Maxence Lasbordes
maxence.lasbordes@dauphine.eu

Lab session # 2
ALTEGRAD 2023

10/17/24

# 1 Question 1

The role of our square mask is to generate a triangular matrix, which allows the model to be auto-regressive. When predicting the next token, the model only considers the previous tokens and not the future tokens; otherwise, the task would be too simple.
The positional encoding is used to enable the model to understand and learn the significance of word positions in sentences, as the relative positions of words add context and meaning.

# 2 Question 2

We want the model to first understand the language in general, so we initially train it to predict the next token. Once the model has grasped the meaning of tokens, we utilize that understanding for classification. Language modeling involves pre-training a model to generate the next token based on a given context, while classification focuses on categorizing a sentence, which is a process of fine-tuning.

# 3 Question 3

- The **classifier** model consists of a single linear layer with the following parameters:

$$\begin{aligned} \text{Weights:} \quad & n_{\text{hid}} \times n_{\text{classes}} = 20000 \\ \text{Bias:} \quad & n_{\text{classes}} = 100 \end{aligned}$$

Thus, the total number of parameters for the classifier is $20100$.
- The **transformer** model has the following parameters:

$$\begin{aligned} \text{Embeddings:} \quad & n_{\text{token}} \times n_{\text{hid}} = 20000 \\ \text{Transformer Layers:} \quad & \text{Number of parameters} = 968000 \end{aligned}$$

Therefore, the total parameters of the transformer model is $988000$.

# 4 Question 4

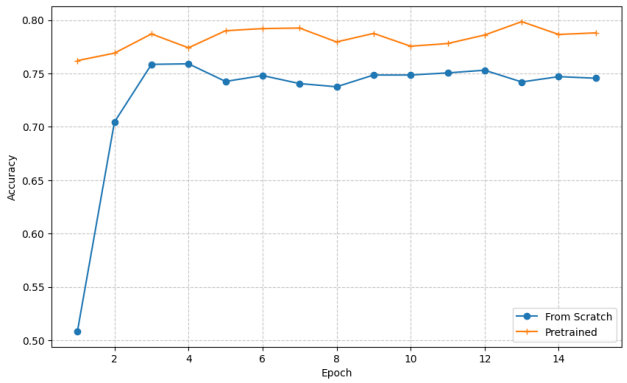


Figure 1: Validation Accuracy for the classifier trained from scratch and from pretrained weights

The graph shows that using pre-trained weights results in better accuracy, regardless of the training duration. This indicates that pre-training the model, enabling it to learn the language through next-token prediction, followed by fine-tuning, is more efficient than training both the transformer and the classifier from scratch.

# 5 Question 5

One of the limitation is the unidirectional language modeling compared to the masked language model objective introduce by BERT which allow the model to capture context from both directions. In the BERT training we mask some random words in the sentence and the model tries to predict those missing words by looking at the whole sentence rather than only looking at the tokens on the left.