References

1. Bohdan Dmytrishin Ukranian IT without filters. – 24th channel, 2023.
2. Anastasia Zanuda How the Ukranian IT works in the war. – BBC Ukraine, 2022.
3. Stepan Mitish EPAM vice-president – EPAM SYSTEMS, 2022.
4. Nemchinsky S. Current state of IT market. – Foxminded Ltd, 2023.

NEURAL NETWORKS. GPT TECHNOLOGY. MIDJOURNEY

Mamochka E. I., student,
Gerasymchuk T. V., Associate Professor,
Kharkiv National University of Radio Electronics

A neural network can be called a program, which is based on the principle of the brain function.

A neural network is a type of machine learning in which a program works on the principle of the human brain. No one knows 100% exactly how the brain works, but it is believed that this is the most approximate but simplified version. The neural network itself consists of a combination of neurons – layers. There are incoming, hidden and outgoing layers.

We should note that neural networks can be:

1. Single-layer (perceptron) structure of a neural network. It is a structure of neuronal interaction in which signals from the input layer are immediately directed to the output layer, which, in fact, not only converts the signal, but also immediately outputs a response.

2. Multi-layer (Deep) Neural Network. Here, in addition to the output and input layers, there are several other hidden intermediate layers. The number of these layers depends on the degree of complexity of the neural network. It is more like the structure of a biological neural network.

In addition to the number of layers, neural networks can be classified according to the direction of information distribution along synapses between neurons, but first you need to understand what neurons and synapses are.

A neuron is a unit that performs calculations. It receives data from the input layer by performing simple calculations with it, and then transmits it to the next neuron.

A synapse is a connection between neurons, and each synapse has its own weight. This is why the input data is modified during transmission. During processing, the information transmitted by the synapse with a high weight indicator will become overwhelming.

That is, the result is influenced not by neurons, but specifically by synapses, which give a set of weights of input data, because the neurons themselves constantly perform exactly the same calculations. The scales are set in random order. Currently, neural networks are developing very rapidly, and such systems can be distinguished, for example, GPT, MidJourney.

MidJourney system is a system for generating images. The user provides the system with a text or image, based on which another image is generated in accordance with the request. The system consists of several neural networks already trained on large amounts of data, performing their own functions for generating, processing, and improving images.

If the user enters the text, the first neural network converts it into a vector (an array of data), then the diffusion model (used to convert embedding an image into an image) creates a small initial image, and then the convolutional neural network performs an improvement (magnification) of the image, which is already ready to be provided to the user. Four options are created for such images so that the user can choose the best option.

The process of generating an image into an image uses only a diffusion and magnification model. When entering an image and text, the system performs vectorization of the text and image separately (the image vector is larger and more important than the text vector) and then, thanks to the diffusion model, it is combined into a single whole. The improvement process is similar to other options.

The GPT system is a text generation system created by OpenAI. There is such a version of GPT as ChatGPT, based on the system version 3.5. The user enters a

request, and the system analyzes the text and creates a response based on it and its knowledge base. Briefly speaking about the principle of operation, we should note that when generating the text continuation using GPT the following happens:

1. the input text is tokenized into a sequence of numbers (tokens).

2.the token list passes through the Embedding Layer and turns into an embedding list (very similar to word2vec).

3. A positive embedding is added to each embedding.

Unlike recurrent networks, the transformer architecture is not sensitive to the order of input tokens, that is, even if you mix words in places, the output will still be the same (permutation invarience).

But in speech word order is very important! To take it into account, I had to come up with a crutch - positive encoding. This mechanism allows transformers to" see " the order of incoming tokens.

4.Next, the list of embeddings begins its journey through several identical blocks (Transformer Decoder Block).

5. after the list of embeddings passes through the last block, the embedding corresponding to the last token is Matrix multiplied by the same input, but already transposed Embedding Layer, and after applying SoftMax, we get the probability distribution of the next token.

6. from this distribution we select the next token (for example, using the argmax function).

7. then we add this token to the input text and repeat steps 1-6.

Currently, the development of such systems is bringing humanity closer to creating a strong artificial intelligence that will help us solve everyday problems and the development of humanity as a whole. Neural networks are our future!

Literature:

1. Neural networks by Simon Heikin.

2. Artificial Neural Networks: calculus-Novotarsky M. A., Nesterenko B. B.