VizWiz Visual Question Answering

Jack Campbell, Ishan Ranjan, Rani Shah

Abstract

Visual Question Answering (VQA)[1] is a crucial emerging technology that has many real-world use cases, from assisting those who are visually impaired to providing accurate information about image features that humans might not be able to discern. To promote the development and testing of VQA models, VizWiz has formulated a set of challenges to develop models to perform VQA tasks[3]. We have implemented a multimodal transformer model that accurately predicts answers to questions about a given set of images. Our model provides a high accuracy of 0.751 as well as an answerability score of 0.798.

Introduction

VizWiz has challenges to build algorithms that assist people who are blind or visually impaired. Visual question answering (VQA) is a computer vision problem where the goal is to teach machines to understand the content of an image and answer questions about it. In the VizWiz challenge, the VQA challenge is specifically designed to answer questions that blind people may have about images. The challenge also asks how answerable a question is, which is an important metric in determining whether a given image is suitable for a VQA task and is not blurry or otherwise obscured. This kind of challenge helps educate people about blindness and creates assistive technology to provide solutions for them.

Figure 1.

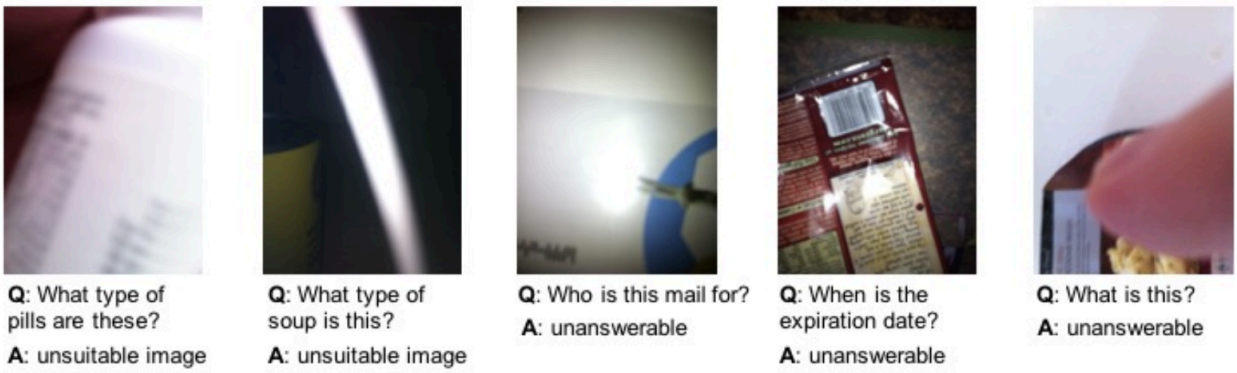

Figure 2.

Related Work

Previous research on Visual Question Answering (VQA) using different datasets has shown that certain transformer models are effective in generating accurate predictions. Meanwhile, some researchers currently participating in the VizWiz challenge have not yet published their findings in this area. The article titled “Less is More: Linear Layer on CLIP Features of Visual Question Answering” by Deuser et al.[2] builds on existing research in the field of VQA. It explores an efficient approach by applying linear layers on top of features extracted from pre-trained CLIP (Contrastive Language–Image Pre-training) models. Moreover, it incorporates a text encoder in addition to the classical use of an image encoder for VQA tasks. This implementation is much simpler than the use of complex transformer models that have been proven effective in prior literature. This approach was shown to be effective, where Deuser et al. were able to achieve an 60.15% accuracy and an 83.78% answerability with their implementation.

Methodology

To create a model that could accurately answer questions about a given image, and provide an accurate score that gauges the answerability of a question about the given image, we utilized CLIP (Contrastive Language-Image Pre-Training)[5], a ViT-like multimodal transformer. Following Deuser et al.’s “Less is More” framework, we adhered to the key insight that there is no need to re-train the entire CLIP model but instead to utilize the pre-trained image and text encoders and only add and train additional linear layers with dropout rates of 0.5 and 0.5, respectively. Since this approach does not train CLIP on the entire VizWiz dataset, it allows for faster and more efficient training while requiring less computational resources.

In addition, a suitable answer vocabulary was curated so that accuracy could be maximized. Following Deuser et al.’s approach, we selected the answer that “returned the highest score per image-question pair”. If multiple candidates were selected, the answer that appeared most often in the training dataset was used, and in the case of a tie, pairwise Levenshtein distance was utilized to “find the most representative answer to all others”. This vocabulary is then multiplied by the sigmoid result of predictions from an auxiliary loss function that helps improve answer type prediction and fed back in with the linear layers to help improve answer generation.

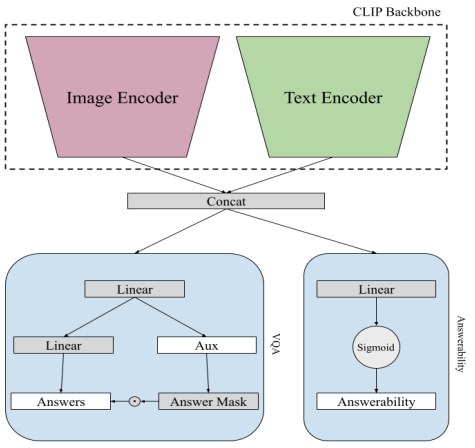

Figure 3.

Datasets

The VizWiz Visual Question Answering dataset comprises images submitted by blind mobile phone users, each accompanied by a spoken question regarding the photo. These were further annotated with 10 crowdsourced answers per question. Additionally, the images have been evaluated for their answerability. The dataset includes 20,523 training image/question pairs and 205,230 train answer/answer confidence pairs. It also contains 4,319 validation image/question pairs, 43,190 validation answer/answer confidence pairs, and 8,000 test image/question pairs.

Dataset: [Visual Question Answering – VizWiz]
Github: https://github.com/jack-campbeli/VizWiz-VQA-ds598

Evaluation Results

Metrics

1. **Accuracy**: $min(\frac{\#\ humans\ that\ provided\ that\ answer}{3}, 1)$
2. **Average Precision Score (Answerability)**: $AP = \sum_n (R_n - R_{n-1}) * P_n$ where P_n and R_n are the precision and recall respectively.

Expected Results

- **High Accuracy:** Obtain a high accuracy score. Current accuracy scores lie between 42 and 74 on the evaluation server, so the goal is to reproduce a score equal or higher than this. The 'Less is More' paper had an accuracy of 0.68.
- **Balanced Precision and Recall:** This answerability metric produces a confidence score between 0 and 1, so the ideal score would be one near 1.

Final Results

- **Accuracy**: 0.751 on the Test VizWiz dataset
- **Answerability**: 0.798 Answerability score
- **Leaderboard Overall Score:** 61.14 (7th Place)

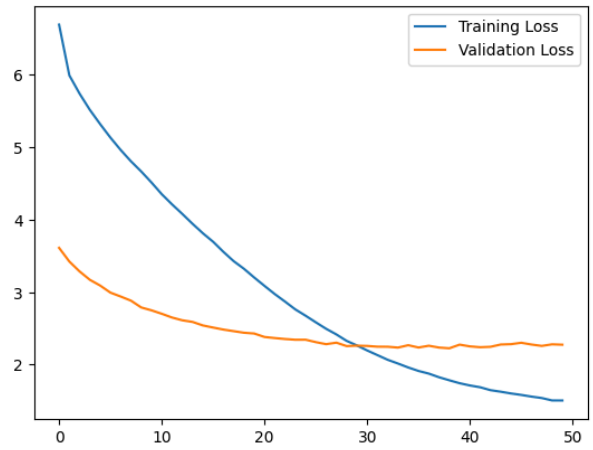

Figure 4a.

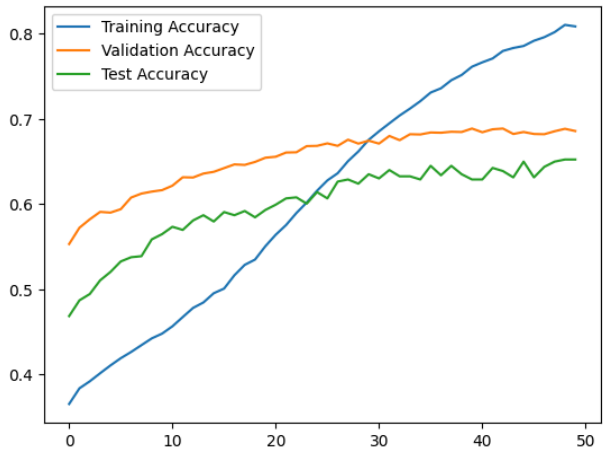

Figure 4b.

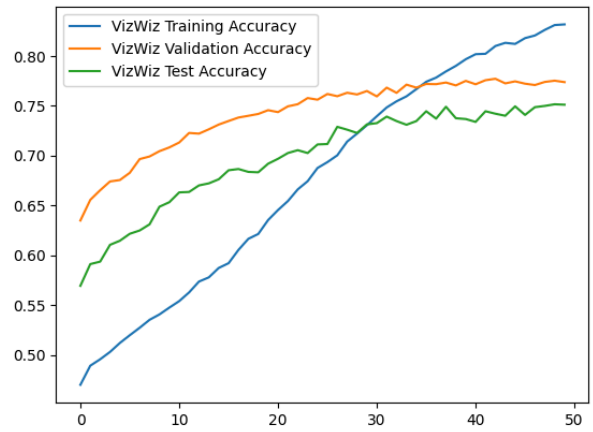

Figure 4c.

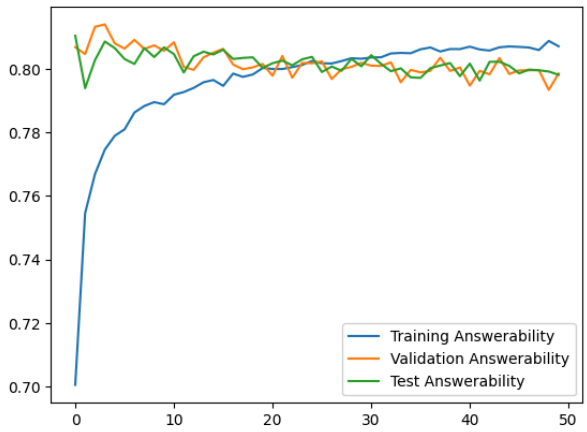

Figure 4d.

Conclusion

Our project focused on implementing a multimodal transformer model for Visual Question Answering for the VizWiz challenge. Our model achieved impressive results with an accuracy of 0.751 on the test VizWiz dataset and answerability of 0.798. We also were able to get into the top 10 on the VizWiz leaderboard. These findings highlight that our approach, inspired by the “Less is More” paper, was effective at accurately predicting the answers to questions about a set of images. This showcases its potential in assisting visually impaired people and providing valuable information on image features.

Our contributions demonstrate how using CLIP, adding hidden layers designed for VQA and answerability, and doing fine-tuning on the model were effective ways to approach this problem. Choosing the correct optimizer and learning rate helps improve the model. We found that it is also helpful to try out different weight decays as you change some of these other parameters to make sure we get the optimal loss and accuracy. Our work underscores the importance of

ongoing research and development in the field of computer vision and NLP, particularly to advance technologies for accessibility.

For future work, there are a few different areas to be explored. First, the hidden layers could be added to or changed since this is an area we have not explored in our research project. Additionally, expanding the dataset with more varied images could be helpful for training. Another area that could be interesting to work on is exploring methods of interpretability of model predictions that could be useful in addressing biases we did not anticipate in the original data. Overall, continued efforts in this direction hold the promise of further advancing VQA methods and technology for practical applications in the real world.

Appendix

Figures

References

1. Agrawal, Aishwarya, et al. *VQA: Visual Question Answering*. arXiv:1505.00468, arXiv, 26 Oct. 2016. *arXiv.org*, https://doi.org/10.48550/arXiv.1505.00468.
2. Deuser, Fabian, et al. *Less Is More: Linear Layers on CLIP Features as Powerful VizWiz Model*. arXiv:2206.05281, arXiv, 10 June 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2206.05281.
3. Gurari, Danna, et al. *VizWiz Grand Challenge: Answering Visual Questions from Blind People*. arXiv:1802.08218, arXiv, 9 May 2018. *arXiv.org*, http://arxiv.org/abs/1802.08218.
4. Le, Tung, et al. "Vision And Text Transformer For Predicting Answerability On Visual Question Answering." *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 934–38. *DOI.org (Crossref)*, https://doi.org/10.1109/ICIP42928.2021.9506796.
5. Radford, Alec, et al. *Learning Transferable Visual Models From Natural Language Supervision*. arXiv:2103.00020, arXiv, 26 Feb. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2103.00020.