NLP Report

Data and model analysis

Theorical questions

1. What is the purpose of subword tokenization used by transformer models ?

Subword tokenization follows two principles:

- No splitting for common words
- Splitting of the rare words into smaller meaningful subwords

For example the word « theorical" is split into « theoric" and "al"

By doing so we have a vocabulary size and a out-of-vocabulary size way less important than a basic word tokenization. And It allows the vocabulary to have less different meaning for similar words.

2. When building an encoder-decoder model using an RNN, what is the purpose of adding attention ?

What problem are we trying to solve ?

Encoder-Decoder models needs to compress all informations of a sentence into a vector with a fixed length. But when the sentence become longer it is harder for the model to keep all the information needed. So the problem is to handle long sentences successfully.

How does attention solve the problem ?

Attention models focus on important part of the input. In fact the model puts a weight on each part : a bigger one on the most important ones and a smaller one on the less important ones.

3. In a transformer model what is the multihead attention used for ?

Multihead attention is a module of the model which runs many attention model in parallels and compute a result based on all the parallels run. It is used in a transformer model in order to increase performance by have an input way more precise.

What are we trying to achieve with self-attention ?

Self attention allows input to interact between themselves

Why do we use muliple head instead of one ?

It allows the model to embed the input data by multiple ways so it gets a better representation for the input.

4. **In a transformer model, what is the purpose of positional embedding ?**

We have seen in class that depending on the language the grammar change so the order between verbs subject etc also changes. To ensure that the input stays clear we need to use positional embedding which keep words order by assigning a representation for all the words in a matrix.

What would be the problem if we didn't use it ?

If we didn't use it the meaning of our sentences could change during the embedding and the model could understand the opposite meaning of an input sentence.

5. **What are the purpose of benchmarks ? And are they reliable ? Why ?**

Benchmarks are used to evaluate and compare performance of a model according to several metrics so we can improve it.

Like any others test it must be robust to be reliable. Thus, as long as the measurements, the data, the problem that the benchmark uses are meaningful it will be reliable.

6. **What are the differences between BERT and GPT ?**

They are kinda the same because they are based on a transformer architecture but BERT only uses the encoder part whereas GPT uses the decoder part.

What kind of transformer-based model are they ?

BERT stands for Bidirectional Encoder Representations from Transformers so it is in his name and GPT is based on an autoregressive language model.

How are they pretrained ?

Bert has been train as a masked language modelling and a next sentence prediction different use case we have seen in class. Whereas GPT was train as a Casual Language Modelling.

How are they fine-tuned ?

Bert is mainly fine-tuned for reading comprehension where GPT is fine-tuned for text generation

7. **How are zero-shot and few-shots learning different from fine-tuning ?**

Few-shots learning use a very small dataset to answer a specific problem while zero-shots learning waits until test time to try to predict classes without training

How do fine-tuning, zero-shot, and few-shot learning affect the model's weights ?