input image can be represented either using the region-based visual features extracted from Faster R-CNN or the grid-based feature map from pre-trained ResNet. In the case of region-based visual features, up to 100 detected objects (m regions) are used as the representation of the input image, and each region x_i has the dimension of 2048 (dx):

$$FI = FasterRCNN(Image)$$

where $FI \in \mathbb{R}^{m \times dx}$ is the feature matrix of the input image.

4.1.2 Question Representation

The question text is first tokenised into a sequence of words q = [q1, ..., qn], with each word is represented as a one hot vector from the index of the dataset's vocabulary. Each word vector is then transformed into a distributed vector representation using 300D GloVe word embedding:

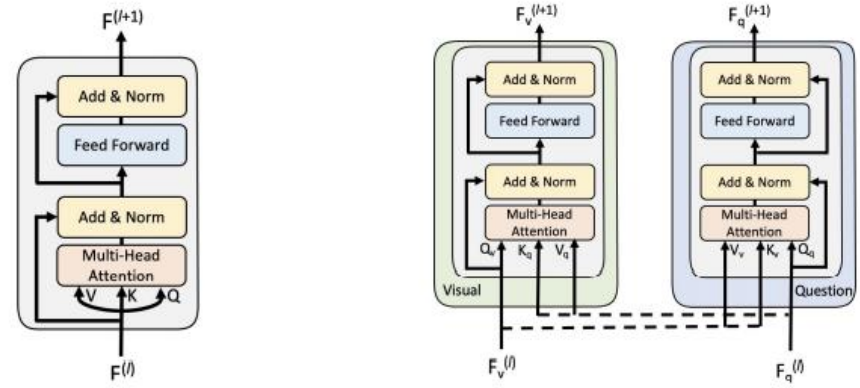
$$E = qWe$$

where $E \in R^{n \times 300}$ is the embedded question, n is the number of words in the longest question, and We is the GloVe embedding weight.

4.2 Multimodal Transformer

The multimodal Transformer encoder block consists of two components: standard Transformer layer (TRM), which is the same as the original BERT model, and Co-attention Transformer layer (Co-TRM). However, in Co-TRM layers, the keys

and values from each modality are swapped. The attention mechanism in Co-TRM layers can be interpreted as question-conditioned image attention in the textual stream and image-conditioned question attention in the visual stream. The image attention conditioned on the question is similar to visual attention found in many existing VQA models



(A) Standard Transformer layer (TRM)

(B) Co-attention Transformer layer (Co-TRM)

Figure 2 Multimodal Transformer block

4.3 Initial Design

Initially, our VQA model was designed to be based purely on self-attention mechanism from standard Transformer layer, which would be similar to single-stream pre-trained vision and language models, such as Pixel-BERT and VisualBERT. In this architecture, the features extracted from image and question are simply concatenated then fed through multiple layers of standard Transformer encoder. Similar to the original BERT model, segment embedding is introduced to distinguish between the two modalities. In this design, [CLS] token is used to extract information related to the answer, and the hidden state of [CLS] token from the final layer of Transformer is used as an input to the classifier layer. Figure 3 shows the overall architecture of our initial design.

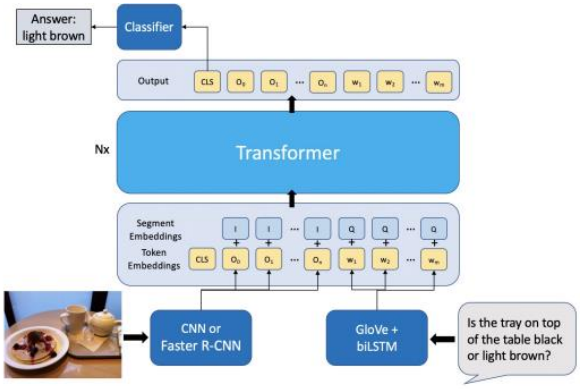


Figure 3 Initial model design

5 EVALUATION

5.1 CLEVR Dataset

We first evaluated our VQA model on the CLEVR dataset [5]. Introduced in 2017, CLEVR is a diagnostic dataset designed to test a range of visual reasoning abilities. The dataset consists of synthetic images and questions generated from functional program. It was considered to be quite challenging, as many standard deep learning based VQA models performed very poorly.

CLEVR: right
Ours: right

CLEVR: right
Ours: right

5.2 Our own dataset

Inspired on the OK-VQA dataset [6], I generated a mini VQA dataset to test a wide range of vehicles and transportations. The dataset consists of 30 real-world images downloaded from website with one question for each image.

	CLEVER-counting	Our dataset
Accuracy	95.6	83.3

Table 1 results of our model

6 DICUSSION

We believe that the model's performance on both datasets could be further improved. Due to time constraint and resources requirement to train the model, we could not perform extensive hyperparameter search. It would have been interesting to know the effects of different model sizes and depth of the Transformer layers. In addition, there are further analysis of model which could be performed. Furthermore, there is still much research on improving the stability and applicability of our own dataset.

As future work, better visual features could be incorporate to help with the performance of the model. In VQA, ResNet architecture, which is pre-trained on classification are typically used to represent the image. However, it is reported that changing the pre-training task to object detection can boost the accuracy of the VQA model. We believed that this would definitely increase the model's performance, especially on the GQA dataset, as it contains real-world images. Moreover, it would be interesting to incorporate curriculum learning as the training strategy to see whether the model can learn faster. Another interesting research direction is to apply the model on other input modalities, and other tasks which require multi-step reasoning, such as textual question answering and reading comprehension.

[1] G. Marcus, “Deep learning: A critical appraisal,” arXiv preprint arXiv:1801.00631, 2018Sam Anzaroot and Andrew McCallum. 2013. UMass Citation Field Extraction Dataset. Retrieved May 27, 2019 from http://www.iesl.cs.umass.edu/data/data-umasscitationfield

[2] F. Chollet, “The limitations of deep learning.” https://blog.keras.io/the-limitations-of-deep-learning.html, Jul 2017.Chelsea Finn. 2018. Learning to Learn with Gradients. PhD Thesis, EECS Department, University of Berkeley.

[3] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, “Building machines that learn and think like people,” Behavioral and brain sciences, vol. 40,2017.

[4] M. Garnelo, K. Arulkumaran, and M. Shanahan, “Towards deep symbolic reinforcement learning,” arXiv preprint arXiv:1609.05518, 2016.

[5] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2901–2910, 2017.

[6] D. A. Hudson and C. D. Manning, “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, 2019

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, pp. 5998–6008, 2017.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., “Language models are few-shot learners,” arXiv preprint arXiv:2005.14165, 2020.

[10] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, “What does bert look at? an analysis of bert’s attention,” arXiv preprint arXiv:1906.04341, 2019.

[11] R. Yampolskiy, “Ai-complete, ai-hard, or ai-easy: Classification of problems in artificial intelligence,” The 23rd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, OH, USA, 01 2012.

[12] R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[13] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, “Vqa: Visual question answering,” in Proceedings of the IEEE international conference on computer vision, pp. 2425–2433, 2015.

[14] S. Manmadhan and B. C. Kovoor, “Visual question answering: a state-of-the-art review,” Artificial Intelligence Review, pp. 1–41, 2020.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[16] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[17] S. Saha, “A comprehensive guide to convolutional neural networks — the eli5 way.” https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neuralnetworks-the-eli5-way-3bd2b1164a53, 2018.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, Ieee, 2009.