YOU Borachhun
borachhunyou@gmail.com

Lab session # 3
ALTEGRAD 2023

10/31/23

1 Question 1

The implemented square mask ensures that the prediction at the current position only depends on the known outputs at the previous positions. The positional encoding gives information of the location of words in the sequence. It encodes the position of the word into a vector using sine and cosine, and sums the vector with the word embedding.

2 Question 2

We have to replace the classification head because the output of both tasks are different. For the language modeling task, the classifier maps from the hidden dimension to tokens of the vocabulary, whereas the classifier maps to classes of text for the classification task.

3 Question 3

The base model has the same number of trainable parameters for both tasks:

- Embedding: vocabulary size × hidden dimension (embedding dimension)
- Positional encoding: 0
- Transformer encoder: number of layers × (parameters of multi-head attention mechanism + parameters of feed-forward network + (2 × parameters of layer normalization))
  - Multi-head attention mechanism: 4 × (hidden dimension × hidden dimension), where 4 indicates the 4 linear layers in the mechanism (see figure 1)
  - Feed-forward network: 2 × (hidden dimension × hidden dimension)
  - Layer normalization: hidden dimension

The number of trainable parameters for the classification head for:

- Language modeling task: hidden dimension × vocabulary size
- Classification task: hidden dimension × number of classes.

Note: All biases are excluded.

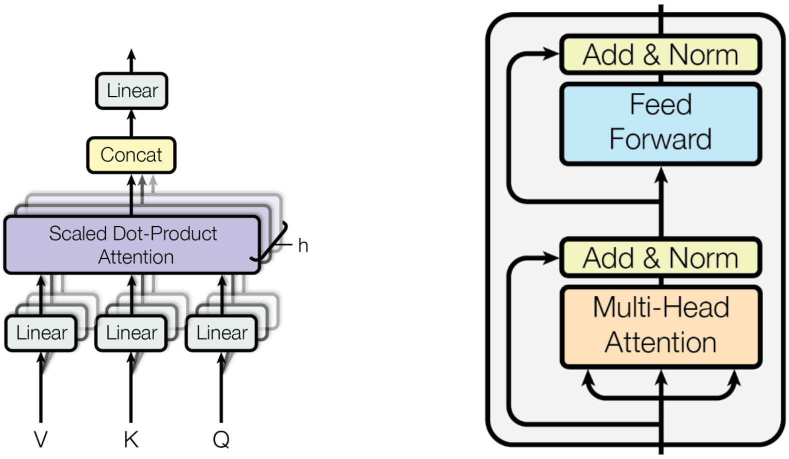

Figure 1: Left: Multi-head attention mechanism. Right: Transformer encoder architecture. Source: [2]

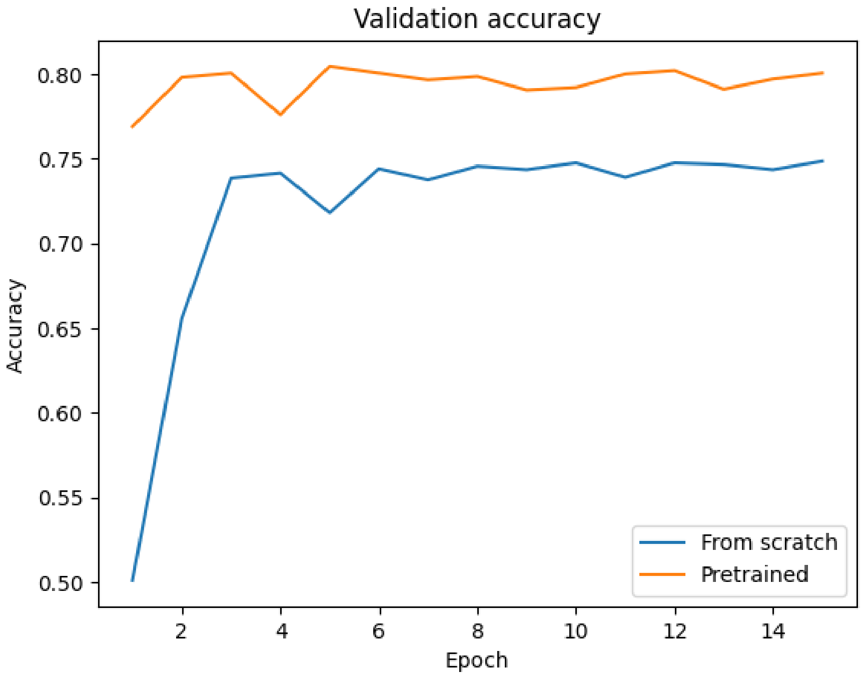

Figure 2: Evolution of the validation accuracy of training from scratch and transfer learning cases

4 Question 4

Figure 2 shows that the transfer learning setting has a better accuracy than the training from scratch setting from the first to the last epoch. The accuracy of both settings fluctuate for the first 6 epochs before stabilizing for the rest of the epochs, at just under 80% for the transfer learning setting, and just under 75% for the training from scratch setting.

5 Question 5

The language modeling objective used in this notebook is to predict the next word of the sequence based on the previous words (left context). In contrast, the masked language model objective introduced in [1] considers both the left and right context for the prediction. This bidirectional approach allows the latter model to capture more contextual information and thus have a better understanding of the relationships between words compared to the unidirectional approach of the former model.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arxiv:1706.03762*, 2017.