Zakaria El Founoun
zakaria.el-founoun@student-cs.fr

Lab session # 3
ALTEGRAD 2022

11/22/22

# 1 Question 1

Number of parameters in the model:
* We have 32 000 vocab size with an embedding dimension equal to 512, so we have:

$$n_{embedding} = n_{vocabsize} * n_{embeddingdimension} = 32000 * 512 = 16384000$$

$$n_{feedforward} = 2 * n_{ffnembeddim} * n_{embeddingdimension} * n_{layers} = 2 * 512 * 512 * 4 = 2097152$$

Let's compute the number of parameters in one layer for the four layers :

$$We have: \quad n_{hid} = 512; n_{layers} = 4$$

$$n_{transf} = n_{layers} * (dim(K) + dim(V) + dim(Q) + dim(Projection_{ext})) = n_{layers} * 4 * n_{hid}^2 = 4194304$$

$$n_{total} = 16384000 + 2097152 + 4194304 = 22675456$$

# 2 Question 2

The two frameworks has many advantages and could be used in approximately the same tasks. Fairseq requires a manual tokenization and binarization which gives us more flexibility regarding this point, while HF transformer only requires to store data in a specific format and does not require any manual tokenization. In our case we can also notice that Hf transformers outperform Fairseq.
Choosing one over the other depends on the task, one is more flexible and the other requires less time for preprocessing the data.