Denis Duval
denis.duval13@gmail.com

Lab session # 3
ALTEGRAD 2022

11/22/22

1 Question 1

To count the number of parameters we refer to [1]. The encoder part consists of :
- an embedding layer and a positional encoding layer. The first one has $ntokens \cdot dmodel$ parameters. In the case of RobertaSmallFr, $dmodel = 512$ and $ntokens = 32000$ as said in the handout. If we use the architecture in the paper then the positional encoding layer is a parameter-free layer.
- $nlayers$ of the encoder part. The encoder part consists of a multi-head attention layer and a feed-forward network (which consists itself of two feed-forward operations). The multi-head consists of $nhead$ of scaled dot-product attention which is getting Q, K, V as inputs after each has been passed into a linear layer. The scaled dot-product contains no parameters. Then the outputs of all the heads are concatenated and passed again in linear layer. So for one single layer of multi-head attention we have:

$$n = nhead \cdot (3 \cdot dmodel \cdot dmodel/nhead) + (nhead \cdot dmodel/nhead \cdot dmodel)$$

where we use that $dv = dk = dmodel/nhead$. For the feed-forward network, we have simply $2 \cdot dmodel \cdot dff$ parameters where dff is the dimension of the inner layer. With Roberta Small it is equal to 512. In RobertaSmall we have $nhead = 8$. In this lab we have $nlayers = 4$. All in all:

$$nparams = 22675456.$$

2 Question 2

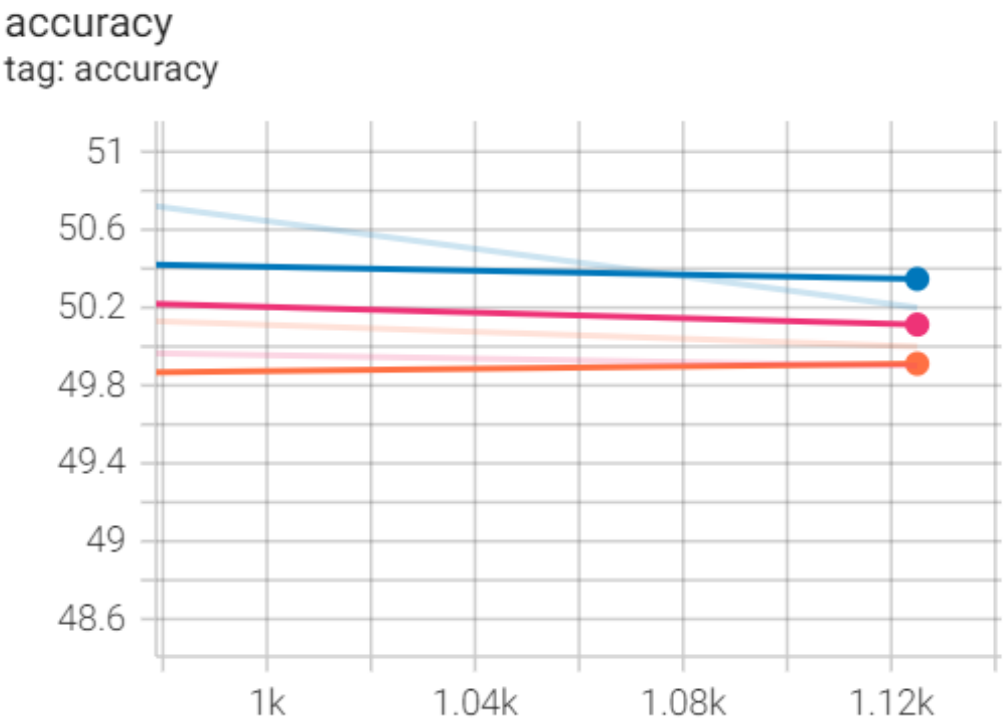

Figure 1: Accuracy of three seeds of Roberta Small Fr with fairseq

Overall the two frameworks are similar to use in practice. Both are easy to use because we don't have much work except the data preprocessing. Each of the two methods have practical pros and cons : with fairseq we have a little work to do to tokenize and binarize the data but then the scripts preprocess.py and train.py are easy to use. With the HuggingFace method, we only have to jsonify the dataset and then we can skip the steps of binarization and tokenization. From this point of view it might be a bit more convenient than the first

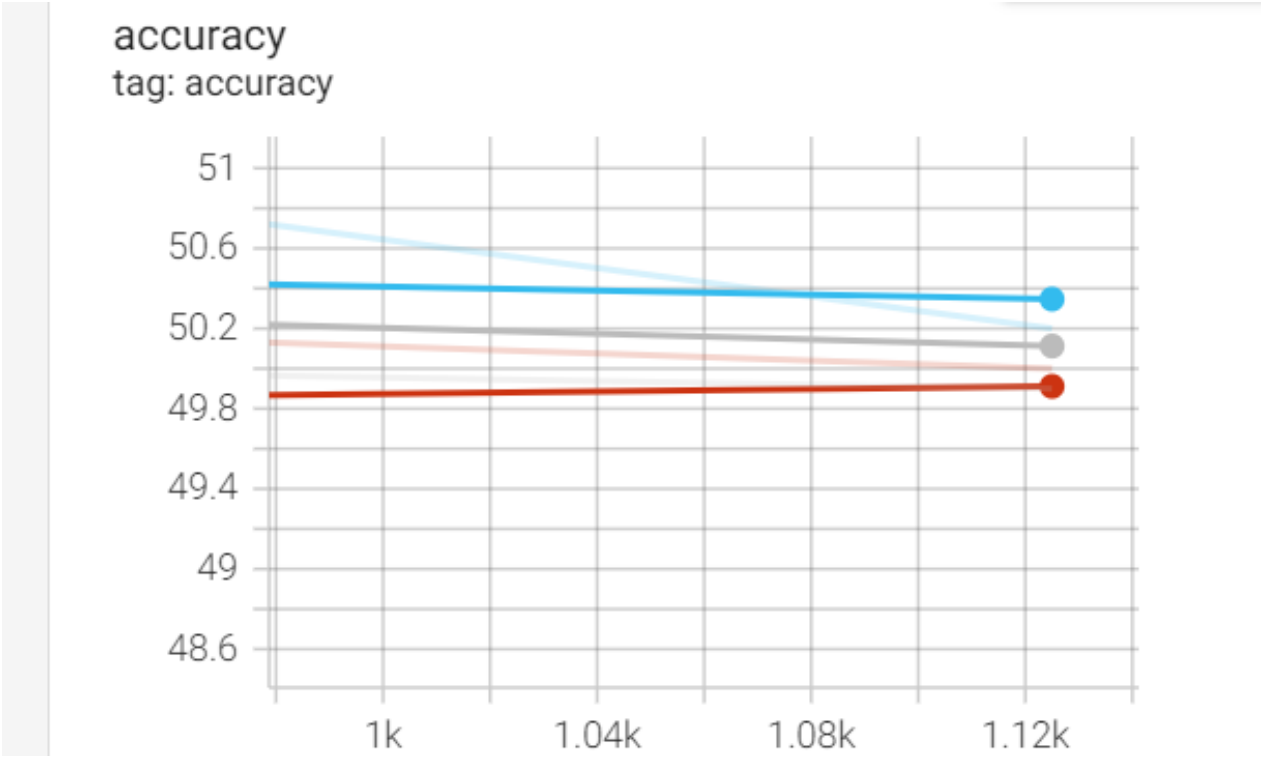

Figure 2: Accuracy of three seeds of Random checkpoint of Roberta Small Fr with fairseq

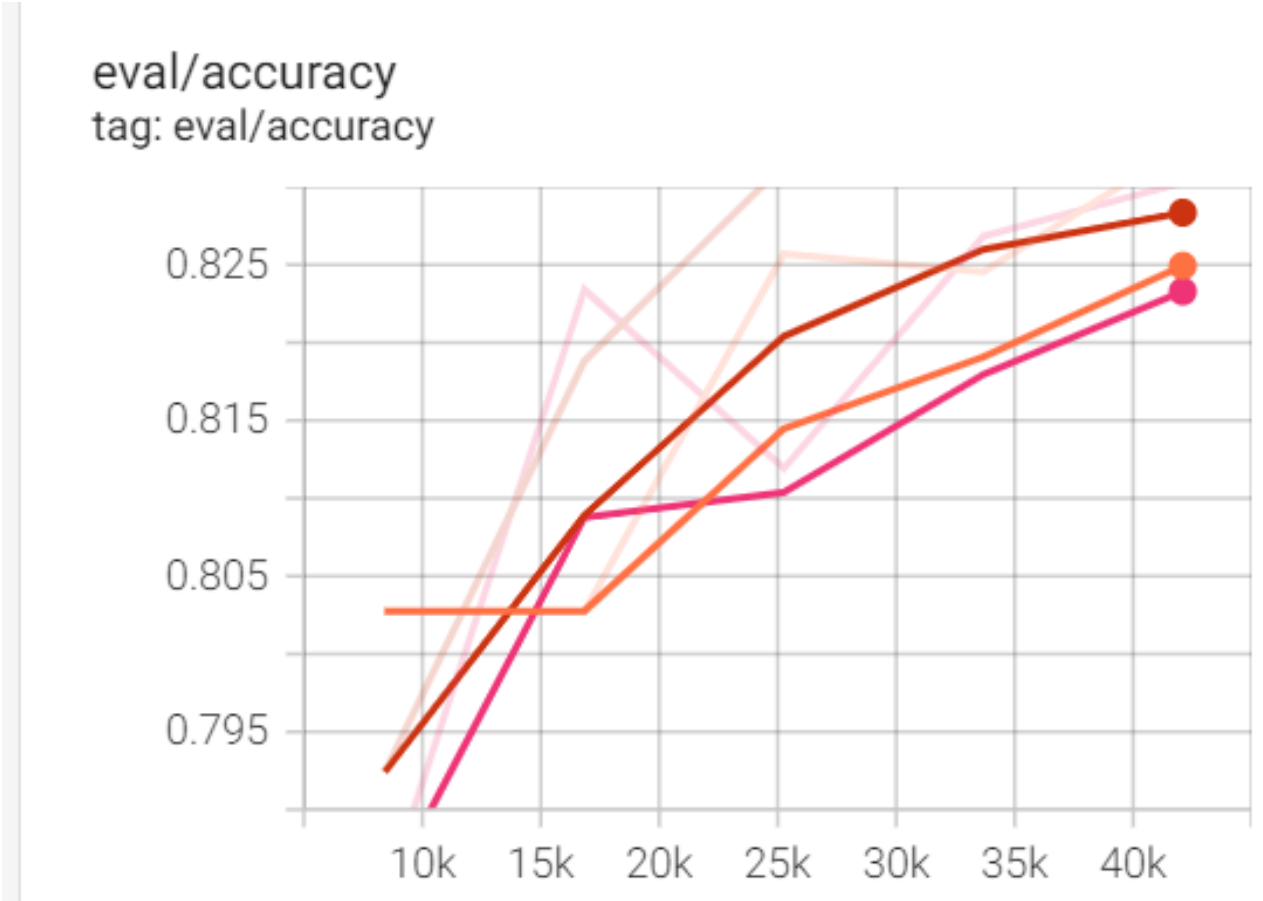

Figure 3: Accuracy of three seeds of Roberta Small Fr with Hugging Face

framework. The script run glue.py is as easy to use as fairseq's train.py.
Then for the results, we should get the same accuracies for the two frameworks however the three seeds of each model of the Fairseq framework don't really learn as we can see on the plots. I'm assuming that I've left

some mistake in my code in the first part of the notebook. Nevertheless we obtain satisfying results for the HuggingFace framework with a final accuracy around 0.825 for each of the three seeds.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.