

NEURAL NETWORKS AND DEEP LEARNING

Homework 3

Laura Iacovissi

September 7, 2020

1 Introduction

In the following report it is described the design process of the `trained_model.py` script as it has been submitted. The program performs a text generation task, with training based on the Bible. The final architecture is based on the Transformer Encoder class from PyTorch.

1. In the first section is described the data preprocessing;
2. in the second section is a about the architecture definition and tuning;
3. in the third section the results are analysed.

2 Data preprocessing

The dataset used for this homework is the Bible, retrieved from the Project Gutenberg page¹. It has been chosen for its high extension and because of its peculiar text structure.

The file has been preprocessed by removing non-unicode characters, lowercasing, removing single newlines, removing useless punctuation or numbers between words (e.g. underscores) and at the beginning of the paragraphs.

In order to prevent the network to learn from noisy sentences (i.e. too short sentences), a threshold equal to 18 on the length has been introduced. The resulting dataset, visualized in Fig. 1, has size 21374. The extracted vocabulary has size 10775.

The dataset is split in training and validation sets, with a percentage of 90% – 10%, shuffling the data. Tests will be performed by feeding random seeds to the final model.

The words are encoded defining a map between the vocabulary and a finite subset of the natural numbers. During the text generation, when the seed sentence has a word outside the vocabulary, a random seed in the set of known words is generated.

3 Architecture and hyperparameter tuning

The chosen architecture is the Transformer network, but employing only the Encoder part. As shown in Fig. 2, its main modules are the Embedding Layer, the Positional Encoding, the Multi-Head Attention and the Feed Forward. A square attention mask is used when passing the input to the model in order to mask also the future tokens, since the standard Transformer structure is only allowed to work on the earlier positions in the sequence. The Gradient Clipping technique is used to avoid problems with the size of the gradient norm.

Having this structure, the hyperparameters to be tuned are:

¹<http://www.gutenberg.org/>

ED	240
heads	4
layers	4
hidden units	2^6
batch size	1200

Table 1: Tuned parameters.

- **Embedding dimension (ED)**: dimension of the latent space on which the inputs (words) are projected;
- **Heads number**: the number of heads in the multi-head attention model. This number must divide the ED;
- **Number of layers**: number of hidden layers in the Feed Forward module;
- **Number of hidden units**: hidden units in each Feed Forward layer;
- **Batch size**: number of training examples in each batch.

Given the high dimension of the dataset and the nature of the network, the hyperparameter tuning has been done with an ad-hoc trial and error method.

Other parameters fixed to some constant values are: dropout probability = 0.4, evaluation batch size = 10, sequence length = 35 (the sentences are split in chunks, when not shorter), clipping constant = 0.25, loss function = Cross Entropy Loss.

During training, the training and validation losses are analysed. The initial number of training epochs was fixed to 100, but then reduced to 40 in order to avoid overfitting.

The final architecture has the parameters show in Tab. 1. The number of heads, layers and hidden units plus the ED have been increased until the losses were satisfactory; the batch size have been set to an high value in order to prevent overfitting. The optimized model performance are shown in Fig. 3.

4 Results

The tests for the generation performance of the trained model have been done by feeding different sentences to the model and asking to return 500 words. Here partial results are reported, full ones are available in the Jupyter Notebook "bible_transformer.ipynb".

Test n.1:

I want you to know that thing; and i am here.

And the priests of the levites, bigthan and anathoth with burdens, and caves, and together in all the heathen, and that chew the lump of the cud, and the coulters which did do it is christ; and for this bond of his own conceit.

Sith thou hypocrite, and musical instruments of the mire of much exceeding throat is full of the just man knoweth any work of it whispered, suffering the mizar. And it be unclean until the elect's sake there is a dream also, and milk of a lion ran through the first day of a strong drink of the king's chamberlains, and to me under foot of the third year; and the synagogue. And in a vile person, whereby the pharisees hath done to heart to the shaking of the menservants, and the beasts, and the grecians against the waters which is defiled herself will so do: and the serpent's youngest son of the cunning men laid equity shall be well the good courage: why did thyself.

Test n.2:

The immigration issue can be solved: let them be cast out a very great tumults in them in the midst thereof round about. Afterward thou shalt make the hindmost of thee: for thou wentest: and to their iniquities are fenced cities unto thee out of burning fenced cities for thou to hear the egyptians, and thou up against israel, saying, and solomon as soon as the canaanites, and they have ye in the ark of them: this day of them into captivity, o lord thy name of thy god for the lord.

So much less and the abomination unto thee, but let them home again the communication from us, and the camp also now the mountains, and read it on whom also not turn my the instruction of the lord my words of the heart and i have called amasa out of the lord, and it upon not down?

Then tell thine enemies, and when the temanite and the judges shall not so that thou from thy foot over all the great: as for thou hast ravished my seed after samaria of the ark of my name of the blast of the altar, and manoah lizard, and thy seed shall come up to recompense the cummin, wherewith i beat vehemently upon you by the lilies.

Both the examples show that the model is not able to generate sentences with clear meaning and logical reasoning. However, the grammatical structure is acceptable and the punctuation well positioned; the sentence length is sufficiently coherent with the Bible style.

A Appendix: Images

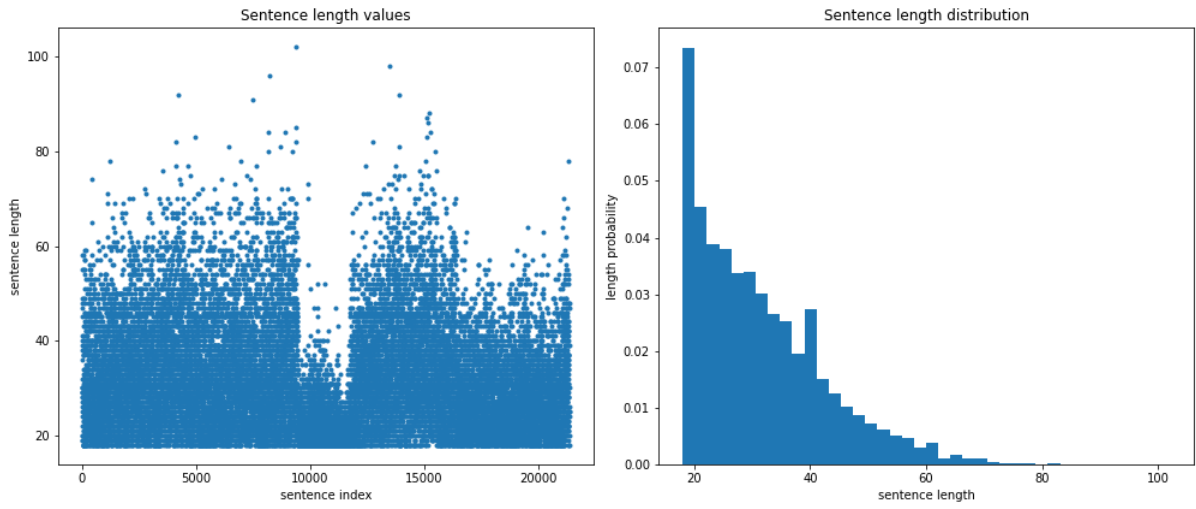


Figure 1: Filtered sentence distribution. Mean sentence length ≈ 31 ; max sentence length = 102; min sentence length = 18 (fixed threshold).

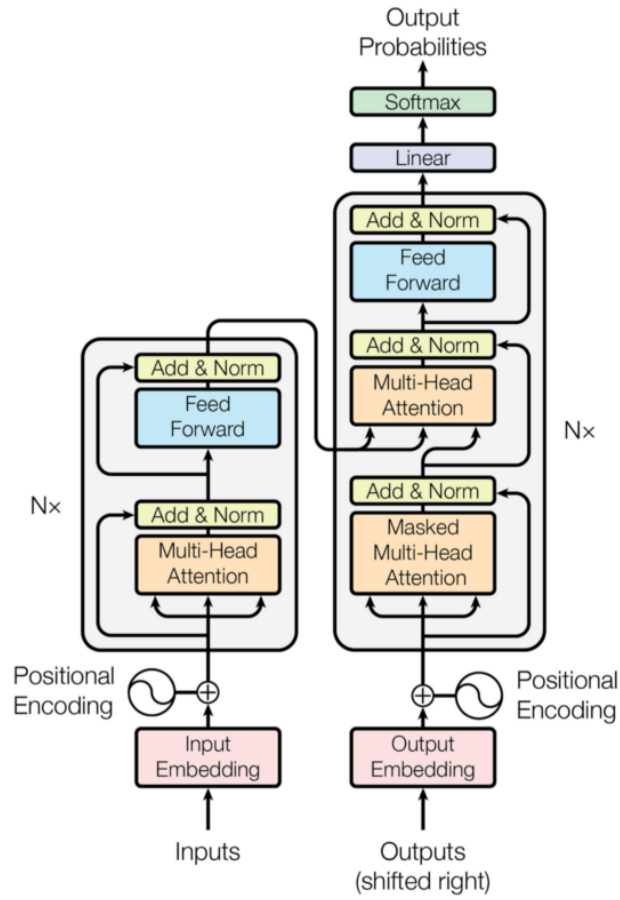


Figure 2: General Transformer Architecture.

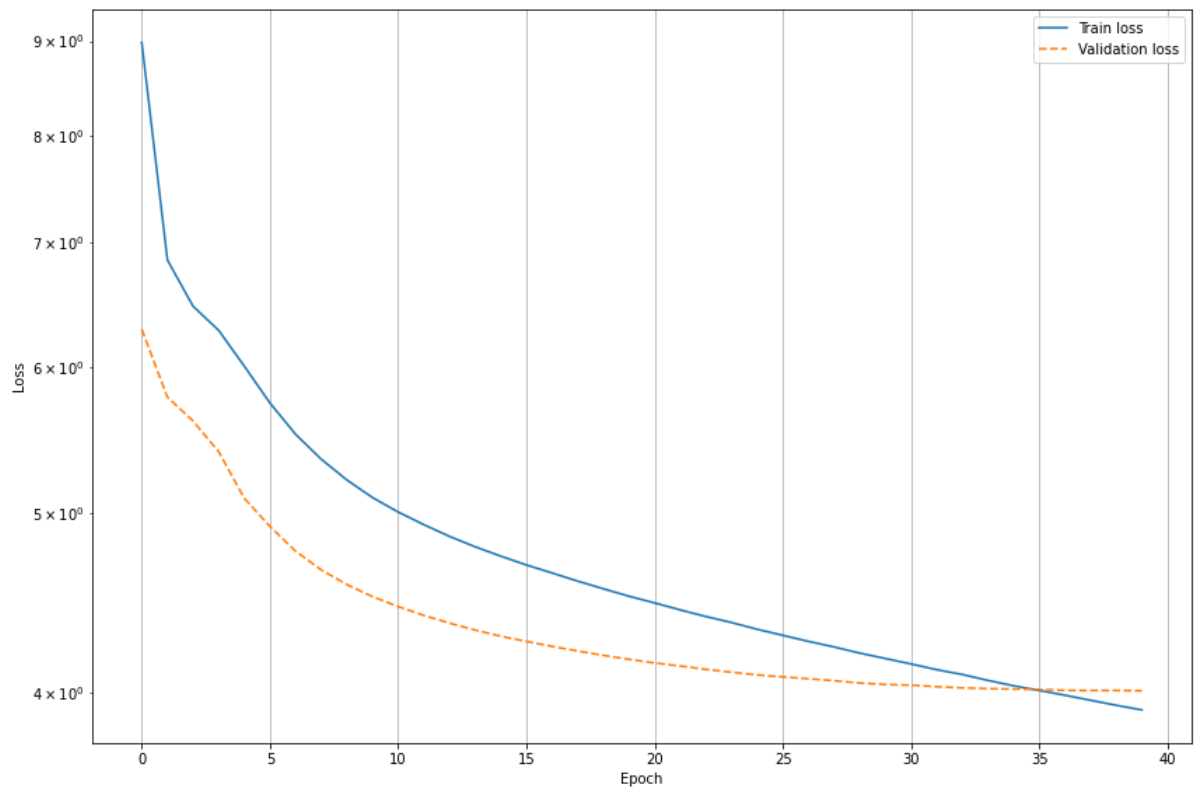


Figure 3: Model performance on the training and validation sets.