Advanced NLP Exercise 1

1 Open Questions

Question 1

1. QNLI - The dataset aims to measure a model's ability to discern whether the answer to a question is explicitly mentioned within the provided context sentence. Understanding whether a paragraph contains information that answers a question, is a facet of language understanding.
2. TriviaQA - The dataset attempts to measure a model's ability to understand and successfully answer relatively complexly formatted questions given a set of documents that may contain the answer. Essentially, this dataset aims to evaluate the reasoning capabilities of models when dealing with complex questions in a Natural Language Inference (NLI) context.
3. Quoref - This is an coreference dataset using QA. The model is expected to be able to link different references to the same entities. Being able to connect between different references is clearly a facet of language understanding.

Question 2

1. Interactive Summarization
   (a) Task Definition: create a summary of a document with user feedback. Feedback can be about length of the summary, summary focus etc.
   (b) I've seen the reddittifu dataset used for this with rouge scores domain is reddit with approximately 125k samples.
   (c) Challenges: Evaluating the model output. how to use user feedback for he summary.
2. Multi-document summarization
   (a) Task Definition: create a static summary of the information contained in a set of documents that is nonrepetitive and concise.
   (b) Multi-News, news summarization, approximately 45k.
   DUC 2004, news, approximately 50 samples.
   In both datasets samples are sets of articles
   (c) Inherent challenges:
   removing repetitiveness - recognizing information that appears in more than one document and making sure it appears only once in the summary.
   Temporal correctness - recognizing when information from one article is more up to date than another and modifying the summary accordingly.

Question 3

The efficient parallelization benefit of transfrmers applies both at training and at inference.

Encoding individual words is slower for an RNN, as representations for words are compted sequentially. In the transformer model word representations are computed non-sequentially, and therefore can be parallelized. This effects both train and inference, as for both one must process sentences.

The one big advantage of RNNs is that the self attention in the transformer architecture is around $O(N^2)$ in compute if the self attention is unrestricted in length, where N is the number of input tokens. This implies that the parallelization may not be worthwhile for very long input sequences. Of course, this must be weighed against the parallelization benefit.

Tl;dr inference and train both benefit from the parallelization.

Question 4

1. I would finetune ELECTRA-base as I don't have money for InstructGPT or the resources to run T5 XXL.

2. Since I only know if the sentences in the pairs are the same or not I would build a bert classifier that would receive concatenated pairs of sentences and classify them as same/not same. Sentences would be concatenated using a [sep] token and the classifier would be trained on the [CLS] vector made by said said sentences. Basically, I would add a classification layer for the same/not same prediction on top of bert and then train on the sentence pairs.
   Tl;dr build a classification model using bert.

3. Reasons:

   (a) For: ChatGPT is state of the art, and you want to compare yourself against the best in class.
   For: You like banging your head against the wall and this is an excellent way to do so.
   For: interacting with OpenAI's API is a pleasent experience.

   (b) Against: ChatGPT is closed source and you have no idea what your actually comparing yourself against(changes in model version, unknown training data/methodology).
   Against: The version of ChatGPT that your comparing against could be depracated without notice while you're still running your experiments and you'll be SOL(not to mention your experiments won't be reproducible).

2 Programming Exercise

https://github.com/elkanatovey/anlp1