Tutorial 10

Question 1

The Masked Language Model (MLM) pretraining objective is central to the functionality of models like BERT. MLM randomly masks out tokens in the input data and trains the model to predict these masked tokens. Given the sentence "The quick brown fox jumps over the lazy dog", assume that during MLM pretraining, 2 tokens are being masked.

1) Construct an example input sequence for the MLM pretraining given the sentence "The quick brown fox jumps over the lazy dog".
2) For the input sequence you have created, specify what the expected output labels would be during the training. What does the objective function look like?
3) Identify two downstream tasks where a pretrained MLM model like BERT would be expected to excel and provide reasons based on the features of the MLM pretraining process.
4) Describe the steps involved in fine-tuning an MLM-pretrained model for the downstream tasks you have identified. What modifications, if any, are typically made to the model architecture during fine-tuning?

Question 2

Decoder-only models like GPT (Generative Pretrained Transformer) are designed to predict the next word in a sequence, making them well-suited for generative tasks.

1) Explain the concept of masked attention in a decoder-only model. How does it ensure that the model predicts each subsequent word based only on the words before it?
2) Suppose the decoder generates a partial sequence: “<START> I love”. Describe the functions used in masked self-attentions when predicting the next token (assuming a single head without position encodings).
3) Illustrate how a decoder-only model, traditionally used for generative tasks, can be adapted for a classification task such as sentiment analysis. Discuss the modifications needed to convert the model's output from sequence generation to a classification format using this example: “This movie was fantastic.”

Question 3 (related to knowledge in last lecture)

Prompting is a technique used to guide language models, particularly those based on the Transformer architecture, to perform specific tasks by feeding them with a carefully crafted input.

1) Discuss how prompting (or in-context learning) is different from traditional learning paradigms such as finetuning.

2) Using the example of sentiment analysis, compare zero-shot and few-shot learning approaches in the context of prompting. How would the prompts differ in each case, and what are the expected outcomes?

3) What does "instruction finetuning" refer to? Discuss how instruction finetuning is used to train and improve a language model.

Code Exercise: Transformer Models

https://colab.research.google.com/drive/1k2mXI486kCSDXKrlahgxmgin8iGBWJXL?usp=sharing