Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»

Кафедра иностранных языков технических факультетов

РЕФЕРАТ

по дисциплине «Иностранный язык»

Тема: Attention Is All You Need

Рецензия:_____

_____

Выполнил:

Студент: Чумаков И.В.

Группа: АТМ-25

_____
подпись

«_____»_____20___г.

Проверил:

Преподаватель: Ридная Ю.В.

Балл: _____ , ECTS_____

Оценка _____

_____
подпись

«____»_____20___г.

Новосибирск 2025

Текст реферата.

The paper is entitled "Attention Is All You Need". The article is written by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.

The paper under consideration is devoted to a new architecture for neural machine translation. The paper deals with the problem of sequential computation in standard models like RNNs and CNNs, which makes training slow. The paper is concerned with creating a faster and better model.

In recent years attention mechanisms have become very important in sequence modeling. However, they were still used with recurrent networks. The authors point out that recurrent computation is slow. Thus, the Transformer model, based only on attention, is proposed. The purpose is to create a model that is easy to parallelize and trains faster.

The paper begins with the discussion of Transformer model architecture. The first part concentrates on its encoder and decoder, which do not use RNNs or CNNs.

The mechanism of self-attention was investigated using a new method called "Scaled Dot-Product Attention". The structure of the model was studied by using multiple attention heads in parallel (Multi-Head Attention).

Then the way to give the model information about word order (Positional Encoding) is analyzed. Sine and cosine functions are used for this.

In addition, the problem of why self-attention is better than RNNs or CNNs is considered. It is shown that self-attention has shorter paths between words and is easier to parallelize.

Special attention is paid to the training process and results. Much attention is focused on testing the model on translation tasks.

The process of training was examined making use of large translation datasets (WMT 2014). The Adam optimizer and dropout for regularization were applied.

Finally, the parameters of different model versions were calculated. Experiments

with different numbers of heads and model sizes were done. Some information concerning the best configuration was obtained.

It was found that the big Transformer model got a new best score of 28.4 BLEU on English-German translation, beating all older models.

Besides, the paper touches upon the problem of using the Transformer for other tasks. An experiment on English grammar parsing is described. It was concluded that the model works very well even with little data.

In conclusion, the author emphasizes that the Transformer model is faster to train and gets better results than older models. To summarize, the author says that this is the first model using only attention, and it works for translation and other tasks.

All things considered we can come to the conclusion that the Transformer is a very important new model. To sum up, the paper under discussion provides us with a new way to build neural networks for language that is more efficient and powerful. These results are very significant for the future of machine learning.

Далее будет представлена оригинальная статья.