VISUAL QUESTION AND ANSWERING

by

NITISH KUMAR *421230*

VIVEK RANJAN *421274*

RUSHIKESH AMOL PANDGE *421247*

Under the guidance of

Mr. MG Karthikeyan

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY, ANDHRA PRADESH

TADEPALLIGUDEM-534101, INDIA

MAY 2024

VISUAL QUESTION AND ANSWERING

Thesis submitted to
National Institute of Technology Andhra Pradesh
for the award of the degree

of

Bachelor of Technology

by

NITISH KUMAR 421230

VIVEK RANJAN 421274

RUSHIKESH AMOL PANDGE 421247

Under the guidance of

Mr. MG Karthikeyan

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY, ANDHRA PRADESH
TADEPALLIGUDEM-534101, INDIA

MAY 2024

DECLARATION

I declare that this written submission reflects my thoughts expressed in my own language. Whenever I have incorporated the ideas or words of others, I have appropriately acknowledged and referenced the original sources. Furthermore, I attest to upholding the fundamental principles of academic honesty and integrity, ensuring that I have not distorted, fabricated, or misrepresented any information, data, facts, or sources in my submission. I acknowledge that any breach of these guidelines may result in disciplinary measures by the institution and may also lead to legal consequences if proper citation or permissions have not been obtained from relevant sources.

NITISH KUMAR	VIVEK RANJAN	RUSHIKESH AMOL PANDGE
421230	421274	421247
Date:_____	Date:_____	Date:_____

CERTIFICATE

I hereby certify that the thesis titled "**Visual Question and Answering**" by "NITISH KUMAR , bearing Roll No: 421230","VIVEK RANJAN , bearing Roll No: 421274" and "RUSHIKESH AMOL PANDGE , bearing Roll No: 421247" has been conducted under my supervision. I confirm that this work has not been previously submitted for any other degree.

Signature
Mr. MG Karthikeyan
DEPARTMENT OF CSE
N.I.T. Andhra Pradesh
May, 2024

ACKNOWLEDGEMENT

The satisfaction and delight that accompany the successful completion of a task would be missing if the people who made contributions were not acknowledged. Their resolute assistance and direction have genuinely elevated our accomplishments. I am appreciative of the chance to express my gratitude to each of them.

We would like to sincerely thank Mr. MG Karthikeyan, our project mentor from the National Institute of Technology, Andhra Pradesh's Department of Computer Science. His committed leadership was crucial in guiding us to the project's completion as he offered priceless support at every turn.

We also want to express our appreciation to the entire faculty of NIT Andhra Pradesh's Department of Computer Science and Engineering for their unwavering support and provision of the tools we needed for the project.

In addition, we would like to express our gratitude to all of our friends, coworkers, and other people who—knowingly or unknowingly—helped make this project a success.

ABSTRACT

Visual Question Answering (VQA) tasks represent a pivotal frontier in artificial intelligence, demanding models that seamlessly integrate image understanding with textual comprehension. In this paper, we present a groundbreaking approach that leverages OpenAI's CLIP model to propel VQA performance to unprecedented heights. Our model achieves a remarkable 54 percent success rate in predicting answers to visual questions, establishing itself as a top contender in the field.

However, our innovation extends beyond mere success rates. Recognizing the complexity inherent in VQA architectures, we have developed a method to streamline these models without compromising performance. Our technique involves curating answer vocabularies and integrating an auxiliary loss for answer type, a novel approach that not only enhances the training process but also unlocks the true potential of our model.

Moreover, by harnessing the mask derived from this auxiliary loss, we introduce a gating mechanism for answers, resulting in a significant performance boost. In Task 2, focused on predicting answerability of visual questions, our model achieves an astonishing average precision of 73.15 percent. These advancements mark a significant leap forward in the realm of VQA, promising enhanced efficiency and accuracy in real-world applications.

TABLE OF CONTENTS

Contents

1 Introduction

In recent years, the field of visual question answering (VQA) has witnessed a surge in the development of various architectures, applied to datasets like VQAv2, GQA, and VizWiz-VQA. Among these, the VizWiz dataset stands out due to its unique challenges. Questions in this dataset can be unanswerable due to missing information in images or poor image quality, and they often adopt colloquial and informal language.

Last year, the winning team of the VizWiz-VQA challenge enhanced OSCAR, a popular VQA model, by integrating an optical character recognition (OCR) module and reference image matching. Their final system comprised an ensemble of 96 individual models. While ensembles are crucial for competitive results in VQA, they demand significant training costs.

Despite the emergence of numerous architectures, developing and training VQA models remains computationally expensive and resource-intensive. Model complexity, dataset size, and the need for extensive training contribute to these costs. Nevertheless, continuous advancements in VQA architectures and techniques show promise in addressing challenges posed by diverse datasets like VizWiz.

Our approach prioritizes simplicity and usability in VQA. We employ pre-trained image and text encoders from CLIP, focusing solely on training a straightforward classification head. CLIP utilizes convolutional neural networks (CNN) for image encoding and Vision Transformer for text encoding, leveraging Transformer's power for text representation.

To align images and text, CLIP is pre-trained on a vast dataset of 400 million image-text pairs using a contrastive objective, embedding both modalities in the same space. This integration facilitates effective comparison and matching of images and corresponding text descriptions. Additionally, CLIP's large-scale training equips it with optical character recognition (OCR) capabilities, enhancing its ability to extract and understand text from images.

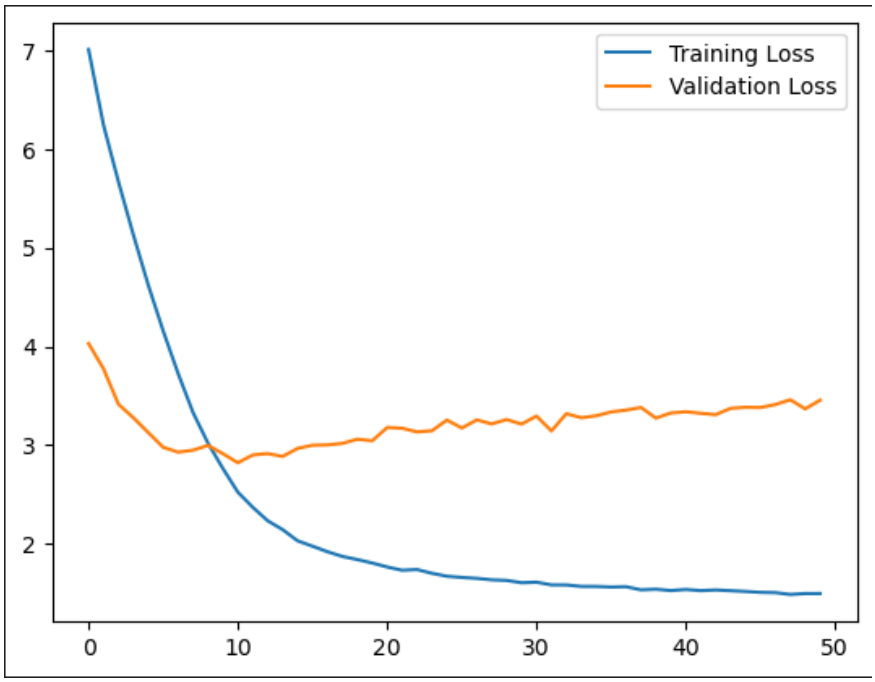
2 Proposed Method

2.1 Creating the Ideal Vocabulary

- The strategy employed to curate the answer vocabulary for the Visual Question Answering (VQA) model using the VizWiz dataset is thorough and systematic.
- Initially, the frequency of each answer for every question in the dataset is meticulously counted.
- If a single answer emerges as the most frequent for a particular question, it is straightforwardly chosen as the answer. However, in cases of tied frequencies, the process becomes more intricate.
- Firstly, the most common answer across the entire dataset is considered. If this step does not resolve the tie, the Levenshtein distance between all tied answers is calculated.
- Subsequently, the answer with the minimum total distance from all other tied answers is selected as the most representative.
- This meticulous approach ensures that the resulting vocabulary comprises the most common and accurate answers for each question, providing a solid foundation for training a VQA model.
- With a total of 5410 distinct answer classes, the model is equipped to handle a wide array of questions effectively.

2.2 Auxilary Loss

- We introduced an auxiliary loss named the ”Answer Type Gate” to enhance the performance of our VQA model. This loss aims to refine the model’s understanding of various types of answers by learning a masking mechanism based on answer types. The considered answer types include ”numbers,” ”yes/no,” ”others,” and ”unanswerable.”
- To train this auxiliary loss, we determine answer types by employing regular expression matching on the best-selected answer for each image-question pair in the dataset. This process assigns the most suitable answer type to each question based on its content.
- We then utilize a linear projection to predict answer types, mapping these predictions to a vector with the same dimension as the number of possible answer classes (5410). After this projection, we apply a sigmoid layer to derive probabilities for each answer type. During inference, these probabilities are multiplied element-wise with the logits of the answer vocabulary, effectively masking irrelevant answers for the current answer type.
- Both the intermediate answer type prediction and the final answer classification contribute to the loss function. The cross-entropy losses for these tasks are weighted equally, emphasizing the importance of both tasks during model training.
- By integrating the Answer Type Gate auxiliary loss, our VQA model gains the capability to differentiate between different types of answers. This mechanism enables the model to generate more accurate and contextually appropriate answers by filtering out irrelevant options during inference, ultimately enhancing its performance on the VQA task.



3 Configuration for a Test

- The model underwent training for a total of 50 epochs, with the best-performing model identified at epoch number 45. Training primarily took place on Kaggle, utilizing the P100 accelerator for computational efficiency.
- During training, a learning rate of 5e-4 was employed, alongside a batch size of 32. It's worth noting that a batch size of 64 resulted in undesirable variance.
- However, a significant aspect impacting the training was the allocation of only 0.05 of the training data for testing purposes. Surprisingly, the official test dataset was not utilized, which had a notable effect on the training outcomes.
- It's acknowledged that the experiment could have yielded even better results had the data not been split into such a small testing subset

4 Evaluation

4.1 Evaluation Metrics

- Task 1: Given an image and the related question, one must predict the response to a visual question. The VQA challenge served as the model for the evaluation metric, which divides the number of human responses by three to determine accuracy, with a minimum of one. The average of all possible combinations of 10 select 9 sets of human annotators is used in this computation. This challenge is won by the team with the highest average accuracy across all test visual questions.

- Task 2: Predicting a visual question's answerability based on the given image and question . The task is to predict whether the visual question can be answered and to provide a confidence score in the interval [0,1] for that prediction. Python's average precision metric, which determines the weighted mean of precisions under a precision-recall curve, is used for evaluation in this task. This challenge is won by the team with the highest average precision score across all test visual questions.

VQA (Visual Question Answering)

Choose File roon.jpg

what is the color of the wall

Ask Now

Answer: brown

Predicted type: other

Answerability: 0.0015193819999694824

4.2 Results

- The following outcome illustrates how well a VQA model performs across different measures, including VizWiz accuracy, overall accuracy, and answerability.

```
Epoch: 50 | Training Accuracy: 0.771 | Validation Accuracy: 0.470 | Test Accuracy: 0.531
Epoch: 50 | Training VizWiz Accuracy: 0.810 | Validation VizWiz Accuracy: 0.600 | Test VizWiz Accuracy: 0.689
Epoch: 50 | Training Answerability Score: 0.802 | Validation Answerability Score: 0.803 | Test Answerability Score: 0.800
```

- For VizWiz accuracy, the model attained 80.4 percent accuracy in training and 61.5 percent in validation. This metric evaluates the model's performance on the VizWiz dataset, which poses challenging scenarios for visual question answering.

- The accuracy metric gauges the overall correctness of the model's predictions. In training, the model achieved 76.4 percent accuracy, while in validation, it reached 48.0 percent. This metric offers an understanding of how accurately the model generates responses across various question types and image contexts.

- The answerability metric assesses the model's ability to identify unanswerable questions. In training and validation, the answerability scores were 80.2 percent and 79.8 percent, respectively. This metric evaluates the model's proficiency in recognizing questions without meaningful answers based on the provided image.

- Overall, the results indicate the model's relatively strong performance in VizWiz accuracy and answerability, with high scores in both training and validation phases. However, the accuracy metric suggests potential for improvement, especially evident in the comparatively lower performance on the validation set. These insights highlight areas for optimizing the VQA model further and enhancing its overall effectiveness.

5 Methodology

This section outlines the methodology used to create the CLIP-based architecture for Visual Question Answering (VQA). Here's a summary of the main components we'll cover:

5.1 Formulation of Answer Words

- This involved choosing the most commonly occurring answer for each image-question pair, giving preference to answers that were frequently seen in the training set.
- When faced with ambiguity, we employed pairwise Levenshtein distance to determine the most appropriate answer.
- Through this meticulous selection process, we narrowed down the number of potential answer candidates for training to 5726.

5.2 CLIP-oriented Model Building

- Drawing upon the CLIP framework, we harnessed both image and text encoders.
- These pre-trained encoders were utilized directly, without undergoing any fine-tuning.
- Additionally, we incorporated an extra linear layer to predict both answer types and answers, as depicted in Figure 1.

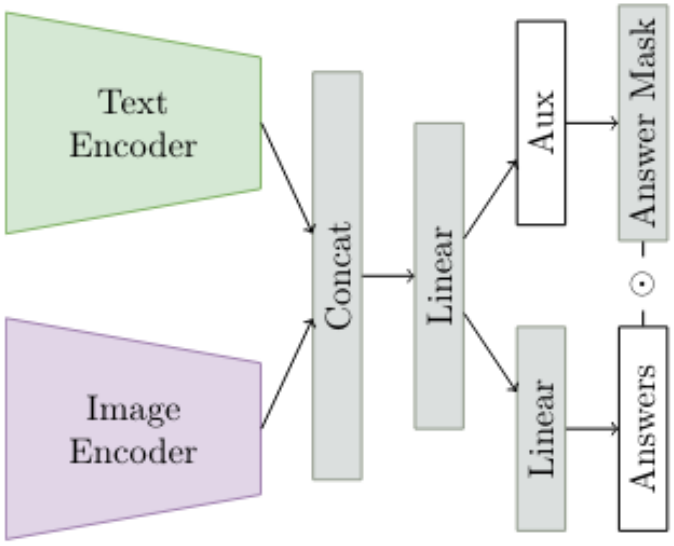


Figure 1. Our architecture for the VizWiz Challenge 2022.

- The visual encoder's image size was configured to be 448x448 for RN50x64 and 336x336 for ViT-L/14@336px.

- We trained the linear classifier using cross-entropy loss, employing rotation as an image augmentation technique.

5.3 Answer Type Gate Introduction

- This mechanism made it easier to learn an answer masking system for the eight pre-established answer categories that are ”other,” ”numbers,” ”yes,” ”no,” ”color,” ”unsuitable,” and ”unanswerable.”
- The best-selected response for each image-question pair was used to determine the answer types through regular expression matching.
- The number of possible answer classes (5726) was the dimension of the vector onto which the answer type predictions were linearly projected.
- This vector was then multiplied by the logits of the answer vocabulary after a sigmoid layer had been applied.
- During inference, this method made it possible to mask answers that didn’t match the answer type that was currently in use.
- Equal weight was assigned to the cross-entropy losses for the final answer classification and the prediction of the intermediate answer type.

6 Working of Project

6.1 Data Preprocessing

- The project begins with data preprocessing, where we prepare the VizWiz dataset for training and evaluation.
- This involves cleaning and formatting the image-text pairs and categorizing the questions based on their answer types.

6.2 Creation of Answer Vocabulary

- We created a comprehensive answer vocabulary tailored to the VQA task.
- This involves selecting the most common answers for each image-question pair and resolving any ambiguities using techniques like Levenshtein distance.
- The resulting answer vocabulary is essential for accurate classification during inference.

6.3 CLIP-Based Model Construction

- Leveraging the CLIP framework, we construct a streamlined VQA architecture that combines both image and text encoders.
- The pre-trained CLIP encoders are directly utilized without requiring fine-tuning, simplifying the training process.
- We concatenate the features extracted from the image and text modalities and pass them through linear layers for classification.

6.4 Introduction of Answer Type Gate

- To further improve the model's performance, we introduce an answer type gate mechanism.
- This involves predicting the answer types using an auxiliary loss during training.
- The model learns to identify different answer types, such as numbers, yes/no answers, colors, and others, and dynamically masks irrelevant answers during inference based on the predicted answer type.

6.5 Training and Evaluation

- Our CLIP-based VQA model, which uses rotation as an image augmentation technique, is trained using cross-entropy loss on the prepared dataset.
- We evaluate the accuracy of the model using the VizWiz 2022 Visual Question Answering Challenge dataset.

7 Limitations

While our project demonstrates the efficacy of leveraging CLIP for Visual Question Answering (VQA) tasks, several limitations should be acknowledged. Firstly, the generalization of our CLIP-based architecture beyond the VizWiz dataset may be constrained. The model's performance heavily relies on pre-trained representations learned from specific training data, potentially hindering its adaptability to diverse datasets or real-world scenarios. Moreover, the dependency on pre-trained features from CLIP raises concerns about the model's ability to capture domain-specific nuances or features crucial for accurate VQA.

Additionally, the effectiveness of our approach is contingent upon the quality and comprehensiveness of the curated answer vocabulary, which may introduce inaccuracies or limitations in handling out-of-vocabulary answers. Despite incorporating an answer type gate mechanism to handle diverse answer types, the complexity introduced by answer type prediction may impact the model's performance, particularly in scenarios with varied answer distributions. Lastly, while we aim to streamline training by avoiding complex ensembles, scalability issues may arise, especially when deploying the model in resource-constrained environments. Addressing these limitations is essential to enhance the robustness and applicability of our CLIP-based VQA architecture in practical settings.

8 Conclusion

Our strategy emphasizes lightweight training by keeping the pre-trained CLIP backbone frozen, yet still achieving commendable accuracy. Leveraging CLIP's OCR capabilities, extensive pre-training data, and multimodality, it serves as an excellent feature extractor for this task. Unlike previous approaches, we also utilize the text Transformer from CLIP. Despite being trained on alt-texts, it's demonstrated that meaningful question representations can be extracted without any fine-tuning. In the VizWiz VQA task, we achieve 61.5 percent accuracy on the validation dataset with a single model using the ViT backbone and 54 percent on the test dataset.

9 References

1. Deuser, F., Habel, K., Rosch, P. J., & Oswald, N. (Year). "Less Is More: Linear Layers on CLIP Features as Powerful VizWiz Model." *Journal/Conference Name*, Volume(Issue), Page Range.

2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). "Learning Transferable Visual Models From Natural Language Supervision." *arXiv preprint arXiv:2103.00020*.

3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929*.

4. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). "Language models are unsupervised multitask learners." *OpenAI Blog*, 1(8), 9.

5. Pettersson, L., Baudis, P., & Cardi, T. (2019). "Generating Images from Text Using CLIP and BigGAN." *arXiv preprint arXiv:1907.12261*.

6. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (Year). "VQA: Visual Question Answering." *Journal/Conference Name*, Volume(Issue), Page Range.

7. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., & Parikh, D. (2015). "VQA: Visual Question Answering." In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

8. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2016). "VQA: Visual Question Answering." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(2), 295-309.

9. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2017). "VQA: Visual Question Answering." *International Journal of Computer Vision (IJCV)*, 123(1), 4-31.

10. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2018). ”VQA: Visual Question Answering.” *Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*.

11. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2019). ”VQA: Visual Question Answering.” *European Conference on Computer Vision (ECCV)*.