ECE 6960 - Deep Learning for Image Analysis
Project Report

Rishanth Rajendhran (u1419542)

Task
Visual Question Answering

Task description
Given an image and a question about that image in natural language, the task is to provide an answer to the question in natural language.

Why is this task important?
With the advent of ChatGPT, there is a growing interest in models that can communicate with lay-users who may use different kinds of inputs such as natural language text, visual images, videos, code, links etc while conversing with these models. While natural language understanding and generation has made significant strides ever since the paper “Attention is all we need” which introduced attention and transformers came out, there is still a large scope for improvement in the performance of multi-modal models, for instance models which take both text and images as input.

There is also a growing interest in extracting common sense knowledge which could be invaluable in many tasks such as machine translation, question answering and word sense disambiguation. One line of work focusses on harvesting common sense knowledge from large-scale language models such as OPT, T5 etc as they have been trained on large amounts of text corpora. Another source of common sense knowledge are images which are also available in large quantities on the internet. Models with visual understanding can thus be used to harvest common sense knowledge as well.

In this regard, visual question answering is a task that not only tests natural language understanding that models possess but also their visual understanding. An ablation study performed by the authors of the paper that introduced the dataset used in this project revealed that while 40% of the questions can be answered with just the question, providing the associated image nearly doubles the performance of the model. While a sizable fraction of the question could be answered with just word knowledge (all the more reason why common sense knowledge extraction is an important research area), being able to understand images helps the models do even better.

Dataset
VQA dataset (https://visualqa.org/download.html)

Example
Image:

Question	How many helmets are being worn?
Answer	1

Question	What store is in the picture?
Answer	caffe nero

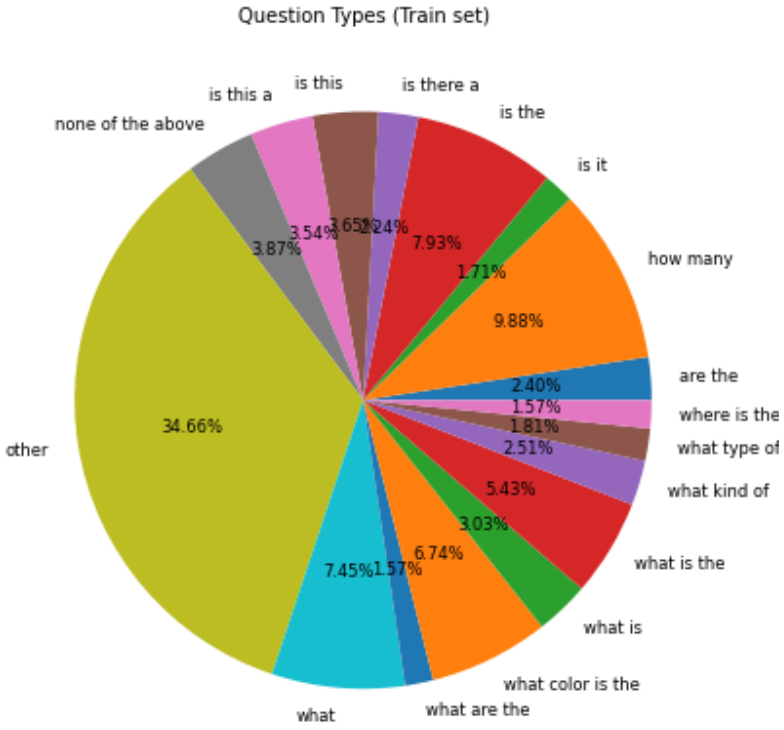
Question	Is the rider using proper personal protective equipment?
Answer	yes

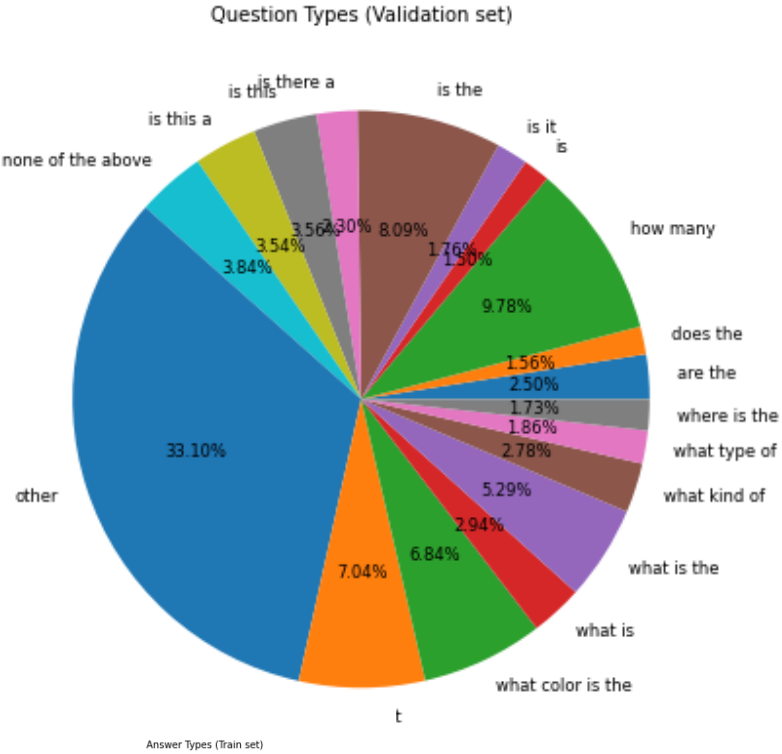
What makes this task challenging?
There has been a significant improvement in the performance of models for visual understanding over the years. While models such as VGGNet, ResNet, InceptionNet etc have nearly mastered the task of classifying images, models such as. U-Net, faster-RCNNs etc have made significant strides in tasks such as object detection and segmentation. The same can be said about models such as GPT and T5 which have made significant strides in natural language understanding and generation.

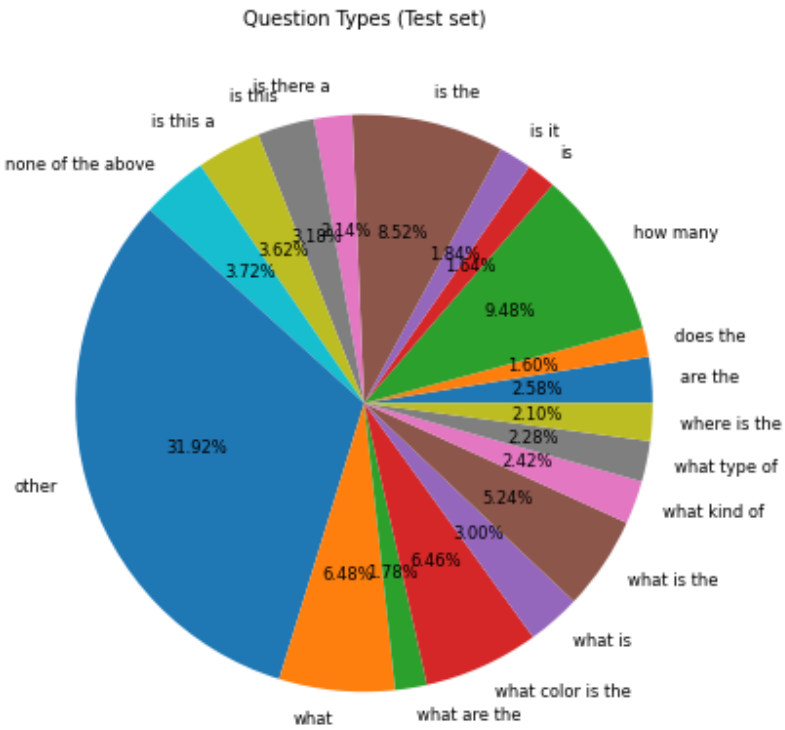
The VQA task requires a model that has the ability to look at an image given the question as context. For example, if the question is about the color of the hat of a person in an image, the model would be better off identifying the people in the image first and then finding the person wearing a hat instead of looking at the image as a whole. This would indicate that muti-task learning would be beneficial in this setting.

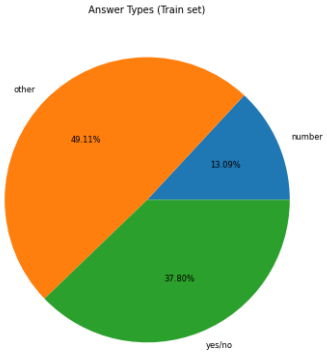
Data Statistics
Due to limited compute resources, a limited number of examples were sampled from the original train and validation sets. The test set was not used as it did not have gold answers. A fraction of the validation set was used as the test set.

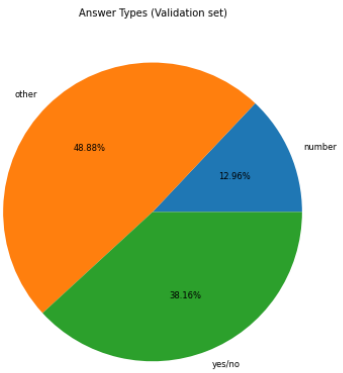
	No. of instances	Average length of answer	Size of vocabulary
Train set	45000	1.05693	2167
Validation set	45000	1.05636	2161
Test set	11250	1.0616	865

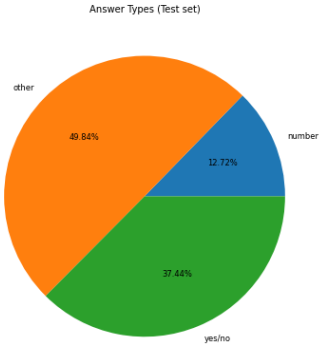






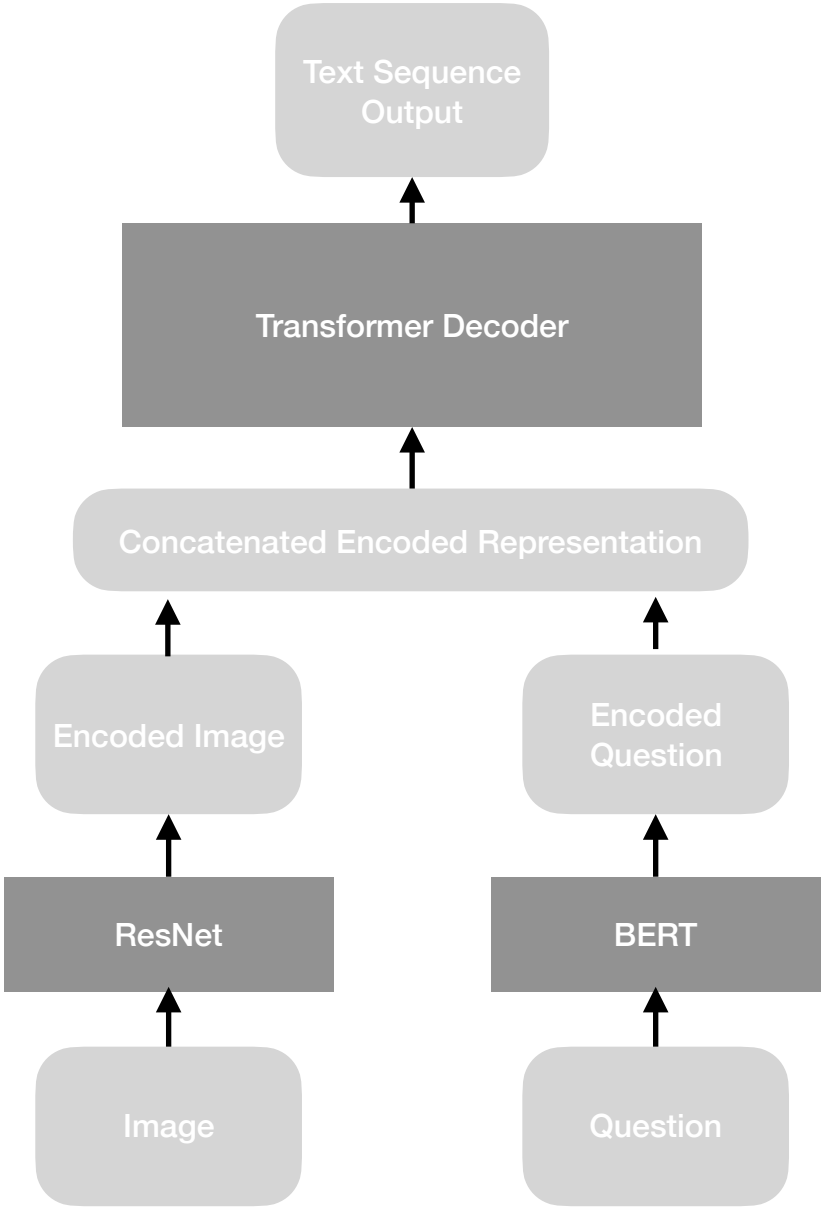






A large majority of the answers (~95%) were one word long and only a few (~3-4%) were two/three words long. About 1% of the answers were longer than three words.

No data augmentation was used as they are not suitable for this dataset. The ground truth may change depending on orientation, angle and color, depending on what the question is about. Random cropping might remove the object the question is about. Horizontal and vertical flipping can change the position of objects in the image which may pose a problem if the question uses that information (eg. what is the color of the hat the person in the top-right corner is wearing?)



Model
The paper that introduced this dataset made use of some convolution layers followed by a multi-layer perceptron (the last hidden layer of VGGNet) to encode the images and two LSTM layers to encode the natural language question. The overall design of the model used in this

project is somewhat similar: ResNet is used to encode the image, BERT model is used to encode the question, the encoded representations are concatenated together before being fed to a transformer decoder which then predicts text sequences.

While majority of the answers in this datasets were only one word long, for the sake of generality a transformer model capable of predicting text sequences of arbitrary length was used in place of a simple softmax layer over the vocabulary as suggested in the paper.

Teacher forcing was employed during the training process. Cross entropy loss was used for training with zero weights given to padding, start-of-sequence and end-of-sequence tokens to prioritize getting the answers right. At inference time, tokens were generated one-by-one until the model generated the end-of-sequence token.

Words from the 300-dimensional Glove embeddings were used to generate the vocabulary for the transformer decoder. All words were lowercased to prevent capitalization issues. An "[UNK]" token was added to the vocabulary to deal with words not in this vocabulary.

Limitations
As discussed above, this task would definitely benefit from a model that takes the question into account while generating a representation of the image. Segmentation models could also. be helpful for this task. However, due to the limitation on compute resources the model used in this project was made extremely simple in its design.

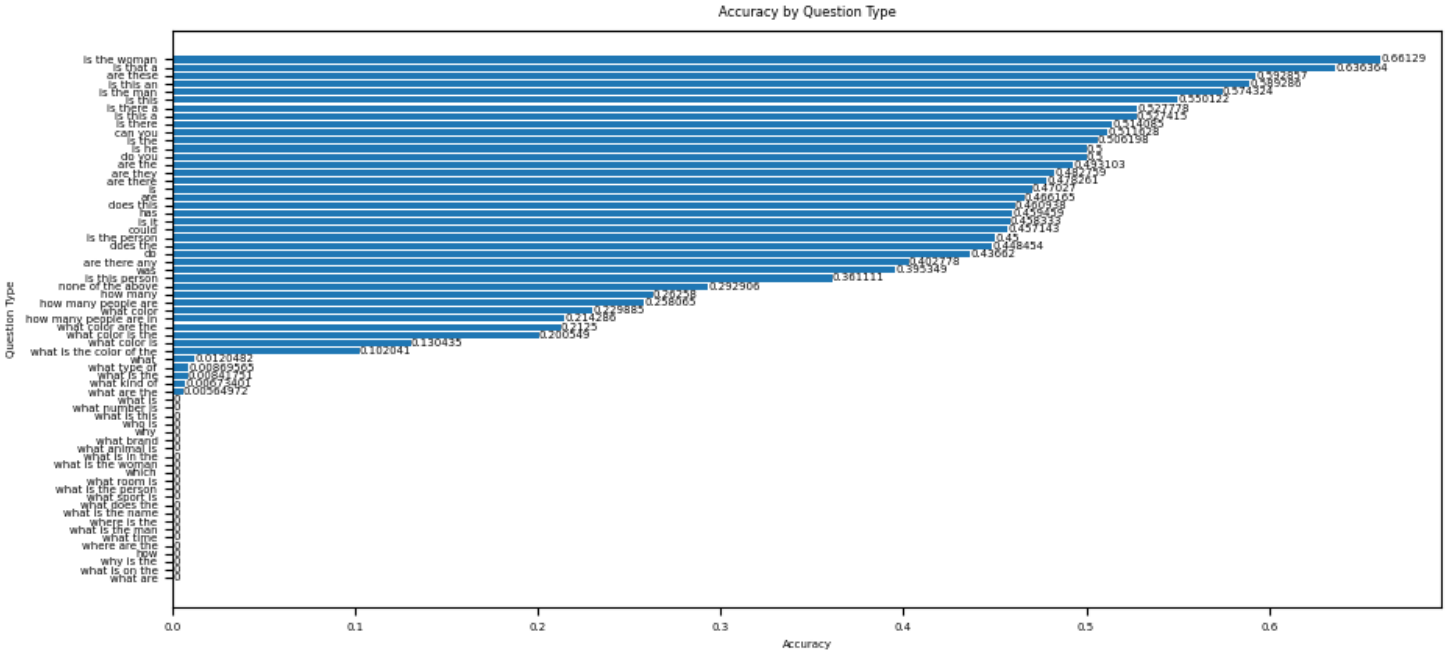
Compute Time
Training: 15 hours on a GPU (for 20 epochs)
Testing: 0.5 hours on a GPU

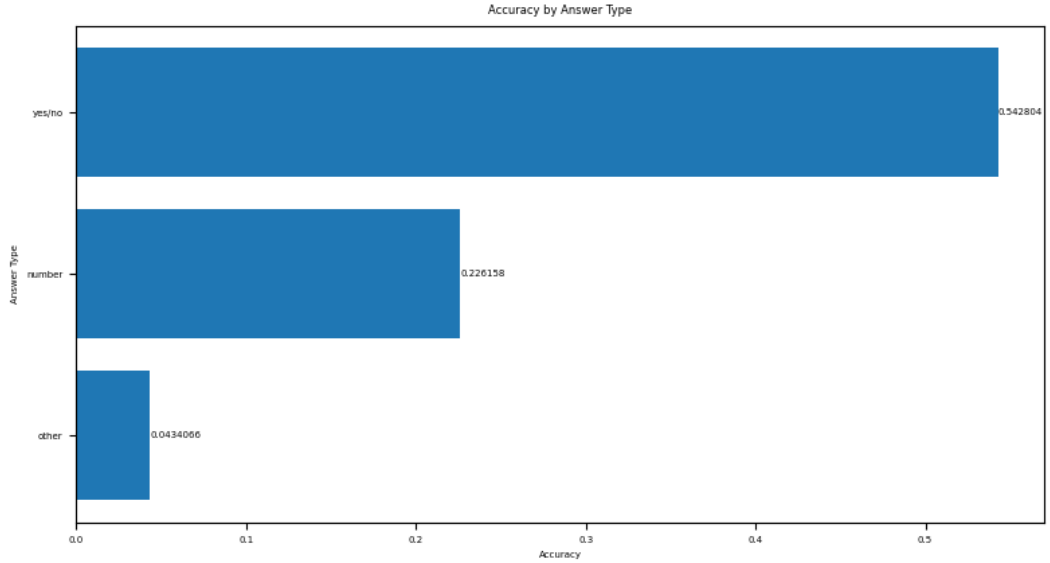
Results
Accuracy was used to evaluate performance as in the paper that introduced this paper. It was proposed earlier to use F1 scores (Precision and recall) instead of accuracy as it could be a harsh evaluation metric, but after looking at the answer distribution it was decided that accuracy is a good enough measure given than a large majority of the questions are only one word long.

Dataset	Accuracy
Train set	26.27%
Validation set	25.78%
Test set	25.91%

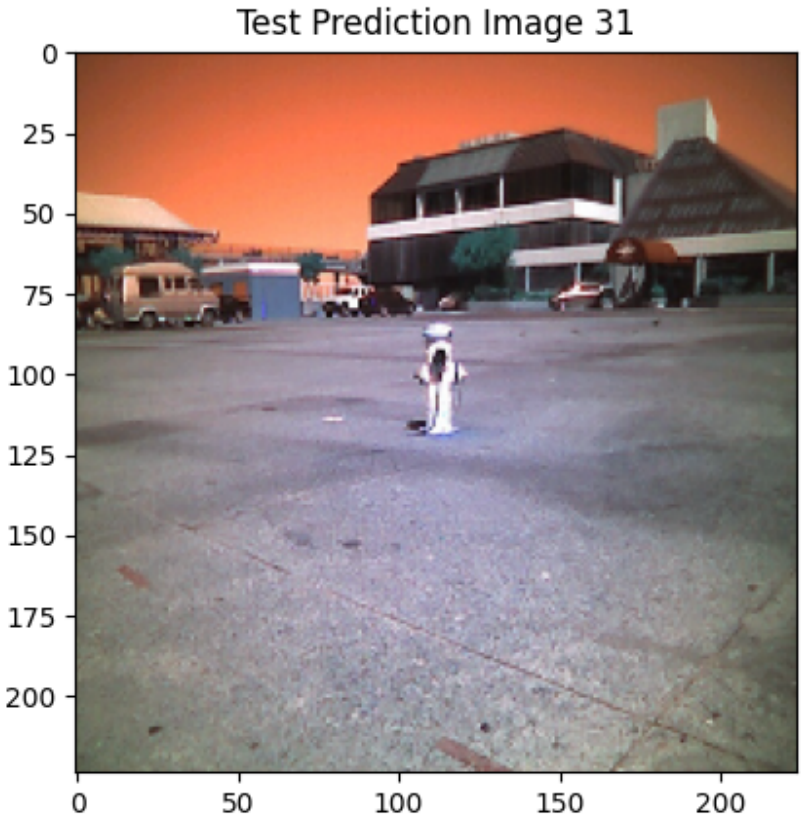
Performance is considerably lower than what was reported in the paper eight years ago. There could be several reasons for the poor performance. One, only a small sample (~10%) for the whole dataset was used for training and testing in this project due to limited compute resources. Two, The smallest ResNet model was used as the image encoder. Better performance could have been achieved with larger ResNet models or better models such as InceptionNet.

Performance on test set: A closer look





Unsurprisingly, the model performs best on questions which have yes/no as their answer. The model gets 54% of such questions right which is only slightly better than chance. While performance on questions which have numbers as their answer is not as bad as the performance on questions with other answer types, it is important to note here that the transformer decoder can only predict numbers present in its vocabulary (which has about 40k tokens).



Question: What color is the fire hydrant?
Answer: White
This is an interesting case since fire hydrants are usually red in color but the model correctly predicted the color of the fire hydrant in the image.

Conclusion
Both the training and validation accuracies continued to improve with more epochs of training. Due to limited compute resources, training had to stopped after a number of epochs even though neither the training nor the validation accuracy had plateaued or dipped. Better performance could have be achieved with more training.

As mentioned before, clearly this task would benefit from using the question to focus on the appropriate objects in the image while generate the feature representation of the image. Again due to limited compute resources, it was opted to simply finteune the pretrained ResNet18 model from torchvision. It would be worth-exploring if using object segmentation models perform better for this task.

Vocabulary of the transformer decoder is extremely limited (only 40k tokens). This severely restricts what the model can predict. Using word-piece tokenization can help improve the range of the decoder which may lead to better performance.

Exploring better ways of concatenated the feature representations of the image and that of the questions could prove to be beneficial. In this project, a simple concatenation was performed. Perhaps, a bit-wise multiplication as suggested in the paper would do better.