V. Korol, Student
Ye. Kanchura, PhD in Phil., As. Prof., research and language advisor
Zhytomyr State Polytechnic University

TRANSFORMER NEURAL NETWORKS

Transformer Neural Networks (TNNs) achieved the state of the art performance in natural language processing (NLP) problems [1, p. 1; 1, p. 9], including machine translation, question answering, sentimental and thematic detection thanks to the attention mechanisms [2]. This study researches the internals of TNNs and extracts the key insights about their algorithms, limitations and further progress.

The key innovation in the transformers is their algorithm of self-attention, which allows them to process speech as one whole, as opposed to the approach of recurrent neural networks that only analyze text sequentially [1, p. 1], and capture long-term dependencies and relationships between words. In the process of training, transformers break the input text into an array of tokens, which can be words and punctuation marks, and for each individual token, they learn a set of three vectors called query, key and value vectors [3, attention]. Transformers compose a matrix that consists of query vectors of each token and multiply the key vector of each lexical unit by the respective query matrix to get the relations between the individual word and the entire text. Afterwards, the result is passed to the softmax activation function that returns another vector that is multiplied by the value vector gaining the context vector that describes the word's meaning, peculiarities and properties [1, p. 3-4].

Upon processing the natural language input in this way, the self-attention stage exits handing the context vector of each token to a feed-forward neural network. It uses the classic multitron architecture to process the resultant vectors with weights and biases [1, p. 5-6]. The FFN layer introduces non-linearity in the model and enables it to perform other linguistic problems that cannot be expressed as linear transformations of input text, including information retrieval, sentiment analysis, machine translation [4, p. 70] and topic recognition.

Finally, for each context vector the FNN network outputs another non-linearly transformed vector called hidden state. When every token is passed through this step, all hidden states are joined together in a single large vector that contains the data about the entire text input. In the end, this final vector goes through another layer that maps the text to every word in the target language as a probability distribution where each output neuron stands for a token in the dictionary. It is then retrieved, and the model generates the first word [1, p. 3]. afterwards, it takes its own generated token and processes it in the same way as described above, appending it to the concatenated hidden states, keeping generating the next tokens until the end of sequence token is generated.

In conclusion, transformer neural networks achieved state-of-the-art performance in a wide range of NLP problems [1, p. 8-9]. The key innovation of their architecture is the self-attention algorithm that represents the token meaning and meta-data with a set of three vectors that allow the model to capture connection between words in language by performing linear algebra operations on them [1, p. 6]. Upon receiving the hidden states, the model maps predicts the most likely next token as a probability distribution in regard to the entire text, both the one it was supplied and the one it generates.

REFERENCES

1. Attention Is All You Need [Електронний ресурс] / [A. Vaswani, N. Shazeer, J. Uszkoreit та ін.]. – 2017. – Режим доступу до ресурсу: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

2. Olah C. Attention and Augmented Recurrent Neural Networks [Електронний ресурс] / C. Olah, S. Carter. – 2016. – Режим доступу до ресурсу: https://distill.pub/2016/augmented-rnns.

3. Rush A. The Annotated Transformer [Електронний ресурс] / Alexander Rush. – 2018. – Режим доступу до ресурсу: https://nlp.seas.harvard.edu/2018/04/03/attention.html.

4. OpenNMT: Open-Source Toolkit for Neural Machine Translation / [G. Klein, Y. Kim, Y. Deng та ін.]. // Association for Computational Linguistics. – 2017. – С. 67–72.