School of Electrical Engineering and Computer Sciences (SEECS)

Class: BSCS 11-B

Submitted to: Momina Moetesum

Date: 22/05/2024

Name	CMS ID
Sarwar Khan	380024
Umer Khan	385574
M.Ahmad Safvi	384629

aEYE

A VQA FOR VISUALLY IMPAIRED PEOPLE

M.Ahmed Safvi | Sarwar Khan | Umer Khan

Problem Statement

Visually impaired individuals lack access to visual information, hindering their understanding of surroundings and objects. Traditional methods like audio descriptions or tactile representations are often limited and lack spontaneity.

Objectives:

A Visual Question Answering (VQA) system tailored for blind users aims to provide access to visual content through natural language questions

This empowers blind individuals to independently inquire about their surroundings, identify objects, and understand scenes.

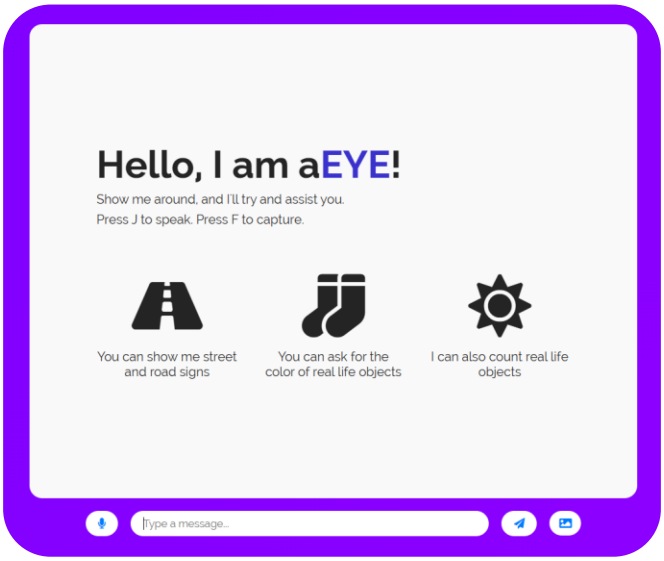
x

Implementation Details

Our solution consists of fine tuning a pretrained version of Microsoft's Git (GenerativeImage2Text)[i] model onto our COCOQA Dataset.

GIT is a Transformer decoder conditioned on both CLIP image tokens and text tokens. It is trained using "teacher forcing" on a lot of (image, text) pairs, which allows it to be used for tasks like image & video captioning, image classification, and, relevant to us, Visual Question Answering on images and videos.

This fine-tuned model is then used to perform inference onto an image and prompt provided by the user using the accessibility focused web application developed on Flask.

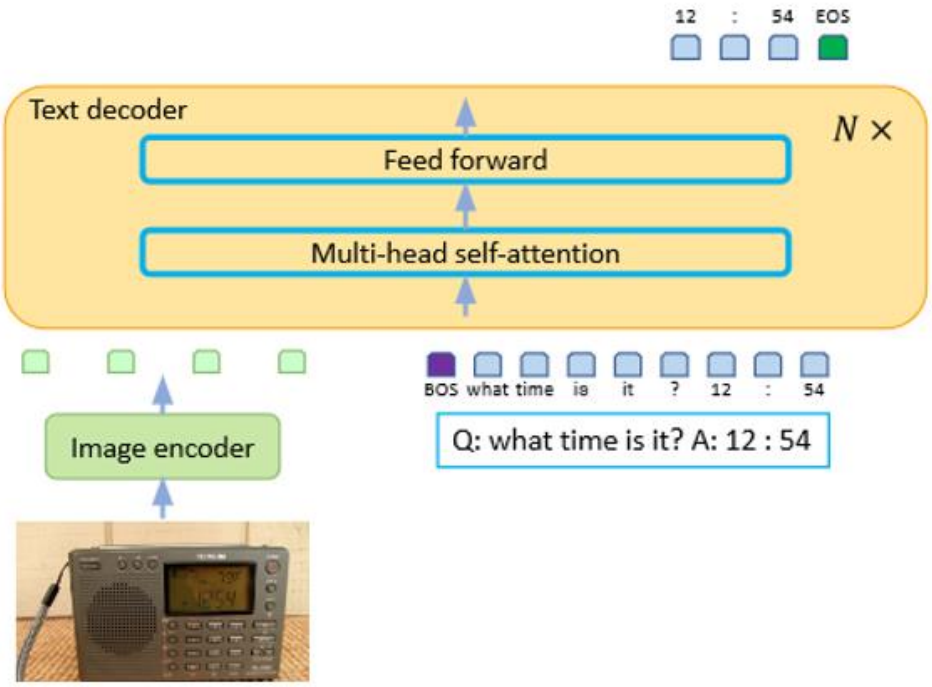


Architecture

The model consists of an image encoder and a text decoder. The image encoder is based on a contrastive pre-trained model. It takes a raw image as the input, and a 2D feature map as the output features. The text decoder is a transformer module, consisting of multiple transformer blocks, each including a self-attention layer and a feed forward layer.

The input text is tokenized and embedded (along with positional embeddings as well), and then gets concatenated with the image features from the image encoder. In our application, the input text is a concatenation of the question and the ground-truth answer as a special caption.

The model uses LM (Language model) Loss, which is Cross Entropy of the actual probability of an output, and predicted probability output given the image and previous words in the sequence. This loss is only applied to the answer and EOS tokens in our VQA application.

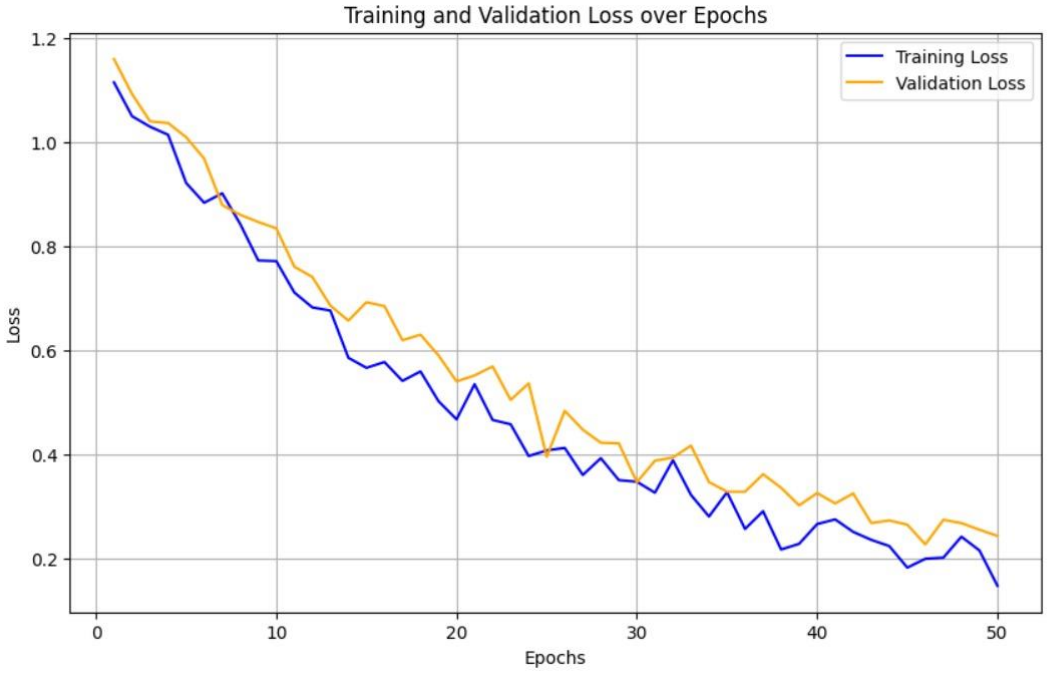


Fine tuning was done on the COCOQA dataset, over 50 epochs using the Adam optimizer, with results shared in the following section.

Experimentation Results & Analysis

The performance of the VQA system was measured using the cosine similarity between the predicted answer and the target answer due to it being a generative response, on which we achieved an accuracy of 0.72.

During training, loss was calculated and plotted, shown in the following:



Epochs	Training Loss	Validation Loss
0	1.11490142	1.15972252
10	0.77140536	0.83439491
20	0.50195270	0.59004251
30	0.39244781	0.42220665
40	0.29047290	0. 36166906
50	0. 19908394	0.226773820

Limitations and Future Directions

- The model currently most commonly provides one word answers, despite the underlying architectures capability to generate more. It is limited by the dataset its fine tuned on, which had one word answers.
- Model complexity and slow/long computation during inference (at least on a CPU) but considering the complexity of a VQA problem as a whole, this isn't too bad.
- Future directions can include fine-tuning on a larger dataset, in particular one with answers as opposed to one-word.
- The model can be optimized for local performance by storing the image features data, while still taking inputs for text prompts and performing inference off of that.

Contributions

Sarwar Khan	-	Carried out fine-tuning on the model
Umer Khan	-	Developed the web app and integrated inference
M. Ahmad Safvi	-	Researched on comparable and applicable implementations

[i] https://arxiv.org/abs/2205.14100