DOI: 10.53469/jtpes.2024.04(05).02

AI Face Recognition and Processing Technology Based on GPU Computing

Huixiang Li[1*], Ang Li[2], Yuning Liu[3],Yiyu Lin[4], Yadong Shi[5]

[1*]Information Studies, Trine University, AZ, USA
[2]Business Analytics, University College Dublin, Dublin, lreland
[3]Business Analytics, Seattle University, Washington,USA
[4]Computer Science and Engineering, Santa Clara University, CA
[5]Computer Science, Fudan University, ShangHai, China
[*]*Corresponding author E-mail: lihuixiang95@gmail.com*

Abstract: ***In recent years, with the development of deep neural network technology, real-time object detection has become increasingly common in mobile applications. However, practical application requirements drive the algorithm to optimize in terms of speed, energy consumption and accuracy. This paper introduces the application of artificial intelligence in the field of face recognition, especially using TensorRT accelerated reasoning technology to improve the speed and performance of face recognition. At the same time, the paper also discusses the key role of GPU computing in face recognition, and expounds the importance of AI chips for optimizing inference tasks. Through the analysis of experimental results and methods, the performance advantages and application prospects of BlazeFace algorithm in mobile applications are demonstrated, which provides a valuable reference for industry.***

Keywords: Real-time object detection; Face recognition; GPU acceleration; TensorRT.

1. INTRODUCTION

In recent years, advancements in various architectures of deep neural networks have made real-time object detection feasible. While research laboratories may focus on developing algorithms that approach the limits of accuracy without constraints, practical applications demand considerations for response speed, energy consumption, and precision [1]. This necessitates algorithms with lower complexity that are suitable for hardware acceleration. In mobile applications, real-time object detection often serves as the first step in video processing workflows, followed by specific tasks such as segmentation, tracking, or geometric reasoning. Therefore, algorithms running inference on object detection models need to be as fast as possible and preferably outperform standard real-time benchmarks in performance.

Artificial Intelligence (AI) [2] aims to enhance the real-time capability of face recognition. Our team has upgraded the traditional neural network framework inference to a unified TensorRT accelerated inference to achieve this goal. Through experimental comparison, face recognition accelerated by TensorRT FP16 mode not only maintains almost lossless accuracy but also improves accuracy in some scenarios [3], with a recognition speed increase of 2.3 times compared to the original. Unified accelerated inference for face recognition not only provides customers with high-quality and responsive services but also enhances the utilization of AI server resources, reduces service costs, and facilitates the integration and unification of various model inferences.

This article delves into the intricacies of AI face recognition and processing technology, elucidating the pivotal role of GPU computing in driving innovation in this domain [4-5]. We explore the implementation of TensorRT accelerated inference as a solution to the challenges of real-time face recognition, highlighting its potential to revolutionize the landscape of facial recognition systems. Furthermore, we examine the broader implications of AI chips, also known as AI accelerators, in optimizing inference tasks and maximizing resource utilization, paving the way for more efficient and cost-effective AI solutions

2. RELATED WORK

AI chips, also known as AI accelerators or compute cards, are modules specifically designed to handle large-scale computing tasks in artificial intelligence applications (with non-computing tasks still handled by the CPU). Currently, AI chips mainly include GPUs, FPGAs, and ASICs [6-7].

2.1 Computer vision

Computer vision stands as a cornerstone of artificial intelligence (AI) and deep learning, representing the quest to endow machines with the remarkable ability to perceive and interpret visual information akin to human vision. At its essence, computer vision endeavors to unravel the intricate contents embedded within both static images and dynamic video streams, extracting invaluable insights and knowledge that can be harnessed to address an array of real-world challenges. Convolutional neural networks (CNNs) serve as the bedrock of modern computer vision, sophisticated architectures meticulously crafted to mirror the intricate processes of the human visual system. Through rigorous training, CNNs learn to emulate the nuanced functions of human vision, empowering machines to discern meaningful features and patterns from visual data.

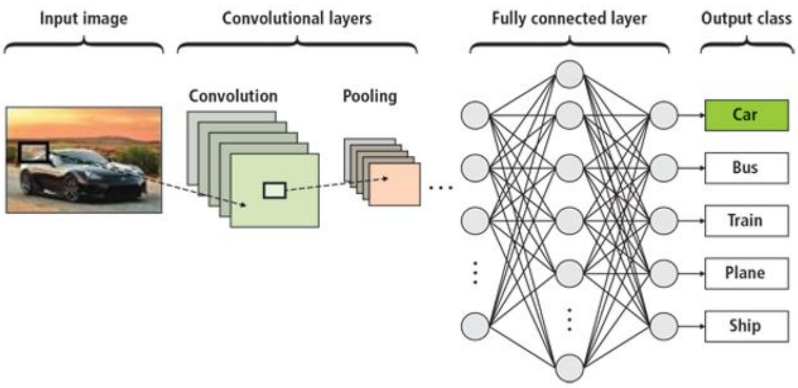

Figure 1: Computer vision architecture diagram

Convolutional neural networks (CNNs) are formidable assets in the realm of computer vision, proficiently handling segmentation, classification, and detection tasks across a spectrum of applications.

In segmentation, CNNs discern and categorize pixels into distinct classes like vehicles or pedestrians, pivotal in contexts such as autonomous driving systems.

For classification, CNNs are adept at discerning image content, enabling precise identification of objects like canines or felines.

In detection, CNNs locate objects within images, encapsulating them in bounding boxes, vital for surveillance or navigation endeavors. By transforming images into numerical representations and harnessing GPU-accelerated processing, CNNs achieve exceptional computational efficiency, enabling swift image analysis for real-time applications [8-9].

A fundamental tenet of computer vision lies in the meticulous training of CNNs to perform an array of tasks, including data segmentation, classification, and detection. Segmentation involves partitioning images into discrete regions to facilitate analysis and comprehension [10]. Classification entails the assignment of predefined labels or categories to images based on their content, enabling machines to recognize and differentiate between various objects, scenes, or concepts [11-12]. Detection, conversely, encompasses the localization and identification of specific objects within images or video frames, often delineated through bounding boxes or keypoint annotations. Through the iterative refinement of CNNs and the relentless exploration of innovative methodologies, computer vision continues to push the boundaries of what machines can perceive and comprehend. Each advancement in training techniques and model architectures heralds a new era of possibilities for computer vision, paving the way for groundbreaking innovations and applications across diverse domains, including healthcare, transportation, robotics, and beyond.

2.2 GPU computing combined with face recognition

Due to their composition of numerous identical neurons, neural networks inherently possess high levels of parallelism. This parallel nature seamlessly aligns with the architecture of GPUs, which excel in providing arithmetic capabilities for data parallelism, resulting in a significant boost in computational speed compared to

CPU-only training. This architectural advantage enables [13-14] GPUs to perform similar calculations on large sets of image data simultaneously, a capability particularly well-suited for computer vision tasks. NVIDIA GPUs, in particular, are renowned for accelerating computer vision operations, thereby relieving the CPU of processing burdens. Moreover, the ability to employ multiple GPUs within the same system allows for the parallel execution of multiple computer vision algorithms, enhancing overall processing efficiency.

GPU-accelerated deep learning frameworks offer robust programming interfaces for widely-used languages like Python. These frameworks empower developers to easily create and explore custom CNNs and DNNs while achieving the ultra-high speeds demanded by both experimental and industrial applications. NVIDIA CUDA-X AI further enhances the performance of popular deep learning frameworks like Caffe, TensorFlow, and Torch, enabling faster execution on GPUs and facilitating scalability across multiple GPUs within a single node. NVIDIA's cuDNN [15] and TensorRT™ frameworks optimize standard routines such as convolution layers, pooling layers, and activation layers, enhancing the efficiency of training and inference processes in convolutional neural networks.

For expedited development and deployment of visual models, NVIDIA provides the DeepStream SDK tailored to visual AI developers. This comprehensive toolkit includes the [16-17] TAO toolkit, empowering developers to create accurate and efficient AI models specifically designed for computer vision tasks.

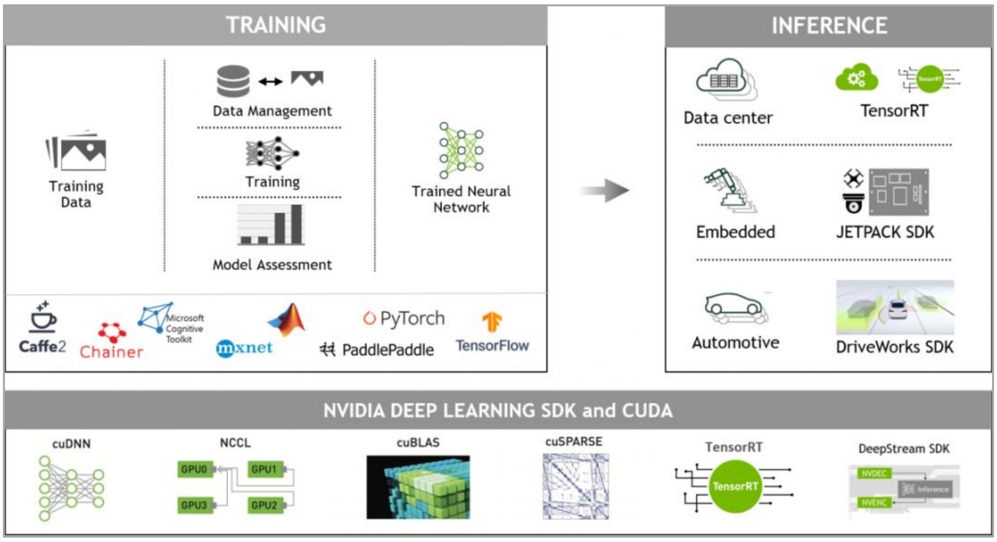

Figure 2: GPU rapid development and deployment of visual models

The performance of the latest version of GPU face recognition algorithm is superior. On the LFW face recognition database, the new algorithm of "DeepImage+" [18] visible light face recognition has achieved a target face recognition accuracy rate of 99.67%, which is ahead of the industry, surpassing the 97.35% announced by Facebook in June 2014. The response time of each face image comparison of the face recognition algorithm of the GPU image processor is 20 milliseconds, focusing on the application field of face recognition technology, and the latest GPU face recognition algorithm has completed the transformation of the algorithm to the application product.

2.3 Practical application of GPU computing

This practical demonstration showcases the utilization of PowerVR GPU hardware for visual processing, employing AI and machine learning algorithms to analyze video footage captured by cameras. By leveraging the computational power of GPUs, multiple convolutional neural networks (CNNs) are deployed to process the video data effectively. Following the application of these neural network algorithms, the screen displays the detected position of each face, along with an associated "identifier" to denote each individual. The networks were trained using the FDDB library, encompassing 5171 face images, and the VGG library, comprising 2622 images with diverse recognition features. The demonstration, exemplified in the accompanying video, underscores the seamless integration of the PowerVR Deep [19] Neural Network (DNN) library, developed collaboratively by the PowerVR Research Team and the Vision Team. This DNN library translates high-level instructions, weights, and biases into data processed by the GPU, enabling real-time execution of the network using the OpenCL library.

Demo shows a practical application case, face detection and recognition need to be integrated into our daily life. On the interface, it mimics a TV system, identifying users while also finding the movie content that suits them.

First, we use the GoogLeNet single lens detection (SSD) [20] neural network to recognize each face in the network camera capture, and each face will be marked with a square frame. We then isolate the face content and run another neural network in the same location, which returns us an "identifier," which is similar to the frame from which each face was captured, which also means that adding markers to the system is no longer important, From the above video, you can see how we identify each face image that has not been trained before, and how to use identifiers to directly identify the corresponding face image next time.

The following is the design block diagram of this Demo:

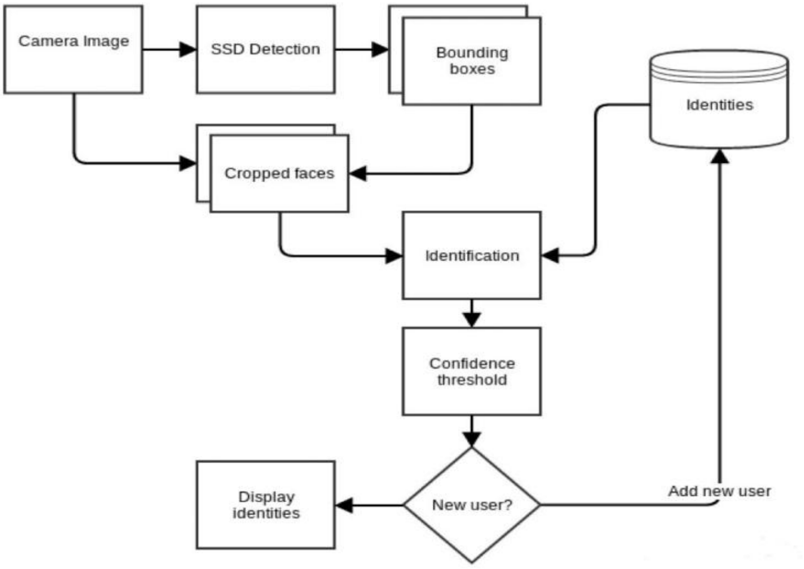

Figure 3: Design block diagram of Demo

The various uses of this technology, such as smart TV [21], itself already has the hardware foundation of running neural networks, users can use face recognition to register, and then the system will automatically load the user's favorite movie content, App applications and shortcuts, which will undoubtedly greatly improve the user experience.

Another potential use is to detect in real time whether the user is looking at the device, so that it can continuously track what the user is interested in and what the user's habits are, etc. This feature can be used for marketing or personalization without the user's subconscious attention.

Another example of an application is the smart doorbell, which determines each visitor through face recognition, and can customize the prompt tone for each user, and these similar products are already on the market, such as AI-based security camera systems.

This is just an image recognition Demo of how to use PowerVR Gpus efficiently and quickly, and I hope you will stay tuned to our blog as we continue to update on machine learning algorithms and PowerVR's advanced performance aspects [22].

TensorRT is a high-performance inference engine for deep learning inference, developed by NVIDIA. It provides fast and efficient inference solutions by optimizing and integrating deep learning models. TensorRT reduces, quantifies, and optimizes models while reasoning to improve reasoning speed and performance. By using TensorRT, deep learning models can be deployed on [23] Gpus for real-time, high-performance reasoning for a variety of AI application scenarios. TensorRT's accelerated reasoning technology has been widely used in face recognition, object detection, semantic segmentation and other fields, providing important support for real-time object detection and processing.

3. METHODOLOGY

A large number of algorithms are proposed in the field of face detection every year, and the spelling accuracy is of course very important, but really to achieve practical applications, the algorithm must also be fast. In algorithm design, the pursuit of low complexity and suitable for hardware acceleration (such as suitable for GPU computing, etc.) are the two main directions of algorithm acceleration. Recently, BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs introduced the BlazeFace algorithm, a lightweight, high-performance face detector tailored for mobile GPU inference [24]. BlazeFace runs at 200-1000 + FPS on flagship mobile devices. This ultra-real-time performance enables it to be used in any augmented reality application where performance is extremely demanding.

Main innovations of the algorithm:

(1) Extremely lightweight feature extraction network, inspired by but different from MobileNet V1/V2;

(2) Modified anchor mechanism for SSD target detection to make it more suitable for GPU computation;

(3) Replace non-maximum suppression (NMS) with the tie resolution policy.

In summary, under the framework of MobileNet-SSD [25] object detection, the author improved the network structure, anchor mechanism, and replaced the NMS post processing, which made the algorithm maintain high accuracy in face detection tasks and fast speed on mobile GPU.

3.1 Experimental model

The design of BlazeFace model architecture mainly considers four aspects.

Increase the receptive field.

In the MobileNet architecture, 5*5 convolution kernel is used instead of 3*3 convolution kernel to enlarge the receptive field, and the increase in computation caused by the increase in the size of convolution kernel in the depth separable convolution is limited.

In addition, in order to promote the transmission of receptive field Size, a double BlazeBlock module is proposed, as shown in the figure below:

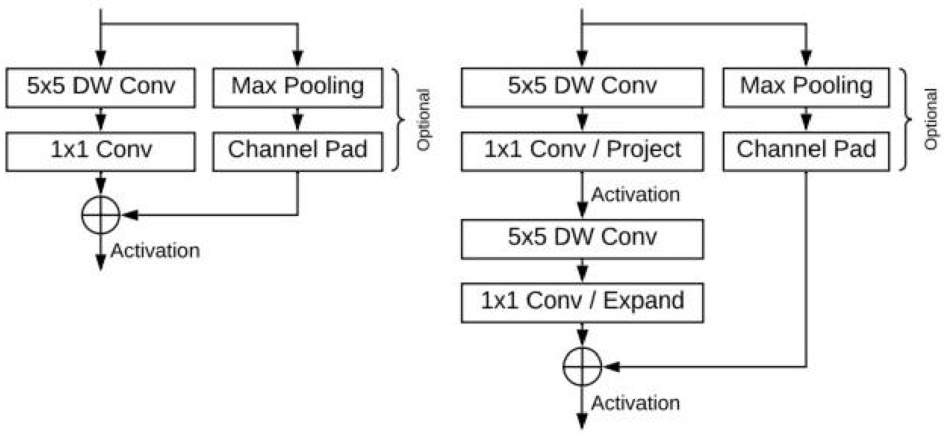

Figure 4: BlazeBlock (left) and double BlazeBlock

3.2 Feature extraction

For the specific needs of face detection in the front camera, the face scale changes less, and a more lightweight feature extraction is defined. The input image is 128*128, containing 5 blazeblocks and 6 double Blazeblocks. The network architecture is as follows:

Table 1: BlazeFace feature extraction network architecture

Layer/block	Input size	Conv. kernel sizes
Convolution	128×128×3	128×128×3×24
Single BlazeBlock	64×64×24	5×5×24×1 1×1×24×24
Single BlazeBlock	64×64×24	5×5×24×1 1×1×24×24
Single BlazeBlock	64×64×24	5×5×24×1 (stride 2) 1×1×24×48
Single BlazeBlock	32×32×48	5×5×48×1 1×1×48×48
Single BlazeBlock	32×32×48	5×5×48×1 1×1×48×48
Double BlazeBlock	32×32×48	5×5×48×1 (stride 2) 1×1×48×24 5×5×24×1 1×1×24×96
Double BlazeBlock	16×16×96	5×5×96×1 1×1×96×24 5×5×24×1 1×1×24×96
Double BlazeBlock	16×16×96	5×5×96×1 1×1×96×24 5×5×24×1 1×1×24×96
Double BlazeBlock	16×16×96	5×5×96×1 (stride 2) 1×1×96×24 5×5×24×1 1×1×24×96
Double BlazeBlock	8×8×96	5×5×96×1 1×1×96×24 5×5×24×1 1×1×24×96
Double BlazeBlock	8×8×96	5×5×96×1 1×1×96×24 5×5×24×1 1×1×24×96

(1) Improved anchor mechanism.

Stopping at the 8×8 feature map dimensions without further downsampling (Figure 2 below), replace the 2 anchors per pixel in the 8×8,4×4 and 2×2 resolutions with 6 anchors in 8×8. Because there is limited variation in the face aspect ratio, the authors found that fixing anchor to a 1:1 aspect ratio was sufficient for accurate face detection.

(2) Post-processing mechanism

Since the feature extractor in the anchor mechanism above does not reduce the resolution below 8×8, the number of anchors that overlap a given object will decrease. In the [26] NMS of [27] SSD, only one winning anchor is used for algorithm output, which causes the face frame to shake significantly when detecting in video.

To reduce this effect, the authors no longer use NMS, a blending strategy, to estimate the regression parameters of the bounding box as the weighted average between overlapping predictions. It incurs almost no cost of the previous NMS part. The authors report that for facial detection tasks in video, this adjustment resulted in a 10 percent improvement in accuracy.

3.3 Experimental result

The experiment focuses on verifying the acceleration effect of the detection algorithm in mobile terminal applications. The accuracy of the detection algorithm is compared with that of MobileNetV2-SSD instead of the most advanced algorithm in the public database. The results showed that the accuracy was better than the MobileNetV2-SSD [28-29], and the speed dropped from 2.1 milliseconds to 0.6 milliseconds on the iPhone XS. Speed comparisons on different phones also show that BlazeFace gets a significant speed boost. The authors note that BlazeFace's output contains six points that can be used for face correction and, due to its rapidity, can be used in combination with other fast face alignment algorithms. Finally, the authors point out that BlazeFace is already being used by Google in its actual AR [30] self-expression application and mobile AR developer API, so this article is of great reference value to industry peers.

4. CONCLUSION

By using TensorRT accelerated reasoning technology, this paper successfully improves the speed and performance of face recognition, and provides an effective solution for the application of real-time object detection in mobile applications. The experimental results show that BlazeFace algorithm has a significant performance advantage in mobile applications. It has excellent performance in both accuracy and speed, and has higher efficiency and faster response speed than traditional methods. This is of great significance for scenarios that require fast and accurate face recognition in practical applications, especially in the field of mobile AR applications and intelligent security.

In addition, the key role of GPU computing and the application of AI chips also provide strong support for the development of face recognition technology. By making full use of the parallel computing capabilities of Gpus, combined with accelerated reasoning technologies such as TensorRT, efficient face recognition algorithm deployment can be achieved to provide users with faster and more accurate services. In the future, with the continuous development and popularization of AI chip technology, face recognition technology will usher in a wider range of application scenarios, and have a positive impact on social life and industrial development.

5. ACKNOWLEDGEMENT

We would like to thank the article "Leveraging Federated Learning and Edge Computing for Recommendation Systems within Cloud Computing Networks" For their in-depth research and contributions to the fields of recommendation systems, joint learning, and edge computing. Special thanks to Yuan Feng, Hanzhe Li and Jingxiao Tian for their dedication and efforts in this study. Their work provides us with valuable reference and inspiration, and has a profound impact on our research work.

We would like to express our special thanks to their articles for their inspiration and guidance to our research. Through their insightful discussion of joint learning and edge computing, we have benefited from deepening our understanding of these areas. Their methods and theoretical framework provide valuable ideas and directions for our research, and provide a solid foundation for our exploration in the optimization of recommendation systems. We would also like to thank them for their in-depth research on the application of recommendation systems in cloud computing networks. Their work provides us with a new perspective and way of thinking, prompting us to further explore the optimization strategies and implementation methods of recommendation systems in complex network environments.

REFERENCES

[1] Qi, Y., Wang, X., Li, H., & Tian, J. (2024). Leveraging Federated Learning and Edge Computing for Recommendation Systems within Cloud Computing Networks. arXiv preprint arXiv:2403.03165.
[2] Xu, J., Wu, B., Huang, J., Gong, Y., Zhang, Y., & Liu, B. (2024). Practical Applications of Advanced Cloud Services and Generative AI Systems in Medical Image Analysis. arXiv preprint arXiv:2403.17549.
[3] Che, C., Zheng, H., Huang, Z., Jiang, W., & Liu, B. (2024). Intelligent Robotic Control System Based on Computer Vision Technology. arXiv preprint arXiv:2404.01116.
[4] Li, Lianwei, et al. "Independent Grouped Information Expert Model: A Personalized Recommendation Algorithm Based on Deep Learning." (2024).
[5] Zhang, Y., Liu, B., Gong, Y., Huang, J., Xu, J., & Wan, W. (2024). Application of Machine Learning Optimization in Cloud Computing Resource Scheduling and Management. arXiv preprint arXiv:2402.17216.
[6] Gong, Y., Huang, J., Liu, B., Xu, J., Wu, B., & Zhang, Y. (2024). Dynamic Resource Allocation for Virtual Machine Migration Optimization using Machine Learning. arXiv preprint arXiv:2403.13619.
[7] Liu, B. (2023). Based on intelligent advertising recommendation and abnormal advertising monitoring system in the field of machine learning. International Journal of Computer Science and Information Technology, 1(1), 17-23.
[8] Yu, Liqiang, et al. "Stochastic Analysis of Touch-Tone Frequency Recognition in Two-Way Radio Systems for Dialed Telephone Number Identification." arXiv preprint arXiv:2403.15418 (2024).
[9] Che, C., Lin, Q., Zhao, X., Huang, J., & Yu, L. (2023, September). Enhancing Multimodal Understanding with CLIP-Based Image-to-Text Transformation. In Proceedings of the 2023 6th International Conference on Big Data Technologies (pp. 414-418).
[10] Zheng, Haotian, et al. "Medication Recommendation System Based on Natural Language Processing for Patient Emotion Analysis." Academic Journal of Science and Technology 10.1 (2024): 62-68.

[11] Che, C., Hu, H., Zhao, X., Li, S., & Lin, Q. (2023). Advancing Cancer Document Classification with R andom Forest. Academic Journal of Science and Technology, 8(1), 278-280.
[12] Yu, Liqiang, et al. "Similarity Matching for Patent Documents Using Ensemble BERT-Related Model and Novel Text Processing Method." Journal of Advances in Information Technology 15.3 (2024).
[13] Zhang, X., Wang, C., Chen, B., Wang, Q., Xu, W., Ye, S., ... & Zhang, R. (2019). Crystal structure of refolding fusion core of Lassa virus GP2 and design of Lassa virus fusion inhibitors. Frontiers in microbiology, 10, 1829.
[14] Chen, B., Basak, S., Chen, P., Zhang, C., Perry, K., Tian, S., ... & Jin, R. (2022). Structure and conformational dynamics of Clostridioides difficile toxin A. Life science alliance, 5(6).
[15] Xu, Kangming, et al. "Intelligent Classification and Personalized Recommendation of E-commerce Products Based on Machine Learning." arXiv preprint arXiv:2403.19345(2024).
[16] Shen, Xinyu, et al. "Biology-based AI Predicts T-cell Receptor Antigen Binding Specificity." Academic Journal of Science and Technology 10.1 (2024): 23-27.
[17] Qian, Kun, et al. "Implementation of Artificial Intelligence in Investment Decision-making in the Chinese A-share Market." Journal of Economic Theory and Business Management 1.2 (2024): 36-42.
[18] Ding, W., Tan, H., Zhou, H., Li, Z., & Fan, C. Immediate Traffic Flow Monitoring and Management Based on Multimodal Data in Cloud Computing.
[19] Chen, B., Zhu, Y., Ye, S., & Zhang, R. (2018). Structure of the DNA-binding domain of human myelin-gene regulatory factor reveals its potential protein-DNA recognition mode. Journal of structural biology, 203(2), 170-178.
[20] Srivastava, S., Huang, C., Fan, W., & Yao, Z. (2023). Instance Needs More Care: Rewriting Prompts for Instances Yields Better Zero-Shot Performance. arXiv preprint arXiv:2310.02107.
[21] Fan, Chao, et al. "Integrating Artificial Intelligence with SLAM Technology for Robotic Navigation and Localization in Unknown Environments."
[22] Bao, Q., Wei, K., Xu, J., & Jiang, W. (2024). Application of Deep Learning in Financial Credit Card Fraud Detection. Journal of Economic Theory and Business Management, 1(2), 51-57.
[23] Chen, B., Liu, Z., Perry, K., & Jin, R. (2022). Structure of the glucosyltransferase domain of TcdA in complex with RhoA provides insights into substrate recognition. Scientific reports, 12(1), 9028.
[24] Biology-based AI Predicts T-cell Receptor Antigen Binding Specificity. (2024). Academic Journal of Science and Technology, 10(1), 23-27. https://doi.org/10.54097/wy28c490
[25] Wu, Binbin, et al. "Enterprise Digital Intelligent Remote Control System Based on Industrial Internet of Things." (2024).
[26] Choudhury, M., Li, G., Li, J., Zhao, K., Dong, M., & Harfoush, K. (2021, September). Power Efficiency in Communication Networks with Power-Proportional Devices. In 2021 IEEE Symposium on Computers and Communications (ISCC) (pp. 1-6). IEEE.
[27] Pan, Y., Wu, B., Zheng, H., Zong, Y., & Wang, C. (2024, March). THE APPLICATION OF SOCIAL MEDIA SENTIMENT ANALYSIS BASED ON NATURAL LANGUAGE PROCESSING TO CHARITY. In The 11th International scientific and practical conference “Advanced technologies for the implementation of educational initiatives”(March 19–22, 2024) Boston, USA. International Science Group. 2024. 254 p. (p. 216).
[28] Li, X., Zheng, H., Chen, J., Zong, Y., & Yu, L. (2024). User Interaction Interface Design and Innovation Based on Artificial Intelligence Technology. Journal of Theory and Practice of Engineering Science, 4(03), 1-8.
[29] Li, L., Xu, K., Zhou, H., & Wang, Y. (2024). Independent Grouped Information Expert Model: A Personalized Recommendation Algorithm Based on Deep Learning.
[30] Chen, Baohua, Kay Perry, and Rongsheng Jin. "Neutralizing epitopes on Clostridioides difficile toxin A revealed by the structures of two camelid VHH antibodies." Frontiers in Immunology 13 (2022): 978858.