MVA / ALTEGRAD 2022 / Lab session 2: Transfer learning

Imane Elbacha

14 November, 2022

1 The Model:

1.1 Question 1:

What is the role of the square mask in our implementation? What about the positional encod- ing?

Square mask: The purpose of the square mask implementation is preventing the ledtward information flow. In other words, the self-attention layers in the decoder allow each position in the decoder to attend to all positions in the decoder up to and including that position. In this case We need to prevent leftward information flow in the decoder in order to preserve the auto-regressive property. The way it works is that we implement this inside of scaled dot-product attention by masking out (setting to -∞) all values in the input of the softmax which correspond to illegal connections.

Positional encoding: We must introduce some information about the relative or absolute location of the tokens in the sequence because the model lacks recurrence and convolution and hence cannot use the sequence's order. For this reason, at the base of the encoder and decoder stacks, we append "positional encodings" to the input embeddings.

1.2 Question 2:

Why do we have to replace the classification head? What is the main difference between the language modeling and the classification tasks?

Regarding our model, it is constructed in two main blocks the base model and the classifier, the base model remains unchanged during both the pre-retraining and the training. The classifier executes two tasks the first being the pretraining using language modelling then the fine tuning using a classifier. Basically the same model can be used for two different tasks.

The main difference between Language modelling and classification, the first is generative and predicts the flow of words, whereas text classification is the process of categorizing text into organized groups

1.3 Question 3:

How many trainable parameters does the model have for languqge modelling the number of parameters is : $n = n_{embedding} + n_{positional} + n_{layers} * n_{tranformer} + n_{classifier}$ $n = size - vocab * n_{hid} + 0 + n_{layers} * (3 * n_{hid}^2) + 2(n_{hid}^2 + n_{hid} + n_{hid} * n_{vocab} + n_{vocab}$ by computing this equation we find the number of parameters to estimate is around 20000000 For classification the only difference is the n classifier which gives us around 10000000 parameters to estimate

2 Supervised Task

2.1 Question 4:

Interpret the results. We notice that the acuracy is higher for a pretrained model

2.2 Question 4:

What is one of the limitations of the language modeling objective used in this notebook, compared to the masked language model objective