Enhancing Visual Question Answering through Advanced Image Encoders and Attention Mechanisms

Yiting Mao
BNU-HKBU United International College
Zhuhai, China
r130026105@mail.uic.edu.cn

Meng Xu
BNU-HKBU United International College
Zhuhai, China
r130026169@mail.uic.edu.cn

Haitong Chen
BNU-HKBU United International College
Zhuhai, China
r130026006@mail.uic.edu.cn

Ruowen Jing
BNU-HKBU United International College
Zhuhai, China
r130026066@mail.uic.edu.cn

Abstract

In this study, we focused on improving multiple-choice questions for visual question answering (VQA). By replacing VGGNet with the advanced ResNet to improve the capture of abstract features, and using Attention mechanisms to make the model more flexible in processing image information, we significantly improved the performance of VQA models in abstract scenarios. We also tried BERT to replace LSTM in the Natural Language understanding part. This highlights the effectiveness of selecting advanced image encoders and introducing attention mechanisms.

1. Introduction

Visual Question Answering (VQA) is an artificial intelligence (AI) task designed to enable a computer to understand and answer questions about the content of an image. In VQA, the system needs to understand both the image and the content of the question, and then generate or select the correct answer.

In our project, we focus on multiple-choice question answers for abstract scenes. In the original paper[1], their approach was to use VGGNet and LSTM to encode images and semantic problems, respectively. Since VQA requires image understanding and natural language understanding, we will improve the model in these two directions by replacing VGG with ResNet and LSTM with BERT. Finally, we consider adding Attention to the image coding process to observe whether it can improve the model's attention to different parts of the image so that the model can improve its performance.

2. Related Work

Visual Question Answering. Visual Question Answering (VQA) has made significant progress in recent years, driven by advances in deep learning techniques. Early approaches to visual question answering [5] relied heavily on traditional computer vision and natural language processing techniques, laying the foundation for subsequent research in the field. For example, in [5], a multi-world approach was proposed to combine semantic segmentation, symbolic reasoning and Bayesian methods to automatically answer questions about real-world scenarios. The advent of deep learning has brought about a paradigm shift in the field of VQA, utilizing Convolutional Neural Networks (CNNs) to extract image features and Recurrent Neural Networks (RNNs) to process textual information. However, this paper is based on the study of real-world scenarios and lacks in-depth consideration of abstract scenarios.

Visual Abstraction. With the depth of research, researchers are gradually realizing the importance of solving problems related to abstract scenes for advancing visual question answering (VQA). Abstract scenes involve the understanding of conceptual and non-concrete scenes, which is important for expanding the application scope of VQA models and improving their generalization ability. Therefore, the study of VQA in abstract scenarios (toy scenarios) was born. [2] created a new dataset for abstract scenes and introduced a multimodal fusion model that effectively combines image and text information. Since the model proposed in [2] is based on a traditional deep learning architecture, we improve the model and introduce an attention mechanism to improve the performance of the VQA model in abstract scenes.

Multimodal Fusion Model. In image feature ex-

traction, we propose the use of the ResNet-18 model as a replacement for the traditional VGG-19 model. For text feature extraction in question processing, we leverage BERT's contextual embeddings instead of LSTM model to enhance the semantic understanding of textual content.

Visual Attention. Recently, numerous research efforts in Visual Question Answering (VQA) have introduced attention models [6], [7], [8], [3]. [6] introduces the principle of self-attention, which in VQA improves the model's ability to model complex relationships between images and questions while maintaining a global focus on the input sequence, thus making the model more flexible and adaptable to different input contexts. However, VQA needs to handle more complex tasks or perform deeper reasoning, as we propose to employ Stacked Attention Networks through self-attention to improve model performance.

3. Methodology

3.1. Dataset

This essay uses a dataset provided by *visualqa.org* [1], which includes images, related questions, and answers associated with the questions. Each sample contains an image, a natural language question, and a manually labeled answer corresponding to that question. The images cover multiple scenes and topics, thus providing a diversity of visual content.

Table 1: Descrption of Dataset

Type	***Name***	***Size***
Annotations	Training annotations 2015 v1.0	600,000 answers
	Validation annotations 2015 v1.0	300,000 answers
Questions	Training questions 2015 v1.0	60,000 questions
	Validation questions 2015 v1.0	30,000 questions
	Testing questions 2015 v1.0	60,000 questions
Images	Training images	20,000 images
	Validation images	10,000 images
	Testing images	20,000 images

This essay chose to use the dataset provided by *visualqa.org* because it covers a rich set of abstract scenes and allows us to explore the complex relationships between images and text. This dataset has been widely used in visual quizzing tasks to evaluate the performance of various deep-learning models.

3.2. Data Processing

To make the original visual quiz dataset suitable for deep learning model training and evaluation , this essay Preparation the data through the following four steps, Constructing the vocab, Pre-processing of data sets, Building the data loader and Image resizing, to provide more reliable and usable inputs to the model.

Constructing the vocab

This part aims to generate a vocabulary of questions and answers and to save the vocabulary as a text file.

Implementation: First traverse multiple dataset files in the specified directory, load the questions and answers in each dataset, segment the questions and answers, and add the segmented words to the question and answer vocabularies, respectively.

Pre-processing of data sets

This section aims to construct the dataset for the visual question and answer (VQA) task.

Implementation: The annotations of the question and a set of valid answers are first read, for which answer extraction is performed and two lists including all answers and valid answers are generated.

Image resizing

This part is designed to resize the image to ensure that the input images have consistent dimensions.

Implementation: This essay uses the resize method of the Pillow Library (PIL) to batch-unify target images to 224*224 size after reading them, and finally save the result in a new directory.

Building the data loader

This part aims to implement a custom dataset class, VqaDataset, and a data loader, get_loader, for loading datasets for Visual Quizzing (VQA) tasks.

Implementation: Image paths, question text and answer markers are first extracted from the VQA data. Images are opened through the PIL library, converted to tensor and normalised. The question text is then converted to the corresponding word index and populated according to the maximum question length of the model. Answer tokens are also loaded if it is a training or validation set. A sample dictionary containing the image, question and answer information is then returned. Finally, the corresponding data loaders are constructed for the training and validation set datasets based on the above data.

3.3. Image Encoder

For reproduction, we used the VGG-19 to extract the features of images. Later, to improve its performance,

we tried ResNet-18.

VGG-19

Unlike other models, this model uses a relatively homogeneous convolutional layer structure and uses multiple 3x3 convolutional layers instead of one large convolutional kernel, as well as increasing the number of full connections. This gives VGG a simple structure and uniform design.

ResNet-18

ResNet-18 is a deep convolutional neural network structure designed to solve the problems of gradient vanishing and gradient explosion in the training of deep neural networks. ResNet-18 makes it easier to train deep neural networks by introducing the structure of Residual Block, which makes it easier for the network to learn identity mapping.

Comparison

The main difference between the two is the network structure and the size of the convolutional kernel; ResNet uses residual connectivity, which allows the network to be trained better, while VGG19 uses a deeper network structure and smaller convolutional kernel, which allows the network to learn the details in the image better. In practice, the choice of network structure depends on the specific task and dataset[4].

Table 2: Comparison of Resnet-18 and VGG-19

	VGG-19	***Resnet-18***
Gradient Vanishing	Can be mitigated or eliminated by residual learning	Unable to solve the problem because direct stacked convolution is used
Converge Speed	Resnet-18 is Faster, which means it could learn more features at the same time	
Parametric efficiency	By using residuals, the correlation between the parameters is reduced and the validity of the parameters is improved. This allows ResNet-18 to achieve better performance with fewer parameters compared to VGG.	
Depth of Learning	ResNet18 introduces the concepts of residual connections and jump connections, making the network structure more complex, but also enabling the network to learn features more deeply.	

3.4. Question Encoder

The original model converts the word sequence of the input problem into a compact fixed-size feature vector through an embedding layer. The model uses the LSTM layer and the fully connected layer to capture the semantic information of the problem and output the fixed dimension vector representing the semantics of the problem, which is suitable for natural language processing tasks. For comparison, the BERT model is used in this experiment, which is a pre-trained language model in natural language processing.

Transformer architecture BERT model is based on transformer architecture, which is a self-attentional deep neural network architecture. The self-attention mechanism allows the model to consider contextual information for all locations simultaneously while processing the input sequence, thus better capturing long-distance dependencies in the sequence. Whereas traditional language models are usually one-way models from left to right or right to left, BERT is two-way. This means that when pre-trained, the model can use both the left and right contextual information to more fully understand the context of each word.

BERT consists of multiple Transformer blocks, typically 12 or more layers, each containing multi-head self-attention mechanisms and feedforward neural networks. This gives the model the ability to capture different levels of semantic information.

Unsupervised training

BERT models also learn rich language representations through unsupervised pre-training on large-scale text data. Pre-training is mainly conducted on two tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). The MLM task is that certain words in the input text are randomly obscured, and the model needs to predict these obscured words. The NSP task receives input from two sentences for the model, predicting whether the two sentences are adjacent in the original text.

3.5. Attention

We implemented a multimodal fusion model based on a self-attention mechanism to solve a visual question answering task.

Self-Attention

Self-attention mechanisms are mainly used for feature association within a modality. In text, the self-attention mechanism allows a word to focus on other words in a sequence to better capture contextual information. In images, the self-attention mechanism allows a region in an image to focus on other regions to capture global contextual relationships.

In practice, the attention function is computed by

combining a set of queries into a matrix Q. Similarly, keys and values are combined into matrices K and V. The output matrix is computed as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

In the project, we defined the self-attention module to compute the attention weights for the images and the problem. Using these attention weights image features were weighted and summarized and the weighted and summarized features were combined with problem features.

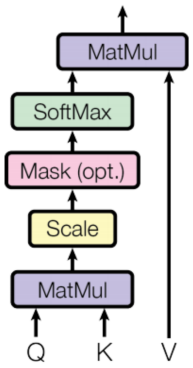

Figure 1: Scaled Dot-Product Attention[6]

External attention mechanism: Stacked Attention Networks

Based on processing more complex tasks or performing deeper reasoning, we proposed to use Stacked Attention Networks instead of self-attention.

The Stacked Attention Networks (SAN) contains multiple attention layers, each of which is a self-attentive mechanism. These layers are stacked together by allowing the model to learn relationships between features at multiple levels.

In the project, we fused image and question features layer by layer through a multi-layer self-attention mechanism. Finally, they were mapped to the answer vocabulary space through MLP.

4. Experimental Results

We started with the reproduction of the article first. To simplify, we use VGG-19 and LSTM as images and text encoders. The result of the recurrent model is our baseline. Then, we trained the models in three cases, which are **ResNet+LSTM**, **ResNet+BERT**, and **ResNet+LSTM+Attention** respectively. Each experiment ran 30 epochs, and the results are shown in the Table3.

Table 3: Results of different Experiments

Experiment	***Loss***	***Accuracy***
VGG-19+LSTM	1.88	0.6223
ResNet+LSTM	1.77	0.6521
ResNet+BERT	3.95	0.3220
ResNet+LSTM+Attention	1.70	0.6615

Based on the results, we can observe a significant performance improvement by replacing VGG with ResNet. In the course of training, the loss value decreases while the accuracy increases. This improvement can be attributed to some of ResNet's advantages over VGG. The residual connections introduced by ResNet allow for deeper networks, alleviating the problem of disappearing gradients and allowing for more efficient learning of complex feature representations. This makes the network more able to capture abstract features in images, which helps to improve the generalization performance of the model. In addition, the ResNet model is more stable when training large deep networks, avoiding some of the common training problems of deep networks. Therefore, by introducing ResNet, we can optimize the model more effectively, reduce losses, and improve accuracy, resulting in better performance on image classification tasks.

By introducing an attention mechanism in the image encoder, further performance improvements were observed. This improvement could result in the model being more focused on key areas in the image, allowing the encoder to capture task-relevant information more efficiently. The attention mechanism allows the model to dynamically focus on specific areas while processing images, helping to improve the model's perception of important features. This task-specific attention mechanism makes the image encoder more adaptable to different scenes and visual contexts, thereby improving the overall performance of the model. Therefore, by combining the attention mechanism and the image encoder, we further optimized the model to achieve better performance.

However, in the results of replacing LSTM with BERT, the performance is very poor, which is beyond our expectations. We suspect there may be several reasons. Since BERT is based on a large-scale pre-trained model, domain-specific data fine-tuning may be required to perform well on a specific task. Our raw data most likely did not match BERT's pre-training data, resulting in a significant drop in performance. At the same

time, BERT tends to require a longer fine-tuning phase because of its larger number of parameters. However, due to time and resource limitations, we may not achieve enough iterations in the fine-tuning process to obtain better results.

5. Conclusion and Discussion

In this project, we focus on solving multiple-choice question answers in visual question answering (VQA), improved by introducing advanced deep learning models. Based on the original paper [1], we adopted a series of model improvement measures. First, we replaced the VGGNet used in the original paper and chose the more advanced ResNet as the image encoder. ResNet introduces residual connections in contrast to VGG, allowing for deeper networks. This change makes our model better able to capture abstract features in images, thus improving the performance of image understanding. Secondly, we replace the LSTM used in the original paper and adopt BERT for natural language understanding. But for various reasons, BERT's performance is not very good. Finally, we introduce the attention mechanism, which gives the model the ability to focus on different parts of the image during the image coding process. The addition of this attention mechanism allows the model to process the information in the image more flexibly, further improving performance. Overall, through these improvements, our VQA model achieved significant performance improvements on multiple-choice questions in abstract scenarios. This suggests that for VQA tasks, choosing more advanced image encoders, and introducing attention mechanisms where appropriate, are both effective strategies. In the future, we will further study and explore the generalization performance of these improved strategies in other VQA scenarios, and explore whether BERT can be adapted to VQA problems.

Contribution

Table 4: Contribution

Members	***Contribution***
Yiting Mao	Coding (Except BERT and Attention) ; Thesis (Introduction, Results, Conclusion, Abstract) ; PPT & Presentation (Results) ; Merge Thesis
Meng Xu	Coding (Attention) ; Thesis (Related Work, Attention) ; PPT & Presentation (Related Work) ; Merge Thesis
Haitong Chen	Coding (BERT) ; Thesis (ResNet, Dataset) ; PPT & Presentation (Method) ; Merge PPT ; Merge Video
Ruowen Jing	Coding (BERT) ; Thesis(BERT, Data Processing) ; PPT & Presentation (Introduction) ; Merge PPT

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2, 5

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015. 1

[3] P. Gao, H. Li, H. You, Z. Jiang, P. Lu, S. C. H. Hoi, and X. Wang. Dynamic fusion with intra- and inter- modality attention flow for visual question answering. *CoRR*, abs/1812.05252, 2018. 2

[4] A. V. Ikechukwu, S. Murali, R. Deepu, and R. Shivamurthy. Resnet-50 vs vgg-19 vs training from scratch: A comparative analysis of the segmentation and classification of pneumonia from chest x-ray images. *Global Transitions Proceedings*, 2(2):375–381, 2021. 3

[5] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *CoRR*, abs/1410.0210, 2014. 1

[6] A. Vaswani, N. Shazeer, N. Parmar, J. U. L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 2, 4

[7] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274, 2015. 2

[8] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian. Deep modular co-attention networks for visual question answering. *CoRR*, abs/1906.10770, 2019. 2