Yunhao CHEN
yunhao.chen@polytechnique.edu

Lab session # 2
ALTEGRAD 2022

11/07/22

Question 1

In the language model, we want to predict the next word based on the previous ones. Therefore, the model could not obtain the information after the current word. The square mask is in fact a lower triangle matrix. Its elements on upper-right side equal to 0 and others equal to 1. Multiplying the attention weights by the square mask before passing the weights to softmax, would ”masking” the information after.

Since the self-attention takes the whole sentence as input, there is no information about order of words. The positional encoding serves to offer the information about word position to the model, expecting to have a better performance.

Question 2

The language modeling is a multi-class classification task (the number of classes equals to the size of vocabulary), While the (document) classification task is a binary classification. We have to replace the last layer to adapt the output shape.

Question 3

We denote by n_t the number of tokens; by n_h the hidden size of attention layer (the embedding dimension). For language modeling task, look at Fig. 1

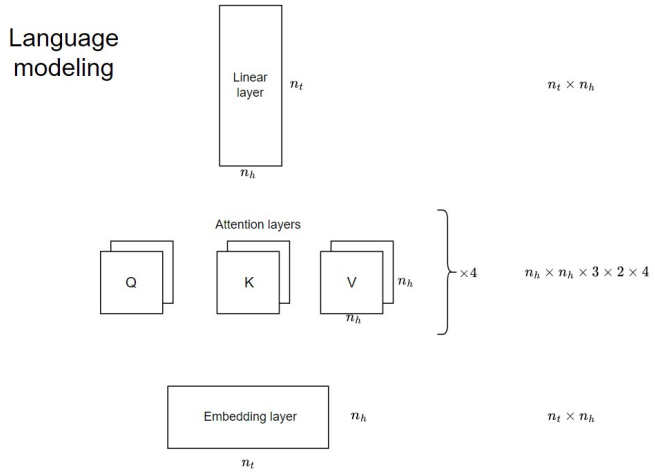

Figure 1: Parameters in language modeling task

For classification task, look at Fig. 2
Taking into account the fact that $n_t = 50001, n_h = 200$, we obtain: the total parameters in language modeling task equals to **20960400**; that for classification task is **10960600**.

Question 4

Look at Fig. 3. The pre-trained model gets a higher accuracy on test set than model trained from scratch. That means the transfer learning works. BERT, GPT and other pre-trained models gains the semantic information from their rich training set. We can make use of them on a specific task by fine-tuning the weights in network.

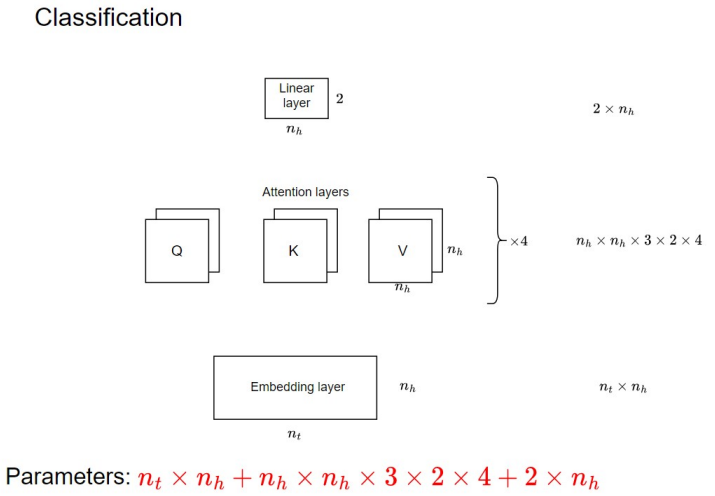

Figure 2: Parameters in classification task

In fact, there is another way to fine-tune: keep the weights in the attention layers frozen, changing the last layers and fine-tune on them.

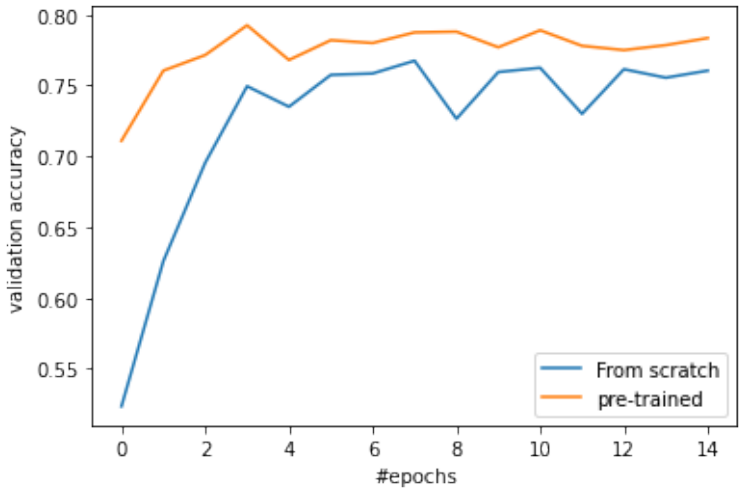

Figure 3: Validation accuracy

Question 5

The masked language model of BERT is bi-directional: it makes use of context words to infer the middle words. Our language model is uni-directional: use just the previous words the infer the next word.