Be My Eyes: Visual Question Answering with CLIP

Xinyuan Liu Xuan Yu Ziyang Wang
Department of Computer Science, Rice University
xl121@rice.edu xy50@rice.edu zw76@rice.edu

Abstract

This project proposes the development of an AI-based application, called Be my eyes, which aims to assist blind individuals in answering questions about images using the CLIP-Linear VQA Model. This model is based on the Contrastive Language-Image Pre-training (CLIP) framework and incorporates a linear model to predict the answer. The proposed approach can accurately predict answers to natural language inquiries about images when asked for detailed information. The project uses the VizWiz-VQA dataset to train and validate the model. The performance of the model is measured using top-1 and top-3 accuracy. Our method reaches comparable performance as the top challengers did in the VizWiz-VQA challenge while using much less computational power. Our approach can easily be deployed on portable devices and has the potential to improve accessibility and independence for visually impaired individuals.

1. Introduction

The ability to perceive and understand visual information is crucial in many aspects of daily life, from identifying objects and navigating the environment to enjoying art and entertainment. However, this ability is greatly limited or absent for visually impaired individuals, presenting significant challenges and barriers to independent living. In recent years, advances in artificial intelligence and computer vision have provided new opportunities for addressing these challenges and improving accessibility. In particular, the development of visual question answering (VQA) systems that can answer natural language inquiries about images has shown promise in providing blind individuals with a way to access and interpret visual information.

In this project, we propose a CLIP-Linear model based on the Contrastive Language–Image Pre-training (CLIP) model to connect images and questions to assist blind individuals to answer questions about images[1]. We use the VizWiz-VQA dataset [2, 3] to train and validate our model (see sample images in Figure 1). We fully utilize the ability of the pre-trained CLIP model. Our method reaches comparable performance as the top challengers did in the VizWiz-VQA challenge while using much less computational power. Our approach can easily be deployed on portable devices and further be developed into an AI-based application to help visually impaired individuals, *Be my eyes*. Given an image, our approach can reliably predict answers when asked about its detailed information, which has the potential to improve accessibility and independence for visually impaired individuals.

Figure 1. Sample images, questions, and answers pair from the VizWiz-VQA Dataset that we use in our work. Our objective is to answer the question about a given input. We select the most confident answer as the ground truth.

2. Related Work

We review the recent works in the field and select proper approaches to efficiently use the available resources.

Recent approaches to VQA have focused on using deep learning models to extract features from images and questions and then combining these features to generate answers. One popular approach is to use convolutional neural networks (CNNs) to extract image features and recurrent neural networks (RNNs) to process question features. However, these models can be computationally expensive and require significant amounts of training data.

[1]The source code is on https://github.com/tigerwang3133/CMPSC646

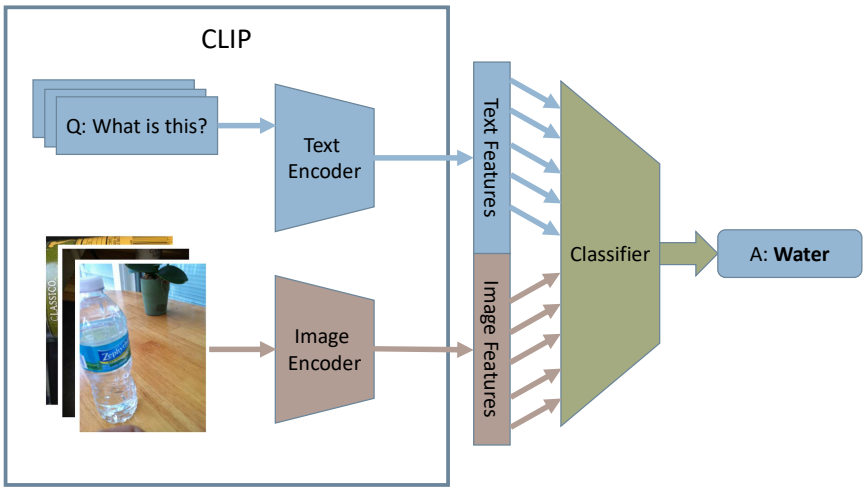

Figure 2. The architecture of **CLIP+Linear**. CLIP encodes images and questions into 512+512 features. A classifier inputs the features and outputs the label and answer.

Early work on visual and language models include [4, 5] had general approaches in connecting visual and language data. More specific work implemented CLIP on vision and language tasks [7]. Anderson *et al.* also performed VQA on robot navigation tasks [1]. However, their works have not been implemented on the VizWiz-VQA dataset. Our work will implement these state-of-art visual and language approaches on the VizWiz-VQA dataset to assist blind individuals in answering questions about images.

Therefore, *Be my eyes* builds on previous research in the field of VQA and employs the CLIP model to perform VQA.

3. Model

Our proposed approach for VQA is based on the CLIP model [6]. Specifically, we applied the classic '*clip-vit-base-patch32*' model, which is a transformer-based model that combines object detection and visual feature extraction with language modeling. Figure 2 depicts our approach's three major components: a vision module that encodes the image into visual features, a text module that encodes the question into textual features, and a fusion module that combines the visual and textual features to predict the answer. Both the text encoder and image encoder produce 512 features. We then concatenate the image features and text features and inputs for the next steps.

To predict the answer, we employed a simple n-way classification method to output the answer, named **CLIP+Linear**. In this case, we use a simple learnable two-layer multilayer perceptron (MLP) that receives image and text embeddings by CLIP as input and learns to predict answers using classification. For the MLP layers, the input dimension is 1024 (image+text features), the hidden layer dimension is 512, and the output dimension is the number of classes in answers.

To compare, we also tested with **Zero-shot CLIP**. We also trained another LSTM network to output the answer from image and text embeddings (**CLIP+LSTM**).

4. Experiments and Results

We use the VizWiz-VQA dataset to assess the performance of our model. This dataset was created with the goal of improving accessibility for visually impaired individuals through the development of image captioning and visual question-answering systems. There are 20,523 training image/question pairs and 4,319 validation image/question pairs in the dataset. Each question has 10 possible answers. We only use answerable questions when preprocessing data. We choose only one answer for each question that is confident and appears the most times out of the ten possible answers. Most questions can be answered with a single word. As a result, in order to simplify the question, we chose answers that contained only one word. We have 12,732 training image/question pairs and 2,419 validation image/question pairs after preprocessing. The testing set is not available to us. As a result, we report the validation accuracy.

We input the image and question separately using the CLIP pre-trained '*clip-vit-base-patch32*', which maps the image to a vector in a high-dimensional feature space and maps the question to another vector in a high-dimensional feature. Specifically, the image is read by Pillow and passed into the CLIP model image preprocessor. The questions are passed into the CLIP model text preprocessor. Then, we use *clipmodel.get_image_features* to extract the image features. we use *clipmodel.get_text_features* to extract the image features. Both the image encoder and text encoder produce 512 features. We then concatenate the two features and pass

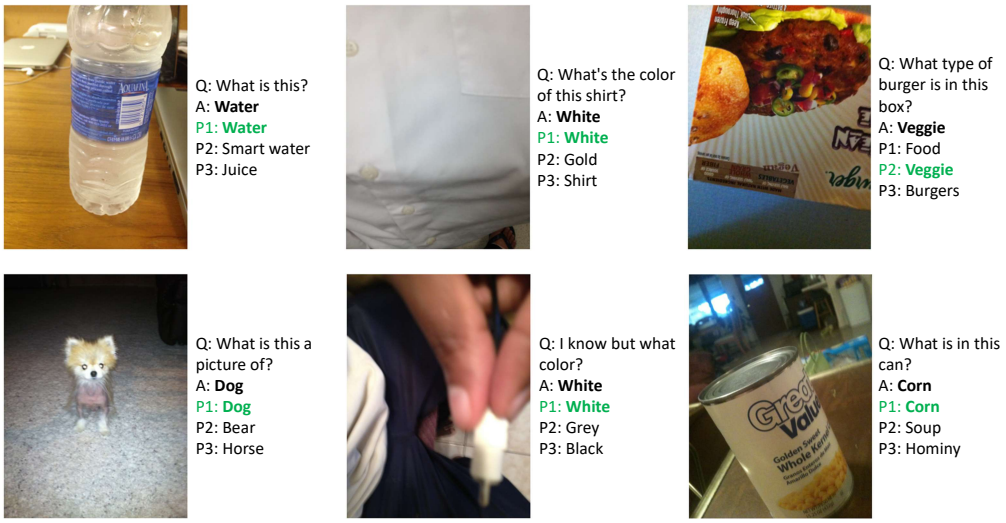

Figure 3. Samples of image, question, answer, and **CLIP+Linear** predicted top-3 answers from the validation set. Q is the question. A is the ground truth. The top-3 predictions are P1, P2, and P3. The correct prediction is colored green.

Method	Top-1 Acc	Top-3 Acc
CLIP+Linear	31.3%	56.6%
CLIP+LSTM	33.15%	47.95%
Zero-shot CLIP	0.00%	0.04%
VILT	25.2%	38.2%
KTLO-top1	/	57.72*%
UIO-top2	/	57.27*%
Katya-top3	/	54.76*%

* VizWiz accuracy

Table 1. Preliminary experimental results. Performance of **CLIP+Linear**, **CLIP+LSTM** and comparison to others.

them to a neural network with two fully connected layers. The model is called **CLIP+Linear** and is shown in Figure 2.

We use the cross-entropy loss to train the **CLIP+Linear** Model on preprocessed training data. We train for 40 epochs using the Adam optimizer with a learning rate of 0.001. To evaluate the model, we use preprocessed validation data and calculate the predicted answer accuracy using the top-1 and top-3 accuracy.

To compare the performance of our model, we run zero-shot experiments on CLIP models, with labels in the form of $\{question\}$ and $\{answer\}$ pairs. We take the dot product to find out the answer with the highest score. We also add a **CLIP+LSTM** version by passing the concatenated features to a simple LSTM network for answer prediction.

Table 1 shows the results of our experiment. We compared **CLIP+Linear**, **CLIP+LSTM** to several other methods. The top-1 and top-3 accuracies were used as the evaluation metrics. First, we compared the performance of **CLIP+Linear** model (our proposed model), **CLIP+LSTM** model to the zero-shot CLIP model and Vision-and-Language Transformer (VILT) model. Additionally, we also compared our model with the top-3 winner models from the 2021 VizWiz Grand Challenge, KTLO-top1, UIO-top2, and Katya-top3. For the 3 winner models, we do not have access to the models, and the accuracy is measured with VizWiz accuracy on another testing set. Therefore, it should be noticed that the performance of the winner models is only for approximate comparison.

The zero-shot CLIP model performed poorly, as expected, with a top-1 accuracy of 0.00% and a top-3 accuracy of 0.04%. However, after incorporating linear and LSTM classifiers, CLIP's performance improved significantly, with top-1 accuracy of more than 30% and top-3 accuracy of more than 45%.

Interestingly, our **CLIP+Linear** model outperformed the **CLIP+LSTM** model in top-3 accuracy while having a slightly lower top-1 accuracy. The VILT model, on the other hand, performed poorly, with a top-1 accuracy of 25.2% and a top-3 accuracy of 38.2%.

Among the winner models, KTLO-top1 and UIO-top2 had very similar accuracy of 57.72% and 57.27% respectively, which slightly outperformed our model. However, our model showed promising results and have better performance compared to Katya-top3.

Figure 3 are some samples of top-3 predictions of **CLIP+Linear**, demonstrating the reliability of our model. Figure 4 depicts some incorrect samples with reasonable error, indicating areas for improvement.

In conclusion, our experiment shows that our model performs well in predicting answers to natural language inquiries about images and that incorporating linear classifiers can significantly improve the CLIP model's performance

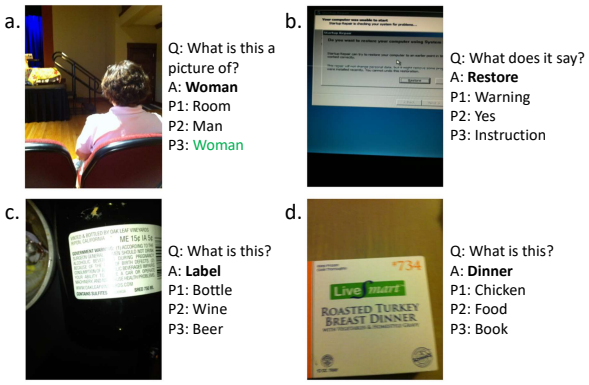

Figure 4. Wrong predictions from **CLIP+Linear**. Q is the question. A is the ground truth. The top-3 predictions are P1, P2, and P3. The correct prediction is colored green.

for simple VQA tasks. Our model is comparable to the best models in terms of accuracy.

5. Disscussion

In this section, we discuss the key findings and limitations of our study.

One of the main strengths of our VQA model is its ability to accurately answer questions about real-world images taken by blind individuals who require visual assistance. Our model achieved a high top-3 accuracy of 56.6% on the VizWiz-VQA dataset, which is a significant improvement over previous approaches that were unable to handle the complexity of the questions and the diversity of the images in this dataset.

Our idea is applicable to a real-world application assisting blind people's daily lives. Figure 5 shows an example of the interface of this *Be my eyes* app. We are sure that a simple interface and practical function like this can assist the blinds well.

However, our study also has several limitations that should be addressed in future work. One limitation is the relatively small size of the VizWiz-VQA dataset with fewer training samples for each object and question, which contains only 20,523 images and questions. For comparison, VQAv2 contains 204,721 and 1,105,904 questions. We also noticed that our approach is limited by the correctness of the annotation in the dataset and the number of classes in the answers. For example, Figure 4a shows an ambiguous question that is hard to answer. Figure 4b, c, and d show n ambiguous answers. However, our model is still able to make reasonable guesses. Additional efforts to identify and address the underlying causes of the errors (shown in Figure 4) can help achieve more accurate results, increasing the model's overall effectiveness in real-world applications. Since our VQA model achieved high accuracy on this dataset, we believe that it would generalize well to other datasets or real-world scenarios with more diverse images and questions.

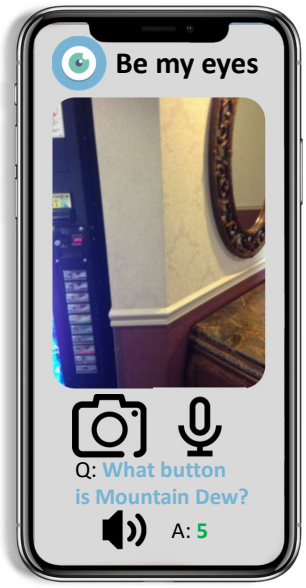

Figure 5. Be my eyes: proposed UI of the application.

References

[1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[2] D. Gurari, Q. Li, C. Lin, Y. Zhao, A. Guo, A. Stangl, and J. P. Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[3] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. *CoRR*, abs/1802.08218, 2018.

[4] Y. Hao, H. Song, L. Dong, S. Huang, Z. Chi, W. Wang, S. Ma, and F. Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022.

[5] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[7] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.