Amir Masoud Shaker

97243081

Part 1: Questions

1. what are the differences between Transformer Models and LSTMs?

Transformers are faster than RNN-based models (like LSTMs) as all the input is ingested once.
Training LSTMs is harder when compared with transformer networks, since the number of parameters is a lot more in LSTM networks.
Moreover, it's impossible to do transfer learning in LSTM networks. Transformers are now state of the art network for seq2seq models. Transformer networks give best accuracy and also comes with less complexity and computational cost.

2. What improvements in the Bert model make it possible to perform well on word processing applications?.

- BERT is the first NLP technique to rely solely on self-attention mechanism, which is made possible by the bidirectional Transformers at the center of BERT's design. This is significant because often, a word may change meaning as a sentence develops.
- By looking at all surrounding words, the Transformer allows the BERT model to understand the full context of the word, and therefore better understand searcher intent.

3. Explain about the difference between Bert model versus GPT-2 model.

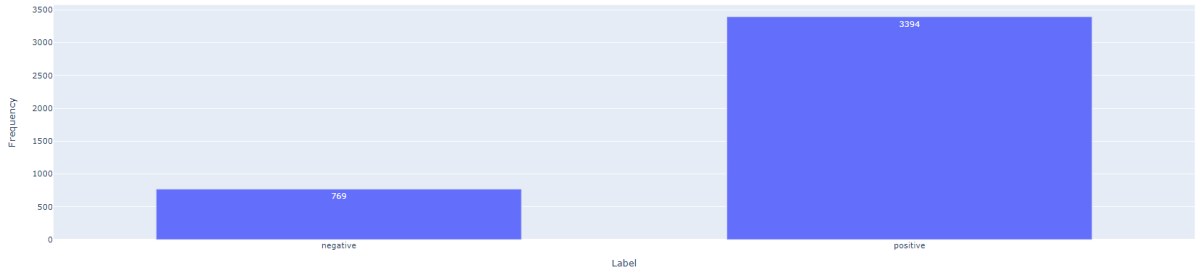
- BERT and GPT-2 perform quite differently on the token prediction task depending on the position of the token being predicted. For a fixed sequence length of 100 tokens, BERT performs best when the masked token is between positions 5 and 95, while GPT-2 tends to continually improve as context length increases. Interestingly, when the final token in the sequence is to be predicted, BERT's performance falls off dramatically, while GPT-2 performance remains stable.
- BERT is purely Bi-directional, GPT is unidirectional.
- GPT uses a sentence separator ([SEP]) and classifier token ([CLS]) which are only introduced at fine-tuning time; BERT learns [SEP], [CLS] and sentence A/B embeddings during pre-training.
- GPT was trained for 1M steps with a batch size of 32,000 words; BERT was trained for 1M steps with a batch size of 128,000 words.

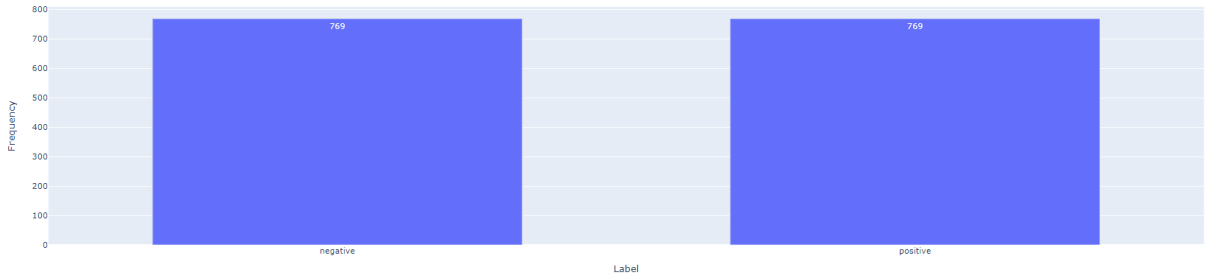
Part 2: Implementation

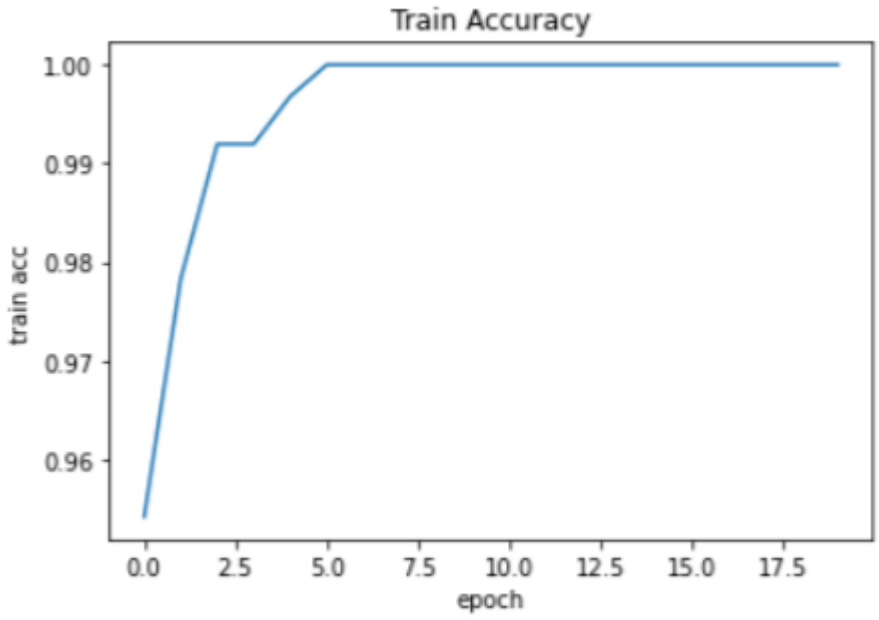
Colab link: https://colab.research.google.com/drive/1rWx0JKOljM8UUjafMSyVDrddsroq-SGf?usp=sharing

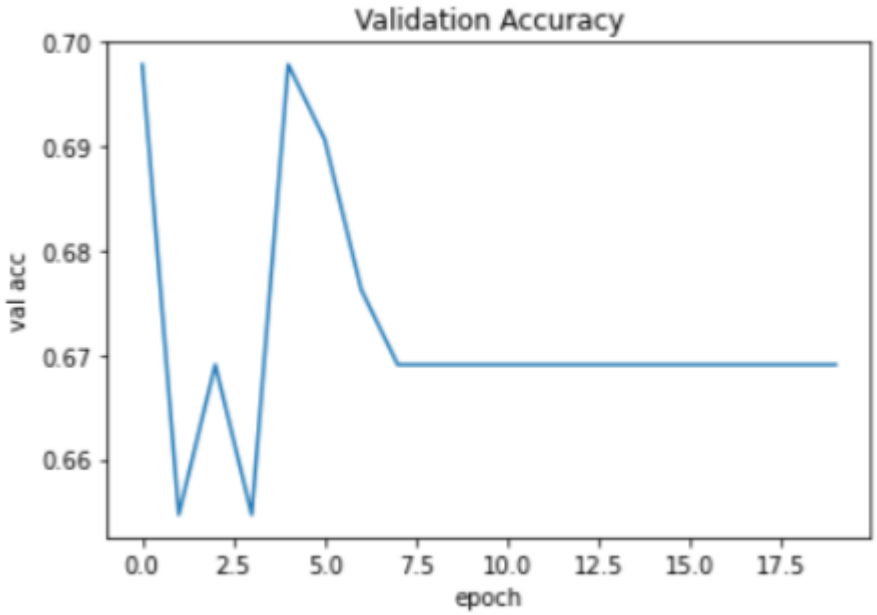
Code Explanation: I've provided comments in my code.

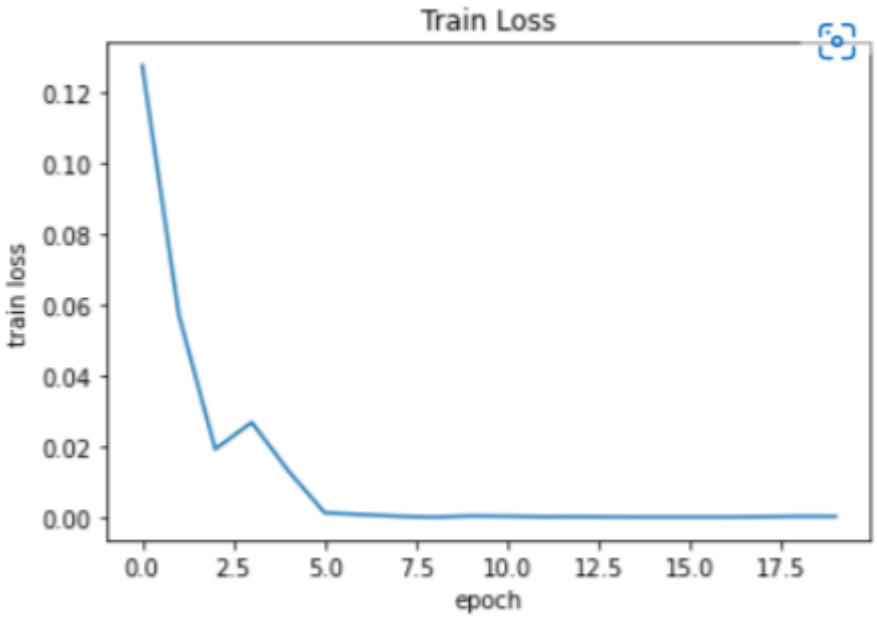
Diagrams:

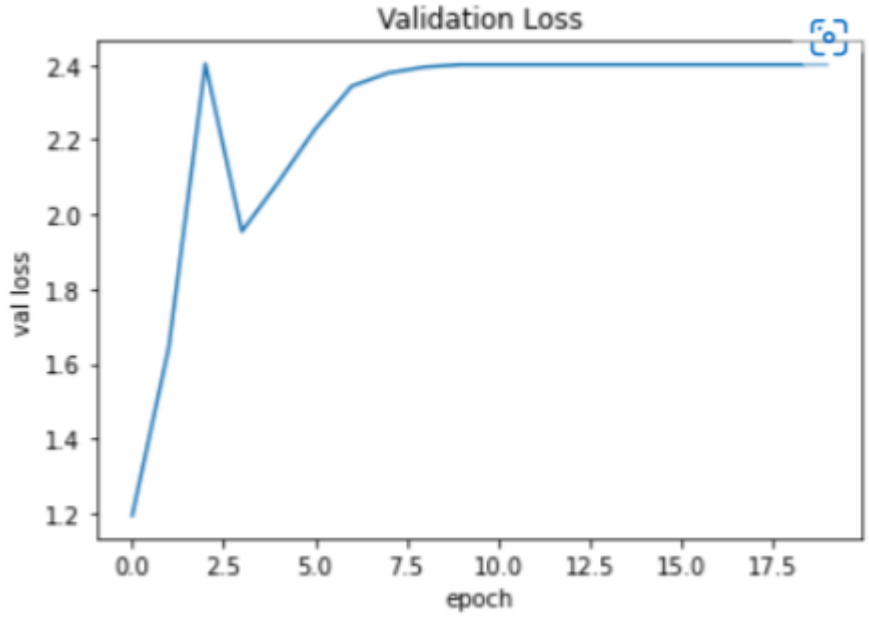







Validation Accuracy
0.70
0.69
0.68
0.67
0.66
val acc
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
epoch





References:

[Ref1]

[Ref2]

[Ref3]

[Ref4]