2023 IEEE 10th International Conference on Cyber Security and Cloud Computing (CSCloud)/2023 IEEE 9th International Conference on Edge Computing and Scalable Cloud (EdgeCom)

2023 IEEE 10th International Conference on Cyber Security and Cloud Computing (CSCloud)/2023 IEEE 9th International Conference on Edge Computing and Scalable Cloud (EdgeCom) | 979-8-3503-1246-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSCloud-EdgeCom58631.2023.00065

NLP Research Based on Transformer Model

Junjie Wu
School of Computer Science and Engineering,Hunan University of Science and technology
China
704597752@qq.com

Xueting Huang
School of Computer Science and Engineering,Hunan University of Science and technology
China
736706309@qq.com

Jingnian Liu
School of Computer Science and Engineering,Hunan University of Science and technology
China
JingnianL1999@163.com

Yingzi Huo
School of Computer Science and Engineering,Hunan University of Science and technology
China
yingzihuo@foxmail.com

Gaojing Yuan
School of Computer Science and Engineering,Hunan University of Science and technology
China
3035200762@qq.com

Ronglin Zhang
School of Computer Science and Engineering,Hunan University of Science and technology
China
15613501713@163.com

***Abstract*—Natural language processing technology is an important research area in artificial intelligence which occupies a pivotal position in deep learning. This paper describes in detail the research of NLP based on Transformer structure, thus showing its ultra-high performance and development prospects. Therefore, this article provides a detailed description of the research on NLP based on the Transformer structure, in order to demonstrate its ultra-high performance and development prospects.**

Keywords: Transformer, NLP, RNN, Transformer XL

I. INTRODUCTION

Today's society is a highly intelligent and informative society, and the popularization of artificial intelligence, such as intelligent transportation system and smart home systems, has brought many conveniences to humanity [1][2]. *Natural language processing* (NLP) technology is an artificial intelligence technology that takes human language as its research [3]. It has applications in many fields such as text classification, machine translation, automatic summarization, etc.; at the same time, it has a very wide research scope and is reused in many fields. In the past, *multilayer perceptions* (MLPs), *convolutional neural networks* (CNNs) [4], and *recurrent neural networks* (RNNs) were commonly used to achieve this goal in order to better process natural language. In 2017, the Transformer model started to be used with good results.

A. Introduction to NLP (natural language processing)

Natural language processing (NLP) is a branch of artificial intelligence and linguistics [5]. It mainly studies the meaning or laws contained in natural language, that is, analyzing and processing natural language to transform it into useful knowledge that can be recognized and processed by computers. This field explores the application and processing of natural language, which refers to any human language, such as Chinese, French, or English, which people use every day[6]. It is a language evolved from human society, but does not include formal languages such as Java, C++, and so on. Natural language has abstractness, generality, and fuzziness, so computers cannot fully simulate human thinking and reasoning processes. Moreover, these natural languages are abstract information symbols, which contain a lot of Semantic information, and human beings can easily understand the meaning of them. Computers can convert natural language into algorithms that can be recognized, interpreted, and even automatically executed by machines, either in numerical or textual form.

However, computers can only process numerical information and cannot directly understand human language, so it is necessary to perform numerical transformation on human language; In addition, natural language contains many complex relationships, which also creates many difficulties in the calculation process. In order to enable machines to better express human thoughts and intentions, it is necessary to further analyze and explain the results generated during the execution of computer statements; Therefore, the research on NLP provides direction for solving such problems. And through NLP learning, we can realize the interaction between people and computer systems using natural language. People can use computers in the most suitable language, without spending a lot of time and energy to study various kinds of computer language that are not natural and not suitable; People can also gain a deeper understanding of human language abilities, intelligence, and other mechanisms through this.

The natural language processing mechanism covers two processes, including *natural language understanding* (NLU) and *natural language generation* (NLG). Natural language understanding refers to the ability of computers to understand the meaning of natural language texts, while natural language generation refers to the ability to express the given intentions using natural language texts [6], thus crossing the communication barrier between humans and machines. Among them, the understanding and analysis of natural language is hierarchical. Many linguists divide this process into five levels, which can better reflect the

979-8-3503-1246-1/23/$31.00 ©2023 IEEE

DOI 10.1109/CSCloud-EdgeCom58631.2023.00065

composition of language itself. These five levels are respectively phonetic analysis, lexical analysis, syntactic analysis, semantic analysis and pragmatic analysis [7]. There are a total of six steps in natural language generation: content determination, text structure, sentence aggregation, grammaticalization, reference expression generation, and language implementation [8]; Through these six steps, all relevant words and identified phrases can be combined to form a structurally complete sentence.

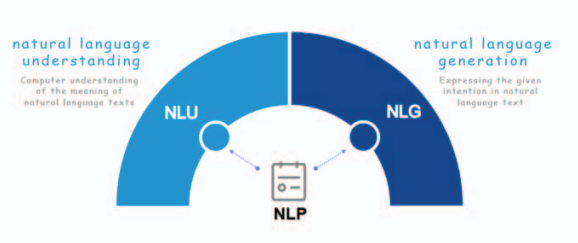

Figure 1 natural language processing Level

B. Introduction to the Transformer Model

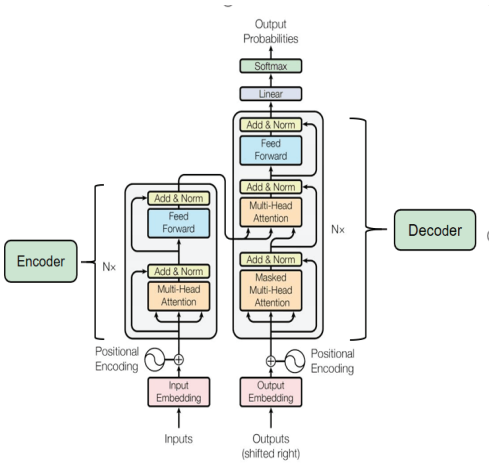

Figure 2 Transformer framework diagram

Transformer is essentially an end-to-end Seq2Seq structure, consisting of two parts: Encoder and Decoder [9]; The Encoder takes on the task of understanding the source text, while the Decoder is responsible for outputting translations. The structure of the model is shown in Figure 2. Encoder consists of N=6 identical layers, where layer refers to the unit on the left side of Figure 2, and there is also a 'Nx' on the far left, where there are x6. Each layer consists of two sub layers, namely multi head self-attention mechanism and fully connected feed forward network; Each sub layer has added residual connection and normalization. The construction of the Decoder is roughly the same as that of the Encoder, but with an additional sub layer for attention, which generally includes the following three modules:

1、 Masked Multi-head Attention Layer

2、 Enc-Dec Multi-head Attention Layer

3、 Feed-Forward Layer

It can be seen that the encoder and decoder blocks are actually composed of several identical encoder and decoder blocks stacked on top of each other, as well as the encoder stack, which has the same number of units as the decoder stack.

II. Related Research on Natural Language Processing and Its Current Status

A. Development of Natural Language Technology

In the field of natural language understanding, the earliest research work is machine translation. The design scheme of machine translation was first proposed by American Weaver in 1949; Subsequently, many people began to engage in exploratory work in this area. In the 1960s, foreign countries carried out large-scale research on machine translation, which cost a huge price. However, at that time, people obviously underestimated the complexity of natural language; Language processing is immature from both theoretical and technical perspectives, so progress is not significant. But since then, with more exploration and deep learning applications, natural language processing technology has also had new development. So far, NLP has achieved certain research results, and its development process can be roughly divided into the following three periods:

1. 1950s to 1970s - adopting rule-based methods

At this time, vocabulary, syntactic semantic analysis, question and answer, chat, machine translation and other systems will be constructed according to the rules. Its advantage is that rules can use human knowledge instead of relying on data and can start quickly [10]; However, the regularization method has unavoidable drawbacks. Firstly, rules cannot cover all sentences. Secondly, this method has a high demand for developers, who not only need to master computers but also linguistics. Therefore, although several simple problems have been solved at this stage, it cannot fundamentally make natural language understanding practical [11].

2. 1970s to early 21st century using statistical methods

Since the 1970s, in the context of the high development of the Internet, the realization of rich corpus, and the increasingly updated and improved hardware, the trend of natural language processing has shifted from empiricism to rationalism, and statistical based research methods have gradually replaced rule-based research methods [12]. At this stage, the research of natural language processing based on mathematical models and statistics has made substantive breakthroughs and has developed from laboratory to practical application[13][14][15] [12].

3. 2008 to present deep learning

Since 2008, inspired by research achievements in image recognition, speech recognition and other fields, people have gradually begun to introduce deep learning into the research of natural language processing, from the original word vector to the emergence of word2vec in 2013, pushing the combination of deep learning and natural language processing

to a climax, and achieving certain success in machine translation, question answering system and reading comprehension [17].

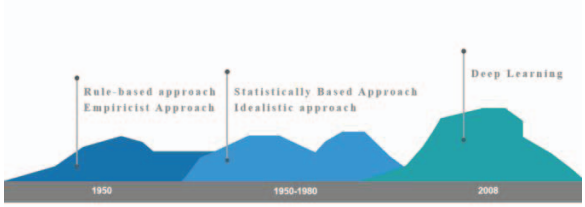

Figure 3 Development History of NLP

After deep learning has been widely use [18], the mainstream method used to model sequences in machine translation or language model tasks is to use recurrent neural networks (RNN). However, because RNN is difficult to process long sequences, it cannot be parallelized, and the training speed is slow, and because of the need to retain historical information at each step, the memory consumption is large, and it is expanded according to time, it cannot be parallelized. In order to solve the non-parallelizability problem caused by the cyclic characteristics of RNN, researchers have proposed various improved models, including the introduction of *convolutional neural networks* (CNN) and attention, but none of them have effectively solved the problem. Finally, after continuous experimentation and research and development, the application of Transformer began to emerge with good results.

B. NLP based on RNN processor

Among them, training based on neural network layers is a commonly used method. However, there is also a problem with the original RNN, which is the use of a linear sequence structure to continuously collect input information from front to back [19]. However, this linear sequence structure can be difficult to optimize during the backpropagation process, as the backpropagation path is too long, which can easily cause serious gradient disappearance or explosion [20].

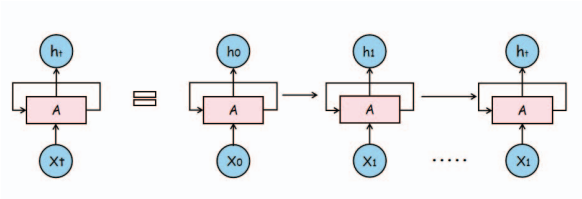

Figure 4 Structure of RNN

In addition, in traditional neural networks, the number of hidden layer neurons is fixed and cannot adapt to the requirements of significant differences between different types of data. In order to solve this problem, LSTM and GRU models emerged. They achieved the goal of directly transmitting information backwards by adding intermediate state information, and thus alleviated the problem of gradient vanishing. They achieved good results, so soon LSTM and GRU models became RNN standard models. However, these models with RNN as the core are difficult to have efficient parallel computing capabilities, which to some extent means that the upper limit value of this model is not as high as other non RNN series models, and both performance and computational efficiency lack competitiveness, destined to gradually withdraw from the historical stage.

C. NLP based on CNN processor

Before Transformer, CNN, as one of the most common deep learning models in natural language processing except RNN [21], was very popular in the image field, and its potential in the NLP field is self-evident. However, due to differences in task processing methods, CNN is commonly used for text data processing in the k-gram mode, similar to the sliding of convolutional kernels in image processing. The k-gram fragments obtained by convolution are features captured by CNN, and k determines how far they can be captured. This is clearly contradictory, ask cannot be set too large, which results in long-distance features that cannot be captured, and long-distance feature information is extremely important for NLP. Thanks to the development of computer power [22-24] and cloud computing [25-27], image processing techniques advanced rapidly [28,29]. We can borrow these techniques, such as increasing residual connections deepening network width and depth [30,31]. However, the cost is enormous and much greater than when processing images. Compared to RNN, its advantages are also quite prominent. Its strong parallel computing ability greatly reduces the cost of depth increase, and its huge success in image processing brings bright prospects for NLP.

D. NLP based on Transformer model

Transformer comes from the paper Attention is All You Need. The article itself is a natural language translation task, so the article calls it Transformer. We usually refer to Transformer, but we prefer to call the Substructure of Encoder or Decoder using Self Attention in the article Transformer. One of the major applications of Transformer is to replace RNN and CNN as model context sensitive feature extractors, which are widely used in NLP fields such as machine translation, reading comprehension, emotion analysis, dialogue recognition, etc. The biggest innovation of Transformer lies in directly abandoning the architecture of RNN and CNN [32] and fully utilizing attention mechanisms. It has strong semantic feature extraction ability, long-distance feature capture ability, task comprehensive feature extraction ability, parallel computing ability, and operational efficiency, and will be in the mainstream network architecture of NLP for a long time [33].

III. Application And Shortcomings Of Transformer Model In NLP

A. Application of Transformer Model in NLP and Case Implementation

a) machine translation

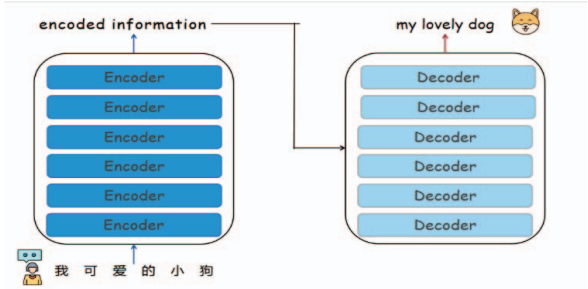

Figure 5 Machine Language Modeling

Machine translation (MT) is a technology that uses computers to translate between various languages. It is a process of automatically transforming two or more segments with the same grammar, semantics, and phonetic features into each other, with the aim of obtaining the desired translation for the target language user[34]. The language of translation is usually called source language, and the resulting language after translation is called target language. Machine translation, which is to transform source language into target language, is one of the most important research directions in natural language processing. Generally, the modeling method of machine translation is shown in the figure below. Given the source language, the model is expected to be translated into the target language.

b) Reading comprehension

Machine Reading Comprehension refers to having a machine read text and then answer questions related to the reading content[35][36]. With the rapid development of science and technology, the way people obtain information has undergone significant changes. Traditional information exchange can no longer meet the needs of social development, and machine reading has become one of the effective ways to solve this problem. Reading comprehension is one of the important frontier issues in the field of natural language processing and artificial intelligence. It is of great value for improving the level of machine intelligence and enabling machines to continuously acquire knowledge. In recent years, it has attracted extensive attention from academia and industry.

c) Emotional Analysis

Natural language, as a medium for human communication, can also express the emotions of interpersonal communication. Emotion plays a crucial role in language, reflecting the characteristics of relationships between individuals and different groups. A conversation or discussion can contain rich emotional colors, such as joy, hate, sadness, etc. In many cases, people express their emotional tendencies through language. The following figure shows an automatic analysis of the above emotional tendencies using a machine, which not only helps businesses understand consumers' perception of their products, but also lays the foundation for product improvement; It can also help businesses analyze the mindset of their business partners in order to make better business decisions. Emotional analysis is a data-driven intelligent technology that can extract effective information from massive amounts of text and make judgments. Usually, we tend to define emotional analysis tasks as classification problems, which involve using computers to determine whether the emotions expressed in a certain paragraph are positive or negative.

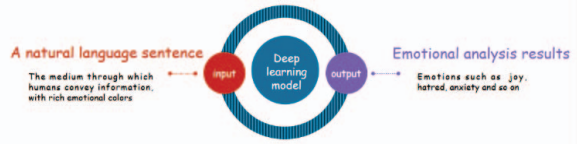

Figure 6 Emotional Analysis

d) Case analysis

There are many use cases of NLP, such as AI news editing, chat robots, automatic report generation, etc. Below, chat robots will be used as a case study for analysis. An important task in chatbots is response selection, which aims to select the most matching response from a set of candidate responses given the conversation context. In addition to playing a crucial role in retrieval-based chat robots, response selection models can also be used for automatic evaluation of dialogue generation and discriminators based on GAN (Generative Adversarial Network) neural dialogue generation. The contextual information in a dialogue system is important because the artificially generated response largely depends on the previous dialogue segments with different granularity (words, phrases, sentences, etc.) in semantics and scenarios. Therefore, as a powerful contextual information extractor, Transformer has broad applications in dialogue systems.

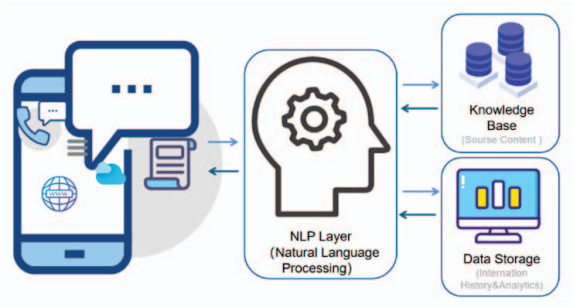

Figure 7 Working Mechanism of Chat Robot

B. Transformer's limitations and solutions

Undoubtedly, Transformer is a significant improvement for the seq2seq model based on recurrent neural networks. It utilizes the advantages of recurrent neural networks such as strong learning ability, associative memory, and generalization ability, and adopts a new training algorithm - dynamic threshold method. However, it also has certain limitations [37]. Firstly, its attention can only handle fixed lengths of text strings; Secondly, it cannot process a large amount of unstructured and semi-structured data at the same time. And before being put into the system, the text must be divided into a certain number of paragraphs or blocks. Secondly, when inputting consecutive characters, there may be misclassification caused by not knowing which paragraph or block the text belongs to. Such text blocking will cause fragmentation of the context; Secondly, after segmenting certain words, it is still necessary to re divide the sentence area. For example, if a sentence is separated from the middle, it loses a lot of contexts. Therefore, it is necessary to repartition the text. That is to say, in the case of text segmentation, the issue of semantic boundaries such as sentences is not considered. In addition, without considering the factors that affect the syntactic structure and meaning of words in different positions, the relationships between concepts cannot be well expressed, resulting in insufficient utilization of information. To overcome this deficiency, a new framework, Transformer XL, has been proposed.

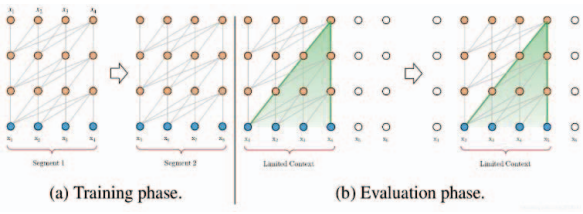

Figure 8 Schematic diagram of Vanilla Transformer

a) Vanilla Transformer

b) Transformer-XL

The Transformer XL architecture introduces two innovations based on the vanilla Transformer: Recurrence Mechanism and Relative Position Encoding [38]. This model can dynamically change the relationships between words based on user input, thereby achieving better grammar analysis results. Compared to Vanilla Transformer, Transformer XL also has an advantage that it can be used in both word level and character level language modeling [39].

V. CONCLUSION

In recent years, the Transformer model has been adopted in NLP research. It has been favored by more and more researchers due to its advantages such as simplicity and strong operability, and has become a new hot topic. This paper summarized and analyzes the development history and latest progress based on existing literature. The Transformer model has been applied to NLP, providing many conveniences for subsequent research. As more and more people devote themselves to the research of natural language processing, the Transformer model will continue to play an indispensable role in it. In the future, Transformer will overcome some existing limitations and make new progress, such as being able to process factory text data, combine multimodal information, and so on. Moreover, this model will still have broad application prospects in the future. Transformer is currently the most powerful NLP in terms of functionality, and it will be a great player leading the maturity of NLP technology.

VI. ACKNOWLEDGEMENT

This work was partially supported by 2021 Fujian Province Education and Research Project for Middle and Young Teachers (Science and Technology) Project Number: JAT210557IV.

REFERENCES

[1] Wei Liang, Yuhui Li, Jianlong Xu, ZhengQin, Dafang Zhang, Kuan-Ching Li. "QoS Prediction and Adversarial Attack Protection for Distributed Services Under DLaaS." IEEE Transactions on Computers, 167-172, 2021

[2] Y Yang, N Xiong, NY Chong, X Defago, A decentralized and adaptive flocking algorithm for autonomous mobile robots, 2008 The 3rd International Conference on Grid and Pervasive Computing, 2008.

[3] Li Huaxu. Review of research on natural language processing based on RNN and Transformer models [J]. Information Recording Materials, 2021,22 (12): 7-10

[4] Chunyan Diao, Dafang Zhang, Wei Liang*, Kuan-Ching Li, Yujie Hong, Jean-Luc Gaudiot. "A Novel Spatial-Temporal Multi-scale Alignment Graph Neural Network Security Model for Vehicles Prediction." IEEE Transactions on Intelligent Transportation Systems, 2021.

[5] Li Zhoujun, Fan Yu, Wu Xianjie. Review of pre training technology for natural language processing [J]. Computer Science, 2020,47 (03): 162-173

[6] Zhang Bo, Dong Ruihai. natural language processing technology enables the development of educational intelligence -- from the perspective of AI scientists [J]. Journal of East China Normal University (Education Science Edition), 2022,40 (09): 19-31

[7] Liu Haitao, Duan Jing, Wang Yanhua, Gu Wei, Yao Sibei. Application Analysis and Architecture Design of Digital Employees in the Power Industry Based on RPA+AI [J]. Power Information and Communication Technology, 2022-20 (04): 88-93, Lv Lucheng, Zhang Bo, Wang Yanpeng, Zhao Yajuan, Qian Li, Li Congcong. Quantitative Analysis of Global Patents in natural language processing [J]. Scientific Observation, 2021,16 (02): 84-95

[8] Xu Chengwei, Jia Xiaoxia, Ran Qingyun. Research on the Application of Intelligent Question Answering Systems in the Equipment Manufacturing Industry [J]. Manufacturing Automation, 2023,45 (03):38-43+75

[9] Chen,Hanting,et al."Pre-Trained lmage Processing Transformer." arXiv preprint arXiv:2012.00364(2020).

[10] Liu Dashan, Liu Luqi, Zhang Guangchi, Xue Chuanqi. Literature review of Attention mechanism based on deep learning [J]. Information Technology and Informatization, 2023 (01): 189-194

[11] Liu Jun, Wang Chunxiao, Dong Hongfei, An Ran, Gao Long. Exploration on the Application of natural language processing Technology in the Aviation Field [J]. Aviation Standardization and Quality, 2021, (02): 27-32+56

[12] Hu Kaibao, Shang Wenbo. Linguistics and Language Intelligence [J]. Journal of East China Normal University (Philosophy and Social Sciences Edition), 2022-54 (02): 103-109+176

[13] Wei Liang, Jing Long, Kuan-Ching Li, Jianlong Xu, Nanjun Ma, XiaLei. "A Fast Defogging Image Recognition Algorithm based on Bilateral Hybrid Filtering." ACM Transactions on Multimedia Computing, Communications, and Applications, 17(42): 1-16, 2020.

[14] Wei Liang, Jing Long, Kuan-Ching Li, Jianlong Xu, Nanjun Ma, XiaLei. "A Fast Defogging Image Recognition Algorithm based on Bilateral Hybrid Filtering." ACM Transactions on Multimedia Computing, Communications, and Applications, 17(42): 1-16, 2020.

[15] Wei Liang*, Dafang Zhang, XiaLei, Kuan-Ching Li*,Zomaya. "Circuit Copyright Blockchain: Blockchain-based Homomorphic Encryption for IP Circuit Protection." IEEE Transactions on Emerging Topics in Computing, 9(3): 1410-1420, 2020.

[16] Lv Lucheng, Zhang Bo, Wang Yanpeng, Zhao Yajuan, Qian Li, Li Congcong. Quantitative Analysis of Global Patents in natural language processing [J]. Scientific Observation, 2021,16 (02): 84-95

[17] Liu Cheng. Learning and research on text sentiment multi classification based on machine learning [J]. Computer Knowledge and Technology, 2020,16 (20): 181-182+186

[18] Wei Liang, Songyou Xie, Dafang Zhang, Xiong Li, and Kuan-Ching. "A Mutual Security Authentication Method for RFID-PUF Circuit based on Deep Learning." ACM Transactions on Internet Technology, 22(2): 1-20, 2021

[19] Ji Zhenyan, Kong Deyan, Liu Wei, Dong Wei, Sang Yanjuan. Research on Named Entity Recognition Based on Deep Learning [J]. Computer Integrated Manufacturing System, 2022-28 (06): 1603-1615

[20] Wu Lifen, Yan Xueyong, Zhao Jiji, Research on Machine Poetry of RNN Model and LSTM Model in IJ1 Technology Innovation and Application, 2021,11 (27): 48-50

[21] Yin Haoran, Miao Shihong, Han Xin, Wang Zixin, Mao Wandeng, Niu Rongze, Anomaly Identification Method for Distribution IoT Based on 3D Convolutional Neural Network. Power System Automation, 2022,46 (1): 42-50

[22] H. Huang, V. Chaturvedi, et al., "Throughput maximization for periodic real-time systems under the maximal temperature constraint", ACM Trans. on Embedded Computing Systems (TECS), 13 (2s), 1-22, 2014

[23] M. Qiu, M. Guo, et al., "Loop scheduling and bank type assignment for heterogeneous multi-bank memory", JPDC, 69 (6), 546-558, 2009

[24] M. Qiu, K. Zhang, M. Huang, "Usability in mobile interface browsing", Web Intelligence and Agent Systems J., 4 (1), 43-59, 2006

[25] K. Gai, M. Qiu, M. Liu, Z. Xiong, "In-memory big data analytics under space constraints using dynamic programming", FGCS, 83, 219-227, 2018

[26] Y. Song, Y. Li, L. Jia, M. Qiu, "Retraining strategy-based domain adaption network for intelligent fault diagnosis", IEEE TII, 16 (9), 6163-6171, 2019

[27] C. Li, M. Qiu, "Reinforcement Learning for Cyber-Physical Systems: with Cybersecurity Case Studies", Chapman and Hall/CRC, 2019

[28] M. Qiu, H. Qiu, "Review on image processing based adversarial example defenses in computer vision," IEEE 6th Conf. BigDataSecurity, 2020

[29] H. Qiu, Q. Zheng, et al., "Deep residual learning-based enhanced JPEG

compression in the Internet of Things", IEEE TII, 17 (3), 2124-2133, 2020

[30] H. Qiu, M. Qiu, R. Lu, "Secure V2X communication network based on intelligent PKI and edge computing", IEEE Network, 34 (2), 172-178, 2019

[31] H. Qiu, Y. Zeng, et al., "Deepsweep: An evaluation framework for mitigating DNN backdoor attacks using data augmentation", ACM Asia Conf. on Computer and Comm., 2021

[32] Liu Yangyang, Dong Tao. Research on the Structure of Chat Robots Based on Dialogue Models [J]. Information Technology and Informatization, 2023, (01): 13-16

[33] Zhang Feng, Chen Wei. ChatGPT and Higher Education: How Artificial Intelligence Drives Learning Change [J]. Journal of Chongqing University of Technology (Social Sciences): 1-12

[34] Wang Jiaqi, Zhu Junguo, Yu Zhengtao. Low resource machine translation based on gradient weight change training strategy [J]. Computer Science and Exploration: 1-10

[35] Yang Zhizhuo, Han Hui, Zhang Hu, Qian Yili, Li Ru. A Study on the Question and Answer of Chinese Reading Comprehension in the National College Entrance Examination by Integrating BERT Semantic Representation [J]. Journal of Chinese Information Technology, 2022, 36 (05): 59-66

[36] L. Shu, Y. Zhang, et al., Context-aware cross-layer optimized video streaming in wireless multimedia sensor networks, The Journal of Supercomputing 54 (1), 94-121, 2010.

[37] Li Qingge, Yang Xiaogang, Lu Ruitao, Wang Siyu, Xie Xueli, Zhang Tao. Overview of Transformer Development in Computer Vision [J]. Small Micro Computer Systems, 2023,44 (04): 850-861

[38] Song Yuhang. Research on Xinjiang Local Medicine Named Entity Recognition Based on Pre trained Models [D]. Supervisor: Tian Shengwei. Xinjiang University, 2021

[39] Zhang Chaoran, Qiu Hangping, Sun Yi, Wang Zhongwei. A Review of Machine Reading Comprehension Research Based on Pre trained Models [J]. Computer Engineering and Applications, 2020,56 (11): 17-25