Less is More: Recreating a simple model to answer VQA

Mohab Mohamed Ali, Rana Mohamed Barakat

***Abstract*—Most architectures used for multi modial problems such as visual question answering rely on transformers or other costly methods for training and evaluation. In this paper, we present a simple solution that utilizes pre trained models to achieve significant accuracies with a simple model that doesn't rely on long training time or expensive hardware based on this architecture[1].**

I. Problem

Visual question answering is where an image question pair is provided and the model outputs an answer to the question with the context of the image. This is usually to help the vision impaired and is a great step towards a more accessible world powered by ai. This task can be achieved in many ways, with one of the most popular ones being transformers considering their recent advancements and magnitude of applications. However, transformers suffer from large complexity and long training times. We address this by utilizing openAI's pretrained CLIP[2] model for image and text encoding and a simple classifier to find a suitable answer to the question.

II. Dataset

We used VizWiz's VQA dataset that has a few flaws. The dataset was composed of crowd sourced answers for image question pairs that were provided by the vision impaired and thus the answers are not always correct and are not always consistent.

III. Vocab selection

For our vocabulary, we tried a few strategies including selecting the answer that most corresponds to a caption of the image using CLIP's ViT-B/32, however that produced too many classes which resulted in a high error rate in the validation set. We also tried the first most common answer, and most common with draws resolved by the previous selection method. However, in the end we adapted the strategy used by the original authors of this architecture, in which we selected the most common answer per image/question pair and draws are resolved by selecting the answer that most resembles the rest of the answers using pair wise Levenshtein distance which proved to produce the best results and we ended up with a vocab size of 4856.

IV. Input and Output

After our vocab set was created, we now tackle the input and output shapes. We started by augmenting the image by rotating 90 and -90, encoding all three images then adding the weighted average and that would be our image representation. We then concatenated the image encoding with the question encoding all of which was done by CLIP with freezing the model to minimize train and evaluation time. The output on the other hand was a label binarized set created by sklearn's LabelBinarizer that took all 4856 classes and returned them as one-hot-encoded vectors. We decided to use this encoding as opposed to label encoding to make sure not to add any bias to the training set. The model also outputs the answer type which we use as an auxiliary loss to help improve training.

V. Model

In training, we used cross-entropy loss with both outputs and Adam optimizer with a learning rate of 1e-4 and wight decay of 1e-5 to avoid overfitting. The model is a simple classifier where the input first goes through a fully connected layer that reduces dimensionality, then the output is normalized and goes through a high dropout layer. The output is then passed to a fully connected layer that maps directly to the vocab size.

Pattern recognition Spring 2023

Answer acc	Answer type acc	Answerability acc
xx %	xx %	xx %

VI. Results

With our architecture and vocab, we achieved xx % in answers, xx % in answer types, and xx % in answerability. Training time took 15 minutes for 200 epochs on Kaggle's TPU1000 which was enough to achieve convergence.

VII. Conclusion

This problem, along many other multipmodial problems can be solved with simpler models that don't require the enormous time and complexity requirements that transformers impose.

References

[1] F. Deuser, K. Habel, P. J. Rösch, and N. Oswald, "Less is more: Linear layers on clip features as powerful vizwiz model," *arXiv preprint arXiv:2206.05281*, 2022.

[2] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.