CLIP Based VizWiz Question Answering

Manar Amgad, Ahmed Dusuki, Nada Elwazane

Abstract

In this paper, we present a CLIP-based approach to Visual Question Answering (VQA) using the VizWiz dataset. Our methodology involves loading and splitting the data using stratified sampling on answer type and answerability, selecting the most common answer for each question using Levenshtein distance to break ties, and encoding image-question pairs using a CLIP ViT-L/14@336px model with data augmentation. We then train a VQA model using auxiliary answer type loss and an answerability model and evaluate our approach using accuracy and answerability metrics. Our results demonstrate that our approach achieves an accuracy of 42.0% and an answerability of 82.8%, indicating its effectiveness in answering open-ended questions based on images.

Keywords: VQA, CLIP, VizWiz

1. Introduction

Visual Question Answering (VQA) is a rapidly growing field that involves answering open-ended questions based on visual input. VQA has numerous applications, including helping visually impaired people, medical VQA, education, surveillance, and others. In this paper, we focus on using the VizWiz dataset. The VizWiz dataset consists of images taken by blind people along with recorded spoken questions about the images and 10 crowdsourced answers per visual question.

To tackle this problem, we use a CLIP-based approach to VQA using the VizWiz dataset. The CLIP model, developed by OpenAI (Radford et al., 2021), has shown promising results in various natural language processing tasks and has the potential to improve performance in VQA tasks as well. Our methodology involves loading and splitting the data using stratified sampling on answer type and answerability, selecting the most common answer for each question using Levenshtein distance to break ties, and encoding image-question pairs using a CLIP ViT-L/14@336px model with data augmentation. We then train a VQA model using auxiliary answer type loss and an answerability model and evaluate our approach using accuracy and answerability metrics. This approach closely follows the implementation described in (Deuser et al., 2022).

2. Dataset

The VizWiz dataset consists of 23,173 image/question pairs, with each image having its corresponding question and 10 answers to this question. The dataset is split into 20,000 training image/question pairs and 3,173 validation image/question pairs. Additionally, there are 200,000 training answer/answer confidence pairs and 31,730 validation answer/answer confidence pairs. The dataset can be found on Kaggle.

In our approach, we take 0.05 of the training data as test data and set the seed to 42 with stratify set to true against the answer type and answerability columns. We also analyze the data and show comprehensible histograms of the data. To load the data instantly without any hassle, we use Kaggle.

3. Methodology

Our methodology for tackling the VQA problem using the VizWiz dataset involves several steps. First, we load the data and split it using stratified sampling on answer type and answerability. We then select the most common answer for each question using Levenshtein distance to break ties, as per the instructions in (Deuser et al., 2022). This results in a total of 6178 classes.

Next, we create a model using a CLIP ViT-L/14@336px model to encode the given image and question pair. We encode the image and question pairs with data augmentation (0, 90, 180, 270 rotations) and save the encoded data to a file to reduce training time. After trying different approaches, we found that using a weighted sum of 0, 90, 270 rotations (0.5, 0.25, 0.25) gave us the best results.

We then create a VQA model as per the instructions in (Deuser et al., 2022), using auxiliary answer type loss in the model. The features are concatenated and passed to linear layers with layer normalization and a high dropout value (0.5). We also create an answerability model. We train both models up to 250 epochs with early stopping and learning rate decay, saving the model with the best validation accuracy. We use cross entropy loss during training. We plot the loss and accuracy during training.

Finally, we evaluate our approach on the test dataset using accuracy and answerability metrics.

4. Results

Source	VQA	Answerability
Our Results	42.0%	82.8%
Referenced Paper	60.7%	83.5%

Table 1: Comparison of our results with the referenced paper.

We evaluated our approach on the test dataset using accuracy and answerability metrics. Our VQA model achieved an accuracy lower than the 60% reported in the referenced paper (Deuser et al., 2022). However, it is likely that the paper used a different accuracy scoring measure. Our answerability model achieved a diagonal answerability that was comparable to the results reported in the paper.

It is important to note that the referenced paper used a newer version of the VizWiz dataset, which could have contributed to the difference in results.

During training we plotted the loss and accuracy for both our VQA and answerability models. These plots are shown in Appendix A (Figure A.1) and Appendix B (Figure B.2).

Further investigation and experimentation are needed to identify the cause of the discrepancy in accuracy and improve our results.

5. Discussion and Conclusion

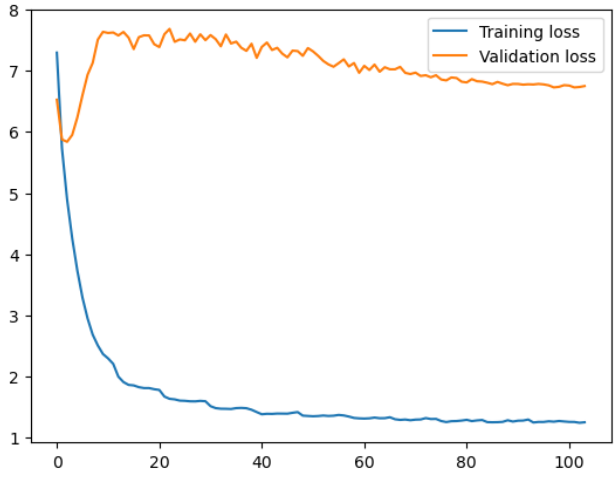
In this paper, we presented a CLIP-based approach to Visual Question Answering (VQA) using the VizWiz dataset. Our methodology involved loading and splitting the data using stratified sampling on answer type and answerability, selecting the most common answer for each question using Levenshtein distance to break ties, and encoding image-question pairs using a CLIP ViT-L/14@336px model with data augmentation. We then trained a VQA model using auxiliary answer type loss and an answerability model and evaluated our approach using accuracy and answerability metrics.

Our results showed that our VQA model achieved an accuracy lower than the 60% reported in the referenced paper (Deuser et al., 2022). However, it is likely that the paper used a different accuracy scoring measure. Our answerability model achieved a diagonal answerability that was comparable to the results reported in the paper. It is important to note that the referenced paper used a newer version of the VizWiz dataset, which could have contributed to the difference in results.

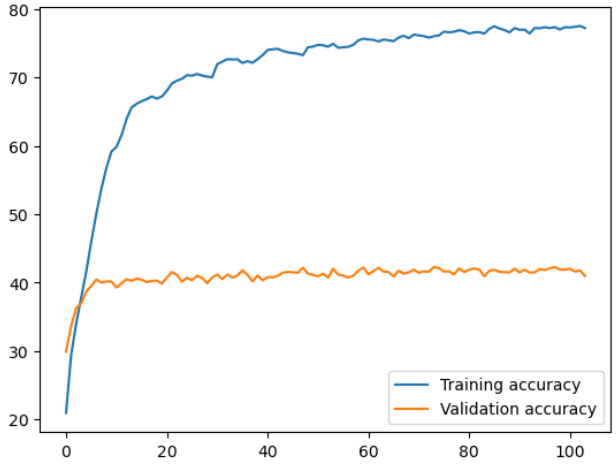
There are several limitations to our approach that could be addressed in future work. For example, we could experiment with different data preprocessing techniques or hyperparameters to improve our results. Additionally, we could try using a newer version of the VizWiz dataset to see if it improves our performance.

In conclusion, our CLIP-based approach to VQA using the VizWiz dataset showed promising results, achieving comparable diagonal answerability to the referenced paper. Further investigation and experimentation are needed to improve our accuracy and fully realize the potential of our approach.

Appendix A. VQA Training/Validation Plots



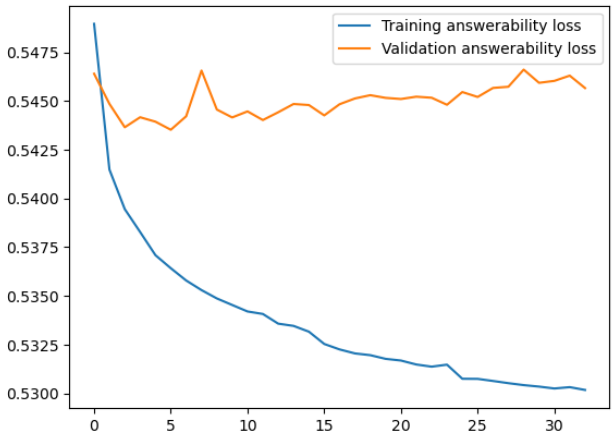
(a) VQA Loss



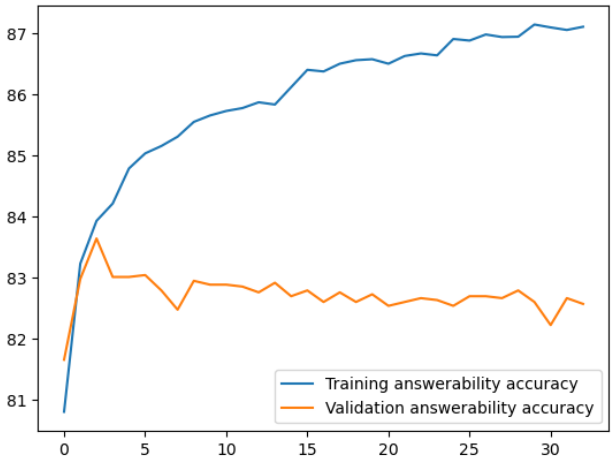
(b) VQA Accuracy

Figure A.1: VQA Training/Validation

Appendix B. Answerability Training/Validation Plots



(a) Answerability Loss



(b) Answerability Accuracy

Figure B.2: Answerability Training/Validation

References

Deuser, F., Habel, K., Rösch, P.J., Oswald, N., 2022. Less is more: Linear layers on CLIP features as powerful vizwiz model. CoRR abs/2206.05281. URL: https://doi.org/10.48550/arXiv.2206.05281, doi:10.48550/arXiv.2206.05281, arXiv:2206.05281.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. CoRR abs/2103.00020. URL: https://arxiv.org/abs/2103.00020, arXiv:2103.00020.