INCREASING THE PERFORMANCE OF THE MULTILINGUAL LANGUAGE MODEL WITH FNET ARCHITECTURE

Syromiatnikov Mykyta
Victoria M. Ruvinskaya, Ph.D., Professor of the Department of Software Engineering
Odessa Polytechnic National University, UKRAINE

ABSTRACT. This paper provides a result of the creation, pretraining, and evaluation of the multilingual language model for Ukrainian and Russian languages. For this purpose, the deep neural network based on the efficient variation of Transformer encoder architecture – FNet was used.

Introduction. Over the past few years, pre-trained language models, networks that determine the probability of characters, words, and n-grams occurrence in the sequence, have gained huge popularity in the field of natural language processing (NLP) due to their significant improvements in text understanding or generation and great generalization ability. Being pre-trained on massive text corpora, large language models, such as BERT, T5, and GPT, demonstrate state-of-the-art accuracy in most downstream tasks of natural language processing and generation. However, since most of the models are variations of the Transformer architecture, along with great accuracy, they also adopt some disadvantages: the need for huge computational resources and quadratic computational complexity due to the attention mechanism.

The objective. The work aims to develop the language model for Ukrainian and Russian languages using the recent advances in the field of deep learning for natural language processing and special attention being paid to the productivity and accuracy of the solution.

The main part. Released in 2017 by Google Brain, Transformer architecture played a key role in the development of natural language processing in the next 5 years, replacing solutions, based on recurrent neural networks (RNN) such as ELMo. This became possible due to the Encoder-Decoder architecture's massive optimization and parallelization: unlike the RNN, the Transformer`s network blocks are not sequential, which allows the processing of the entire input sequence simultaneously [1]. An advanced attention mechanism called self-attention along with standard encoder-decoder attention from sequence-to-sequence models also played an essential role. The self-attention mechanism can be described as the introduction of context from other relevant words into the current one.

One of the first and most popular variations of Transformer is the BERT architecture, published in 2018. This language model, which contains only encoder layers, is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right contexts in all layers [1]. After release, the pre-trained BERT model with various additional output layers demonstrated state-of-the-art results on eleven natural language processing tasks.

Large language models demonstrate ground-breaking results on the majority of NLP tasks, however, most of them contain hundreds of millions (BERT), billions (mT5), or even trillions (Switch Transformer) of parameters. This implies that a huge amount of computational resources are required to train or use such models. One possible solution to reduce the model size without significant loss in quality is knowledge distillation. DistilBERT, a distilled version of BERT, shows that it is possible to reduce the size of a BERT model by 40% while retaining 97% of its language understanding capabilities and being 60% faster [2].

Being a key element of Transformer architecture, the self-attention mechanism, though, has a quadratic computational complexity, which has a significant impact on the processing speed of long sequences. This problem can be partially solved by replacing the attention sublayer with a simpler one, such as a linear layer (MLP-Mixer) or Fourier Transform (FNet). FNet is an attention-free Transformer architecture, wherein we replace the self-attention sublayer of each encoder layer with a Fourier sublayer, which applies a 2D DFT to its (sequence length, hidden dimension) embedding input – one 1D DFT along the sequence dimension, and one 1D DFT along the hidden dimension [3]. FNet model underperforms BERT by 8% on General Language Understanding Evaluation (GLUE) benchmark while training 1.8 times faster on GPU. A hybrid FNet model, the last two layers of which use self-attention instead of Fourier transforms to retain 97% of BERT accuracy, was used in this work. The configurations of all used models are shown in Table 1.

Table 1 – Models configurations

Model	Dimension	Number of layers	Number of attention heads	Parameters (millions)
BERT-base-multilingual [1]	768	12	144	178
DistilBERT-base-multilingual [2]	768	6	72	134
FNet-Hybrid-base [3]	768	12	24	88
FNet-Hybrid-base-multilingual	768	12	24	164
FNet-Hybrid-base-ru-uk	768	12	24	95

Multilingual BERT and DistilBERT models have already been pre-trained by their authors on a text corpus consisting of more than 100 languages, but the existing pre-trained FNet model was trained on English only. For this reason, we built two versions of the model: multilingual and bilingual Russian-Ukrainian. They are pre-trained in an unsupervised fashion on a large corpus of 50 million texts in English, Ukrainian, and Russian languages. As well as for BERT, FNet models were trained simultaneously on two tasks (Figure 1): Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM the model randomly masks 15% of the words in the input, then run the entire masked sentence through the model and has to predict the masked words. In NSP the model concatenates two masked sentences and then predicts if the second sentence continues the first in meaning.

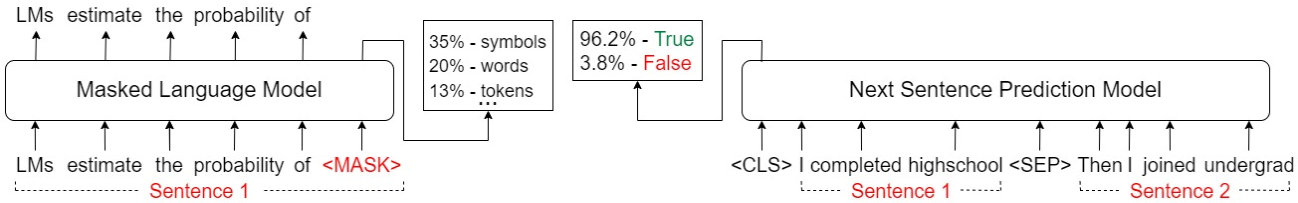

Figure 1 – Visualization of Masked Language Modeling and Next Sentence Prediction tasks

In turn, the model loss function is expressed as the sum of the loss functions for the MLM and NSP tasks (Formula 1):

$$L = L_{cross\_entropy}(\hat{y}_{mlm}, y_{mlm}) + L_{cross\_entropy}(\hat{y}_{nsp}, y_{nsp}) \quad (1)$$

where L is the loss function for the training procedure; $L_{cross\_entropy}$ – cross-entropy loss function; $\hat{y}_{mlm}$ – predicted tokens for the masked positions; y_{mlm} – exact tokens for the masked positions; $\hat{y}_{nsp}$ – two predicted values within [0, 1] for the NSP task; y_{nsp} – two exact values for the NSP task.

A comparison of results of pre-trained language models on the test dataset, which contains entries in Ukrainian and Russian languages, is demonstrated in Table 2.

Table 2 – The results of the models' evaluation

Model	Total loss	MLM loss	NSP loss	Exec. time for 1000 requests, sec
BERT-base-multilingual [1]	**2.05**	**1.69**	0.36	14.5
DistilBERT-base-multilingual [2]	3.66	2.97	0.69	10.3
FNet-Hybrid-base [3]	2.13 (en)	1.79 (en)	0.34 (en)	**8.9**
FNet-Hybrid-base-multilingual	2.09	1.81	**0.28**	9.9
FNet-Hybrid-base-ru-uk	3.90	3.54	0.36	9.3

Conclusions. In this work, the multilingual and bilingual language models for Ukrainian and Russian languages based on FNet architecture were introduced. After training for a million steps, the multilingual model achieves 98% of BERT`s language understanding capabilities and even slightly outperforms it on the NSP task, while being 32% faster. Those metrics generally correspond to the results achieved by the FNet`s authors for the English language model. Further work may include improving the quality of the bilingual model and evaluating language models on the localized GLUE benchmark.

REFERENCES

1. J. Devlin, C. Ming-Wei, L. Kenton, K. Toutanova, BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding (2018). arXiv:1810.04805v2.

2. V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2019). arXiv: 1910.01108v4.

3. J. Lee-Thorp, J. Ainslie, I. Eckstein, S. Ontanon, FNet: Mixing Tokens with Fourier Transforms (2021). arXiv:2105.03824