Online Services Based on Neural Networks: a Review of Modern Implementations

V. D. Ilyin

***Abstract* – The review presents the results of an analysis of six online services based on neural networks (NN-services), which are designed to generate texts, images, videos and websites; search the Internet and answer questions. A brief overview of the functionality of the NN-services is provided. Based on the analysis, it is suggested to enhance statistical approaches in neural network technology development by incorporating methods for training neural networks and interpreting queries using S-symbolic models of concept systems. The proposed methods have been tested in automated parallel program design systems that utilize task knowledge systems.**

***Keywords* – online services; neural networks; online services based on neural networks; S-symbolic systems of concepts; S-symbolic systems of knowledge.**

I. INTRODUCTION

Researchers model the human brain as a neural network consisting of approximately 10^{11} neurons, each considered as a network node [1–3]. The arcs of this network (connections between neurons) and the associative memory of the brain are formed during the perception and processing of audio, visual, tactile, odorous and gustatory S-messages, cognition of oneself and the environment, and solving various S-problems [4–5]. The effectiveness of associative memory (capacity, speed of memorization, retention time, etc.) and the brain's parallel data processing varies based on learning, motivation, problem complexity, and the number of problems solved.

The brain's neural network, when solving S-problems, is often compared to a parallel data-processing computer. Note that the switching time of the neuron is about 10^{-3} seconds, and the transistor is about 10^{-9} seconds; computer memory is not associative, but addressable [3].

The ability of the neural network of the brain to improve itself as a result of learning has long attracted the attention of not only biologists. The first neural network model was created in 1943 by neurolinguist Walter Pitts and neuropsychologist Warren McCulloch [6]. They created and presented a mathematical model of a neuron (a structural and functional unit of the human nervous system). They created and presented a mathematical model of a neuron (a structural and functional unit of the human nervous system).

In 1958, Frank Rosenblatt developed the first neural network with reverse error propagation, called the «Perceptron» [7]. This model was the first one that could independently learn and solve classification problems, which was a significant step in the development of neural networks. Thus, the first implementation of the neural network model can be considered the «Perceptron» created by Frank Rosenblatt.

A. Writing and semantic markup of the text

TSM language with semantic markup is used to record formulas and highlight definitions, remarks and examples in S-messages [4].

To complete the descriptions, a standard keyboard and a set of tools available in the text editors of the LibreOffice, OpenOffice, or other office packages are sufficient.

The article uses the following TSM tools for semantic text markup:

□ <description fragment> □ definition (hereafter the symbol ≈ replaces the word "means");

◊ <description fragment> ◊ ≈ remark;

∘ <description fragment> ∘ ≈ example.

The first occurrences of the names of concepts and fragments of the description to which the author wants to draw attention are highlighted in italics (and bold).

B. About the results presented in the article

The study was conducted within the framework of the research project "Mathematical methods of data analysis and forecasting" (code: 0063-2019-0003, state registration No.: AAAAA-A19-119091990038-2), performed in accordance with the state assignment of the FANO of Russia for the Federal Research Center «Computer Sciences and Control» of the Russian Academy of Sciences.

II. THE STUDIED NN-SERVICES

There are three layers of interconnected artificial neurons in any basic neural network:

– *The input layer*. It is necessary to transfer information from the outside world to the neural network. The input nodes process the incoming data, analyze and classify it, and then transfer it to the next layer.

– *Hidden layers*. They receive data from the input or other hidden layers. Neural networks can have many hidden layers, each of which analyzes the output of the previous layer, processes it and transmits it to the next.

– *The output layer*. It outputs the final result of data processing. The output layer can have one node (for binary logic tasks) or multiple nodes (∘ for classification tasks, there may be quite a large number of nodes ∘).

Knowledge bases integrated with neural networks are usually built on the basis of *homogeneous semantic networks*, and knowledge processing in them is carried out using a *multi-agent approach*.

Direct propagation neural networks (*dpNNs*) are multi-layered networks consisting of several levels of active elements. The input level generates signals that are propagated to the connectors (synapses) of the first layer.

[1]The article was received on October 9, 2024.

Ilyin V. D. The Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences (e-mail: vdilyin@yandex.ru)

All odd layers of the network consist of connectors (synapses), and even layers consist of switches (*neurons*). The number of connectors and switches in each layer corresponds to the number of reference signals – vectors of the training sample.

An experimental study using the example of a six-layer dpNN has shown that a multilayer neural-like direct propagation network learns much easier than a recursive network trained by the method of error back propagation. At the same time, such a network is resistant to significant interference when distinguishing signals.

For example, based on a neural network of direct distribution, it is possible to build a model for predicting the volume of electricity consumption.

Recurrent neural networks (*RNNs*) with feedback are a class of neural networks that process input messages by storing data from previous inputs. This architecture effectively solves tasks that require understanding context over time (∘ when translating messages from one language to another ∘).

The mechanism of operation: RNNs accept an input sequence and process it one element at a time, updating the hidden state at each step. The hidden state is affected by both the current input and the previous hidden state, which allows the network to accumulate information in successive time steps.

Training: The back propagation in time (BPTT) method is usually used. This method expands the network in time steps of the input sequence, which allows you to apply a standard back-propagation algorithm.

One example of RNN with feedback is *Convolutional Neural Networks with Feedback* (*CNN-F*). This framework introduces generative feedback with latent variables into existing CNN-architectures, where consistent predictions are made through an alternative MAP discussion in a Bayesian environment. Experiments have shown that CNN-F significantly increases attack resistance compared to conventional unidirectional CNNs.

Another example is *Feedback Recurrent Neural Network* (*FBRNN*). FBRNN is used for extreme super resolution of images. With the help of a recurrent network, the image is improved in several stages from coarse to fine. In addition, at each stage, the improved image is reverse projected onto the original ones, which allows you to continuously refine the result.

Transformers are called neural networks in which the processing of data sequences is based on attention mechanisms. Transformers are able to efficiently process long sequences (for example, when processing messages in natural languages). Attention mechanisms determine the relationships between the elements of the sequence, highlighting the most significant ones for the current context.

Transformers consist of encoders and decoders interconnected through attention mechanisms. Encoders are designed to process and encode input data, and decoders are designed to generate an output sequence based on information received from encoders. [6–13].

A. Generative Pre-trained Transformer 4 (GPT-4)

The *GPT-4*[2] NN-service is designed for text generating, translating, searching for information on the Internet and answering questions. The GPT-4 neural network refers to *neural networks-transformers*.

Text generating. The *GPT-4* generates text based on the context of the dialog with the user.

Image analysis. The *GPT-4* can recognize photos and screenshots, understand their context, and work with image-based queries.

Search for information on the Internet. The *GPT-4* can generate answers and provide hyperlinks to sources.

Solving standard tasks. Searching for answers to questions, conducting a dialogue, summarising text, creating artistic texts, generating program code based on a text description (including creating websites and mobile applications).

◊ The *GPT-4* is more powerful than the previous version of the *GPS-3.5*: holds 96,000 words in the chat memory compared to 12,000 for the *GPT-3.5*. The *GPT-4* perceives the context better, recognizes nuances in user questions. Due to this, it gives more adequate answers. The *GPT-4* copes better with mathematical and coding tasks. It can generate code snippets and supports debugging existing code. ◊

B. Midjourney

The *Midjourney*[3] NN-service is designed for creating images based on text descriptions. With it, you can develop unique graphic design and visual content. The Midjourney neural network refers to *recurrent neural networks*.

The *Midjourney* neural network is trained on millions of different images and understands popular art styles and trends, and is able to imitate the works of famous artists, illustrators, designers and photographers. It can combine different styles in one image, so it can be used to create complex and creative works.

Users create images by sending commands to the bot in the *Discord* messenger.

C. DALL-E 3

The *DALL-E 3*[4] NN-service is designed for image generating on a text description. Its neural network is trained on more than 500 million images, which allows it to create images in different styles. The DAL-E 3 neural network belongs to *direct propagation neural networks*.

Some *DALL-E 3* features:

– Inserting text (corresponding to the context) into images.

– Auto-completion of prompts.

– The ability to combine different characters and objects in one image.

D. Pika 1.0

The *Pika 1.0*[5] NN-service is designed for video generation. It can convert text to video, images to video and video to video.

[2] GPT-4. https://openai.com/index/gpt-4/.

[3] Midjourney. https://midjourney.com.ru/#rec543942941

[4] DALL-E 3 https://dalle-3.ru/

[5] Pika 1.0 https://pika.art/home

Some *Pika 1.0* features:

– Working with 3D graphics. The *Pika 1.0* creates frames in the style of Japanese animation, Pixar cartoons and generates photorealistic videos.

– Editing dynamic objects. For example, you can ask to remove selected characters from the video and depict the missing scenery.

– Expansion of the staff. The *Pika 1.0* complements the video around the original frame when it is expanded.

– The *Pika 1.0* can rotate and change the aspect ratio. It can convert horizontal video to vertical (or vice versa) and change the aspect ratio of video frames.

E. Craftum AI

The *Craftum AI*[6] NN-service is designed for automated website generation. The *Craftum AI* generates designs, texts and images based on the description of the subject area. The Craftum AI neural network belongs to *recurrent neural networks*.

Some *Craftum AI* features:

– Fast generation of websites with a rational structure, relevant images and acceptable texts.

– A simple visual editor that allows you to customize site elements (texts, pictures, buttons, forms, etc.).

– Built-in tools for SEO optimization, hosting, SSL certificate, DDoS protection, integration with other services.

– Adaptability, rich library of templates and blocks.

To create a website, you need to select its type (∘ landing page, business card site, online store, etc. ∘), text style and enter basic information about the project. If necessary, the generated site can be edited in a visual editor. After connecting the domain, the site is ready for publication.

◊ There are no restrictions on the number of web pages per project, the amount of memory on HDD and traffic. ◊

F. YandexGPT

The *YandexGPT*[7] NN-service is designed for creating and processing texts, taking into account the context of a dialogue with a user. The YandexGPT neural network belongs to *recurrent neural networks.*

Some *YandexGPT* features:

– Creating text content, coming up with titles; generating annotations and announcements.

– Generating simple code for HTML pages.

– Creating article plans.

– Translating texts at a basic level.

◊ The first version, the *YandexGPT 3 Pro*, is designed for business and is available to Yandex Cloud users. According to the developers, on average, the *YandexGPT 3 Pro* responds better than the *YandexGPT 2* in 67% of cases, and when working with simple user queries, its results were 69% higher.

The third generation of the lightweight version of the *YandexGPT Lite* is useful in scenarios where response speed is important: for example, it can be used in chatbots, for spell checking or data analysis. ◊

[6] Craftum AI https://ai.craftum.com/

[7] YandexGPT https://hi-tech.mail.ru/review/109799-kak-polzovatsya-yandexgpt/#anchor171575593902175765

1) About the development of NN-service called Neuro

The markup part of the work on the *Neuro* service was not easy.

In this case, markup is the annotation of data, assigning labels or categories to them, indicating the quality and validity of the text response.

A fragment of a dialogue with Neuro. *Ilyin V. D.*: "What are your chances of success in the Turing test?" *Neuro*: "There is no clear answer to the question, what are the chances of a neural network to succeed in the Turing test. The Turing test has limitations. First, it focuses only on textual communication, which excludes other aspects of intelligence such as visual perception or motor skills. Secondly, passing the test successfully does not necessarily mean that the machine has true intelligence or consciousness.

However, according to a study by scientists from the University of California, San Diego, ChatGPT-4 successfully passed the Turing test, deceiving human subjects in 54% of cases. For comparison, participants correctly identified people in conversations only 67% of the time.

At the same time, it should be borne in mind that stylistic and socio-emotional factors can influence the test results."

III. MAIN RESULTS OF THE ANALYSIS

It is advisable to supplement statistical approaches to neural network training and query interpretation with *methods of message interpretation based on S-models of concept systems*. These methods have been tested in *technologies of automated program design based on knowledge systems about S-tasks* [14–15].

To define a new concept (not related to axiom concepts), it is necessary to describe it with an already existing S-system of concepts.

□ *The definition of the S-system of concepts* is a description of its *S–model*, accompanied by an indication of the scope of applicability.

< S-system *Sc* of concepts > ≈ < the set S^{Sc} of concepts on which the S-object being studied is defined >, < the family *rel* (S^{Sc}) of relationships defined on S^{Sc} >.

The description is presented in the form of an *S-message*, designed for interpretation by IT-(researchers and developers); presentation, preservation, dissemination, accumulation and search in the *S-environment* [4–5]. □

◊ The definition of the S-system of concepts must meet *the necessary requirements of constructiveness*:

– representation in the form of a pair < S-model of the system of concepts >, < definition of the area of S-applicability >;

– the S-system of concepts, considered defined, should not include concepts that do not have definitions (and at the same time do not relate to axiom concepts). ◊ □ *The definition of the area of S-applicability* is a description of the types:

– *the correspondent* (to whom the definition is addressed);

– *the goals in the process of achieving which the definition makes sense* (classes of problems in the study of which the definition may be useful);

– *the stage at which it is advisable to use the definition* (concept, solution methodology, etc.). □

Applicability may belong to a set of areas in which natural and/or inventive objects are studied.

IV. CONCLUSION

1. The NN-services reviewed are based on neural networks that employ statistical approaches for neural network training and query interpretation. This limits their applicability in scientific and technical fields, where a more accurate alignment with the conceptual frameworks of the subject areas is necessary.

2. It is advisable to supplement statistical approaches to neural network training and query interpretation with *methods of message interpretation based on S-models of concept systems*. These methods have been tested in *technologies of automated program design based on knowledge systems about S-tasks* [14–15]. There is reason to believe that this will increase the adequacy of neural network models.

REFERENCES

[1] Walter, W. G. 1953. The living brain. New York, Norton. 311 p.

[2] Freeman, W. J. 1986. W. G. Walter: The Living Brain. In: Palm G., Aertsen A. (eds). Brain Theory. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-70911-1_17.

[3] Kriesel, D. A. 2007. Brief Introduction to Neural Networks. 244 p. http://www.dkriesel.com.

[4] Ilyin, V. D. 2022. Symbolic Modeling (S-Modeling): an Introduction to Theory // Artificial Intelligence Trends in Systems, CSOC 2022 / Ed. R. Silhavy. – Lecture Notes in Networks and Systems ser. – Cham: Springer, 2022. Vol. 502. P. 585–591. DOI: 10.1007/978-3-031-09076-9_54.

[5] Ilyin, V. D. 2023. Teoriya S-simvolov: formalizatsiya znaniy ob S-zadachakh [Theory of S-symbols: Formalization of Knowledge about S-problems] Sistemy i Sredstva Informatiki [Systems and Means of Informatics], 33(1), 124–131.
DOI: 10.14357/08696527230212 .

[6] Neural Network Primitives Part 1 – McCulloch Pitts Neuron Model (1943). 2020. Deep Learning Basics. https://machinelearningknowledge.ai/mcculloch-pitts-neuron-model/.

[7] De La Cruz, R. 2024. Frank Rosenblatt's Perceptron, Birth of The Neural Network. https://readmedium.com/frank-rosenblatts-perceptron-19fcce9d627f.

[8] Haykin, S. 2008. Neural Networks and Learning Machines. 936 p. Pearson. ISBN: 978-0131471399.

[9] Zhang, X. J., Moore, J. M., Yan, G. et al. 2023. Universal structural patterns in sparse recurrent neural networks. Commun Phys 6, 243. DOI: 10.1038/s42005-023-01364-0.

[10] Wegner, S. A. 2024. Neural Networks. In: Mathematical Introduction to Data Science. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-662-69426-8_16.

[11] Yu, W., Guo, H., Xiao, J. et al. 2024. Physical neural networks with self-learning capabilities. Sci. China Phys. Mech. Astron. 67, 287501. DOI: 10.1007/s11433-024-2403-x.

[12] Ortega-Fernandez, I., Sestelo, M., and N. M. Villanueva. 2024. Explainable generalized additive neural networks with independent neural network training. Stat Comput 34, 6. DOI: 10.1007/s11222-023-10320-5.

[13] Pira, L., and C. Ferrie. 2024. On the interpretability of quantum neural networks. Quantum Mach. Intell. 6, 52. DOI: 10.1007/s42484-024-00191-y.

[14] Ilyin, V.D. 1995. A methodology for knowledge based engineering of parallel program systems. In: Forsyth, G.F., Ali, M. (eds.) Proceedings of the Eighth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Melbourne, Australia, Jun 06-08, 1995 (IEA/AIE-95). pp. 805–809. Gordon and Breach Publishers, Melbourne.

[15] Ilyin, A. V., and V. D. Ilyin. 2021. Updated Methodology for Task Knowledge Based Development of Parallel Programs. In: Silhavy, R., Silhavy, P., Prokopova, Z. (eds) Data Science and Intelligent Systems. CoMeSySo 2021. Lecture Notes in Networks and Systems, vol 231. Springer, Cham. DOI: 10.1007/978-3-030-90321-3_25.

V. D. Ilyin. Doctor of Science in technology, Professor, The Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences, 40 Vavilova str., Moscow,119333, Russia, e-mail: vdilyin@yandex.ru.