Zifei Liu

800 Dongchuan Road, Minhang District, Shanghai, China | zejianwang@sjtu.edu.cn | (+86) 157-3500-3219 |

EDUCATION

Shanghai Jiao Tong University, Information Security Sept 2020 – Jul 2024

- GPA: 3.52/4.0 (Transcript)
- **Core Courses:** Core courses: Software Engineering (4.0/4.0) Project Management and Software Design (4.0/4.0), Data Mining (4.0/4.0), Comprehensive Information Security Practices (4.0/4.0), Information Security Science and Technology Innovation (4.0/4.0), Computer Composition and System Architecture (4.0/4.0)

PUBLICATIONS

Automatic Pipeline Parallelism: A Parallel Inference Framework for Deep Learning Applications in 6G Mobile Communication Systems Jul 2023

Shi, H., Zheng, W., ***Liu, Z.***, Ma, R., Guan, H

Journal: IEEE Journal on Selected Areas in Communications

A Combined Multi-Classification Network Intrusion Detection System Based on Feature Selection and Neural Network Improvement Jul 2023

Wang, Y., ***Liu, Z. (co-first author)***, Zheng, W.,Wang, J., Shi, H.,Gu, M.

Journal: Applied Sciences

A Federated Network Intrusion Detection System with Multi-Branch Network and Vertical Blocking Aggregation Sep 2023

Wang, Y.,Zheng, W., ***Liu, Z. (second author)***, Wang, J., Shi, H.,Gu, M.,Di, Y.

Journal: Electronics

EXPERIENCE

The 25th phase of the "Shanghai Jiao Tong University Student Innovation Practice Program", SJTU, Research Assistant Apr 2022 – Apr 2023

Advisor: Prof. Ruhui Ma, School of Electronic Information and Electrical Engineering, SJTU

- **Motivation**: Proposed a new parallel inference framework for deep learning applications, aimed at enhancing model inference efficiency while ensuring reliability.
- **Content**: (i) Modeled the main part of the parallel inference arrangement optimizer problem to determine device for each task;

  (ii) Designed machine learning algorithms and evaluation metrics for task segmentation/ scheduling/ merging/ adjusting;

  (iii) Optimized the main code.

Undergraduate Summer Internship, SJTU, Intern and Research Assistant Apr 2022 – Apr 2023

Advisor: Prof. Lihong Yao, School of Cyber Science and Engineering, SJTU

- **Motivation**: Developed secure compile option detection software for GCC to mitigate potential attacks within the Linux system.
- **Content**: (i) Collaborated with team members to explore available GCC compilers and compilation options;

  (ii) Configurated NX(DEP)/ RELRO/ PIE(ASLR)/ CANARY/ FORTIFY secure compilation options to enhance code security in software development;

  (iii) Conducted experiments on the impact of different secure compilation options on code performance and security.

AWARDS

Zhiyuan Honor Scholarship in Shanghai Jiao Tong University (7%)	2020
“Puyuan Future” Scholarship in Shanghai Jiao Tong University (1/20)	2021
SJTU-HUAWEI “Intelligent Base” Scholarship in Shanghai Jiao Tong University (1/20)	2022
SJTU-HUAWEI “Intelligent Base” Scholarship in Shanghai Jiao Tong University (1/20)	2022
President’s Award in Shanghai Jiao Tong University	2023
“High Performance in Network Security” Scholarship in Shanghai Jiao Tong University (1/20)	2023

SKILLS

Programming Languages: C++, C, python, JavaScript

Algorithms and Frameworks: Well-versed in SVM, KNN, K-Means and GB. Familiarity with GANs and the Transformer framework, with thorough study of reinforcement learning methodologies.

Language: Chinese (Native), English (Proficient,IELTS(6.5))

上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

本科生成绩单

电子信息与电气工程学院　专业：信息安全　班级：F2003602　学号：520021910590　姓名：刘紫菲

2020-2021学年

代码	课程名称	学期	学分	成绩	代码	课程名称	学期	学分	成绩
CHEM1211H	大学化学（荣誉）	1	3	85	EE0502	电路实验	2	2.0	85
CHEM1302	大学化学实验	1	1	87	EE1503	工程实践与科技创新I	2	2.0	91
CS1501H	程序设计思想与方法（荣誉）（C++）	1	3	83	EN063	大学英语（3）	2	3.0	75
					KE1202	体育（2）	2	1.0	90
FL2201	大学英语（2）	1	3.0	70	MARX1202	中国近现代史纲要	2	3.0	91
KE1201	体育（1）	1	1.0	92	MARX1205	形势与政策	2	0.5	92.00
MARX1201	思想道德修养与法律基础	1	3.0	92	MARX1206	新时代社会认知实践	2	2.0	P
MARX1205	形势与政策	1	0.5	90	MATH1202	高等数学II	2	4.0	81
MATH1205H	线性代数（荣誉）	1	5	81	ME1221	工程学导论	2	3.0	85.50
MATH1607H	数学分析（荣誉）I	1	6.0	69	PHY1221	大学物理实验（1）	2	1.0	87
MIL1201	军事理论	1	2.0	85	PHY1251	大学物理(A类）（1）	2	4.0	67
PSY1201	大学生心理健康	1	1.0	89	PU004	现代心理学	2	2.0	94
SI1210	工程实践	1	3.0	83	TY001	“UTJS”体验式教育:大学生演讲与沟通训练	2	2.0	93
AM016	网络环境下的文科信息检索	2	2.0	W					
EE0501	电路理论	2	4.0	80					

2021-2022学年

代码	课程名称	学期	学分	成绩	代码	课程名称	学期	学分	成绩
CH944	且共从容：唐宋词讲读	1	2.0	90	GA903	西方风景园林艺术史	2	2.0	86
CS0501	数据结构	1	3.0	91	ICE2501	信号与系统（B类）	2	3.0	64
CS2501	离散数学	1	3.0	85	JC903	美学	2	2.0	88
EST2501	数字电子技术	1	2.0	81	KE2202	体育（4）	2	1.0	96
EST2502	模拟电子技术	1	2.0	87	MARX1203	毛泽东思想和中国特色社会主义理论体系概论	2	3.0	95
EST2503	电子技术实验	1	2.0	91					
IP021	暑期科研见习岗位	1	1	P	MARX1205	形势与政策	2	0.5	87.50
KE2201	体育（3）	1	1.0	92	MECH2508	理论力学	2	4.0	P
MARX1204	马克思主义基本原理	1	3.0	90	NIS1335	网络信息安全概论	2	2.0	81
MARX1205	形势与政策	1	0.5	89.80	NIS2312	信息安全的数学基础（1）	2	3.0	75.85
MATH1207	概率统计	1	3.0	72	NIS2313	数据库原理	2	2.0	83
MECH2508	理论力学	1	4.0	缓	NIS2328	软件工程	2	1.0	91
PHY1222	大学物理实验（2）	1	1.0	89	NIS2331	计算机组成与系统结构	2	2.0	90
PHY1252	大学物理(A类）（2）	1	4.0	85	PHY1253	大学物理(A类）（3）	2	2.0	92
CHEM1201	化学实验安全	2	1	93.00	EV901	环境与可持续发展	3	2.0	84

2022-2023学年

代码	课程名称	学期	学分	成绩	代码	课程名称	学期	学分	成绩
BI912	遗传学与社会	1	2.0	84	BME1204	生物医学工程研究的伦理及学术道德	2	1	97
DES1350H	创新思维与现代设计（荣誉）	1	3	93					
IP053	上海大学生创新创业训练计划（创新项目）	1	3	B+	IP054	上海大学生创新创业训练计划（创新项目）	2	4	B+
LIS1200	信息检索与利用	1	1	78	NIS2332	嵌入式及FPGA实验	2	3.0	90.18
MIL1202	军训	1	2	P	NIS3316	信息安全综合实践	2	3.0	90
NIS3302	信息安全科技创新	1	2.0	93	NIS3317	数据挖掘	2	2.0	92
NIS3303	信息论与编码	1	2.0	85	NIS3359	应用软件安全	2	2.0	85
NIS3318	数字信号处理（E类）	1	3.0	91	NIS3366	项目管理与软件设计	2	2.0	95.15
NIS3322	计算机通信网络(A类)	1	3.0	82	NIS4301	信息内容安全的理论与应用	2	2.0	82
NIS3323	编译原理（C类）	1	3.0	85	NIS4307	人工智能导论（A类）	2	2.0	93
NIS3325	操作系统（B类）	1	3.0	80	NIS4324	网络安全管理技术	2	2.0	90
NIS3365	现代密码学(1)	1	2.0	80	NIS3309	专业实习（信息安全）	3	2.0	A

2023-2024学年

代码	课程名称	学期	学分	成绩	代码	课程名称	学期	学分	成绩
NIS3607	区块链技术及应用	1	2.0	89	AI2615	算法设计与分析	2	3	W
NIS4315	网络安全攻防实战技术	1	2.0	73.70	NIS4332	毕业设计（论文）（信息安全）	2	4.0	B

注1-△表示该课程尚未通过 注2-↗表示学分转换课程 注3-P（通过）F（未通过）

注4-W表示中期退课　注5-本单加盖经办人及成绩证明章后为原件

注6-毕结业时间：20240630，毕结业证号：102481202405001123

注7-授予学位时间：20240630，学位证号：1024842024100881

学期1为秋季学期 学期2为春季学期 学期3为夏季学期

2024/07/15 上海交通大学电子信息与电气工程学院

经办人：　　主管：

上海交通大学电子信息与电气工程学院 本科生教务办公室

http://jwc.sjtu.edu.cn

IELTS™

Test Report Form

ACADEMIC

NOTE *Admission to undergraduate and post graduate courses should be based on the ACADEMIC Reading and Writing Modules. GENERAL TRAINING Reading and Writing Modules are **not** designed to test the full range of language skills required for academic purposes. It is recommended that the candidate's language ability as indicated in this Test Report Form be re-assessed **after two years** from the date of the test.*

Centre Number	CX208	Date	25/APR/2024	Candidate Number	511781

Candidate Details

Family Name	LIU
First Name(s)	ZIFEI
Candidate ID	142202200204290867

Date of Birth	29/04/2002	Sex (M/F)	F	Scheme Code	Private Candidate

Country or Region of Origin	CHINA (PEOPLE'S REPUBLIC OF)
Country of Nationality	
First Language	CHINESE

Test Results

Listening	6.5	**Reading**	8.0	**Writing**	6.0	**Speaking**	5.5	**Overall Band Score**	6.5	**CEFR Level**	B2

Administrator Comments

idp
Education Pty Ltd

Centre stamp

BRITISH COUNCIL
EXAMINATIONS SERVICES
CHINA

Validation stamp

VALIDATION STAMP
IELTS

Administrator's Signature

Date	27/04/2024	Test Report Form Number	24CX511781LIUZ208A

BRITISH COUNCIL idp CAMBRIDGE English

IELTS is jointly owned by the British Council; IDP IELTS; and Cambridge University Press & Assessment

The validity of this IELTS Test Report Form can be verified online by recognising organisations at http://ielts.ucles.org.uk

Research Proposal

1. Title

Accelerating and Optimizing Multimodal Large Models Using Mixture of Experts (MoE)

2. Background Review and Research Outlook

Current model work primarily focuses on single modalities, particularly in **Natural Language Processing (NLP)**. NLP is a subfield of artificial intelligence that focuses on the interaction between computers and humans through natural language. The goal of NLP is to enable computers to understand, interpret, and respond to human languages in a way that is both meaningful and useful. NLP involves various tasks, including language modeling, text classification, sentiment analysis, machine translation, and question answering[1]. The field has witnessed tremendous advancements, especially with the development of large-scale language models (LLM) like BERT[2], GPT[3], and their variants, which have significantly improved the performance of NLP applications across various domains[4].

However, the application of LLM has certain limitations. Firstly, in the real world, many scenarios involve multiple modalities of data, such as social media content that includes text, images, and videos[5]. Moreover, different modalities of data often contain complementary information. For example, images provide visual information, text provides semantic information, and audio provides speech information[6]. Integrating these types of information can enhance the model's performance in various tasks, such as image captioning, video subtitle generation, and sentiment analysis[7].

To address the aforementioned issues, **Multimodal Learning** has been proposed. The goal of multimodal learning is to create models that can understand and generate more comprehensive and contextually rich representations by leveraging the complementary information provided by different modalities[8]. Multimodal learning is crucial for tasks that require a holistic understanding of complex scenarios, such as video captioning, visual question answering, and audio-visual speech recognition[9]. By combining multiple modalities, these models can achieve better performance and provide more accurate and nuanced results compared to unimodal approaches[10].

Implementing multimodal functionality in LLMs , , faces numerous challenges, including cross-modal data integration and representation learning, acquisition and annotation of high-quality multimodal data, high computational resource requirements and efficiency optimization[11].

Multimodal Large Models (MLM) are advanced AI systems designed to simultaneously process multiple types of data to meet the demands of complex real-world scenarios[12]. These models build on the principles of multimodal learning, aiming to create unified representations that can capture the interactions and dependencies between different modalities[13]. MLMs have shown remarkable success in various applications, such as generating detailed image descriptions, understanding video content, and enhancing human-computer interactions through more natural and intuitive interfaces. Examples of state-of-the-art MLMs include CLIP (Contrastive Language–Image Pre-training)[14], DALL-E (a model that generates images from textual descriptions)[15], and ALIGN (a model that learns image-text alignments)[16].

Despite their impressive capabilities, multimodal large models are computationally intensive and resource-demanding[17]. Training and deploying these models require significant computational power, large memory capacities, and substantial energy consumption. To address these challenges, researchers have explored various acceleration techniques, including model compression (e.g., pruning[18], quantization[19]), efficient model architectures (e.g., transformers[20], convolutional networks[21]), and hardware optimizations (e.g., GPUs[22], TPUs[23]). These techniques aim to reduce the computational burden and energy consumption of MLMs while maintaining or improving their performance[24].

Mixture-of-Experts (MoE) models offer a promising solution to these challenges[25]. MoE models employ multiple expert networks, each specializing in different aspects of the data or task, and select a subset of experts for each input to perform computations, thus optimizing computational resources and enhancing model performance. This dynamic and sparse activation mechanism can significantly reduce the computational burden while maintaining or even improving the accuracy and robustness of the models[26].

Research Outlook:

Although Mixture of Experts (MoE) models have been effectively applied in various fields, their application in optimizing multimodal large models (MLMs) has not been fully explored. This research aims to bridge this gap by investigating the potential of MoE models in accelerating and optimizing MLMs, focusing on the following key areas:

- **Efficient Management of Multimodal Data Fusion and Processing(A)**: Explore how MoE models can be configured to effectively handle and integrate diverse types of data, ensuring seamless and efficient fusion and processing of multimodal inputs.
- **Improving Computational Efficiency through Sparse Activation and Dynamic Routing (B, C)**: Develop and implement strategies such as sparse activation and dynamic routing to enhance the computational efficiency of MLMs. This involves creating algorithms that dynamically select the most relevant experts based on the characteristics of the input data, thereby optimizing processing and reducing computational costs.
- **Enhancing Model Scalability and Robustness through Specialized Experts (D, E)**: Investigate the configuration of expert networks within the MoE framework to identify optimal sizes, numbers, and specializations of experts. This includes assessing how specialized experts can improve the scalability and robustness of the model, ensuring it can handle a wide range of tasks and scenarios effectively.

The underlined parts in parentheses indicate the specific Research Objectives corresponding to this point in Chapter 3. By focusing on these areas, this research aims to advance the application of MoE models in multimodal large models, ultimately leading to more efficient and robust AI systems capable of handling complex, real-world scenarios.

3. Research Objectives

A. Determine the optimal configuration of expert networks:

- In MoE models, optimizing the configuration of expert networks is crucial for achieving efficient and effective processing of multimodal data. The number of parameters plays a crucial role in MoE models because it directly affects the model's capacity and generalization ability. More parameters can provide a richer representational power, allowing experts to capture more nuanced features and patterns within their respective specialized domains. However, an increase in the number of parameters also brings more challenges, including longer training times, higher memory consumption, and the potential risk of overfitting.

- Current models often struggle with finding the right balance between model capacity and computational efficiency. Typically, expert models are not overly complex as they are trained on specific datasets. Determining the optimal configuration of expert models requires extensive experimentation and fine-tuning.

B. Develop an Expert Choice Routing algorithm for dynamic routing:

- Parameter efficiency is also a key factor, because even with a large number of parameters, if the expert model cannot effectively use these parameters to improve performance, it may lead to a waste of resources. Parameter efficiency is influenced by the degree of matching between the expert model's preferences and the tokens. In other words, it depends on whether the tokens routed to each expert can effectively update the expert model's parameters. The Routing Algorithm of Mixture of Experts is the mechanism responsible for dynamically selecting the appropriate expert model for computation when given input data. The main function of the algorithm is to determine which experts will be activated and involved in the processing based on the characteristics of the input data, thus optimizing the use of computational resources. Selecting the most relevant experts for processing based on input data characteristics and expert specialization, through the development of an expert choice routing algorithm, ensures that the model parameters are updated in the most efficient manner.
- The challenge in developing an expert choice routing algorithm lies in balancing the optimal matching of experts with the issue of load imbalance. For a set of tokens, their preference for each expert is likely to be uneven. Therefore, simply routing tokens to their most relevant one or few expert models can easily lead to an imbalance in expert load, resulting in wasted system resources and unreasonable training times for the expert layer.
- The issue of expert load imbalance may be alleviated by the shadow expert strategy outlined in part C.

C. Implement and evaluate the shadow expert strategy:

- The load imbalance caused by the frequent selection of certain experts can reduce system performance. Addressing this issue is crucial for maintaining computational efficiency.
- The difficulty in implementing the shadow expert strategy lies in the need for effective strategies to identify popular experts, replicate and distribute their parameters, and ensure that the additional communication overhead brought by executing this strategy is offset by the benefits provided by the shadow expert strategy.
- After implementing the shadow expert strategy, the next step will be to evaluate its effectiveness in achieving better load balancing and computational efficiency.

D. Evaluate the computational efficiency of the proposed MoE framework:

- Reducing computational costs and improving processing speed are key to the practical application of MoE frameworks. To validate the effectiveness of the proposed model, the computational efficiency of the proposed MoE framework needs to be evaluated.
- Conduct comprehensive evaluations focusing on key metrics such as floating point operations (FLOPs), inference time, and energy consumption. Compare the proposed MoE framework with traditional multimodal large models to determine its effectiveness.

E. Assess the improvement on model accuracy and robustness:

- Ensuring that the MoE framework maintains or improves the accuracy and robustness of multimodal large models is critical for their adoption in real-world applications.
- Investigate the performance of the MoE-based multimodal large model across various multimodal tasks, including image-text alignment, video understanding, and audio-visual fusion. Measure accuracy and robustness using task-specific performance metrics such as BLEU scores, accuracy, and AUC.

4. Methodology

A. Model Design:

a) **Develop a MoE-based architecture specifically designed for multimodal large models**: Design a framework capable of efficiently handling multimodal data, comprising multiple expert networks that specialize in processing different data modalities (such as text, images, audio, and video). Each expert network focuses on a specific modality's data processing tasks, enhancing computational efficiency and model performance. By allocating different modalities to specific experts, computational resources can be better utilized, and processing speed can be improved.

b) **Implementation of the shadow expert strategy**: To address the load imbalance caused by some experts being frequently selected, introduce a shadow expert strategy. This strategy replicates popular experts across all working nodes, thereby reducing the need for cross-node data transmission and lowering network communication overhead. During each training iteration, dynamically evaluate the load of each expert to determine which experts need to be shadowed, achieving better load balancing and computational efficiency.

c) **Use of Expert Choice Routing to implement a dynamic routing algorithm**: Develop a dynamic routing algorithm based on Expert Choice Routing, which dynamically selects the most relevant experts for processing based on the characteristics of the input data and the specialization of the experts. This algorithm leverages reinforcement learning or other optimization techniques to adjust expert selection strategies in real-time, ensuring that each input is processed optimally, thereby improving overall model performance and efficiency.

B. Dataset and Benchmarking:

a) Use established multimodal datasets such as COCO (image-text), YouCook2 (video-text), and AVSpeech (audio-visual) for training and evaluation.
b) Benchmark the proposed MoE framework against state-of-the-art MLMs using these datasets.

C. Training and Optimization:

a) Train the MoE-based multimodal large model using a combination of supervised learning and reinforcement learning techniques to optimize expert selection and activation.
b) Implement model compression techniques such as pruning and quantization to further enhance computational efficiency.

D. Performance Evaluation:

a) Measure computational efficiency through metrics like FLOPs (Floating Point Operations), inference time, and energy consumption.

b) Assess model accuracy and robustness using task-specific performance metrics such as BLEU scores for image captioning, accuracy for video classification, and AUC (Area Under Curve) for audio-visual integration.

5. Discussion

The integration of MoE models into multimodal large models presents a novel approach to addressing the computational challenges associated with these complex systems. By leveraging the specialized capabilities of multiple expert networks and employing a dynamic and sparse activation mechanism, this research aims to achieve significant improvements in computational efficiency without compromising accuracy. The findings from this research could have far-reaching implications for various applications, including autonomous systems, medical diagnostics, and interactive AI systems, where multimodal data integration is crucial.

Key areas for further exploration include:

Scalability: Investigating the scalability of the MoE framework for even larger multimodal datasets and more complex tasks.
Generalization: Assessing the generalization capabilities of the MoE-based MLMs across different domains and unseen data modalities.
Real-World Applications: Implementing and testing the MoE framework in real-world scenarios to evaluate its practical utility and robustness.

6. Reference

[1] Fanni, Salvatore Claudio, et al. "Natural language processing." Introduction to Artificial Intelligence. Cham: Springer International Publishing, 2023. 87-99.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT.

[3] Haupt, C. E., & Marks, M. (2023). AI-generated medical advice—GPT and beyond. Jama, 329(16), 1349-1350.

[4] Khurana, Diksha, et al. "Natural language processing: state of the art, current trends and challenges." Multimedia tools and applications 82.3 (2023): 3713-3744.

[5] Duong, Chi Thi Phuong. "Social media. A literature review." Journal of Media Research-Revista de Studii Media 13.38 (2020): 112-126.

[6] Lahat, Dana, Tülay Adali, and Christian Jutten. "Multimodal data fusion: an overview of methods, challenges, and prospects." Proceedings of the IEEE 103.9 (2015): 1449-1477.

[7] Schnotz, Wolfgang. "An integrated model of text and picture comprehension." The Cambridge handbook of multimedia learning 49.2005 (2005): 69.

[8] Xu, Peng, Xiatian, Zhu, and David A. Clifton. "Multimodal learning with transformers: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.10 (2023): 12113-12132.

[9] Blikstein, Paulo, and Marcelo Worsley. "Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks." Journal of Learning Analytics 3.2 (2016): 220-238.

[10] Fu, Yanwei, et al. "Learning multimodal latent attributes." IEEE transactions on pattern analysis and machine intelligence 36.2 (2013): 303-316.

[11] Tong, Shengbang, et al. "Eyes wide shut? exploring the visual shortcomings of multimodal llms." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[12] Yin, Shukang, et al. "A survey on multimodal large language models." arxiv preprint arxiv:2306.13549 (2023).

[13] Liang, Paul Pu, Amir Zadeh, and Louis-Philippe Morency. "Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions." arxiv preprint arxiv:2209.03430 (2022).

[14] Li, Yangguang, et al. "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm." arxiv preprint arxiv:2110.05208 (2021).

[15] Ge, Yunhao, et al. "DALL-E for Detection: Language-driven Compositional Image Synthesis for Object Detection." arxiv preprint arxiv:2206.09592 (2022).

[16] Jia, Chao, et al. "Scaling up visual and vision-language representation learning with noisy text supervision." International conference on machine learning. PMLR, 2021.

[17] Dumas, Bruno, Denis Lalanne, and Sharon Oviatt. "Multimodal interfaces: A survey of principles, models and frameworks." Human machine interaction: Research results of the mmi program. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. 3-26.

[18] Wang, Maolin, et al. "Large Multimodal Model Compression via Iterative Efficient Pruning and Distillation." Companion Proceedings of the ACM on Web Conference 2024. 2024.

[19] Cao, Yue, et al. "Collective deep quantization for efficient cross-modal retrieval." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 31. No. 1. 2017.

[20] Tsai, Yao-Hung Hubert, et al. "Multimodal transformer for unaligned multimodal language sequences." Proceedings of the conference. Association for computational linguistics. Meeting. Vol. 2019. NIH Public Access, 2019.

[21] Wang, Jinguang, et al. "Multimodal graph convolutional networks for high quality content recognition." Neurocomputing 412 (2020): 42-51.

[22] Vouitsis, Noël, et al. "Data-Efficient Multimodal Fusion on a Single GPU." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[23] Zellers, Rowan, et al. "Merlot: Multimodal neural script knowledge models." Advances in neural information processing systems 34 (2021): 23634-23651.

[24] Lepikhin, D., Lee, H., Xu, Y., Chen, Z., Firat, O., Huang, Y., ... & Chen, C. (2020). GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. arXiv preprint arXiv:2006.16668.

[25] Li, Yunxin, et al. "Uni-MoE: Scaling Unified Multimodal LLMs with Mixture of Experts." arxiv preprint arxiv:2405.11273 (2024).

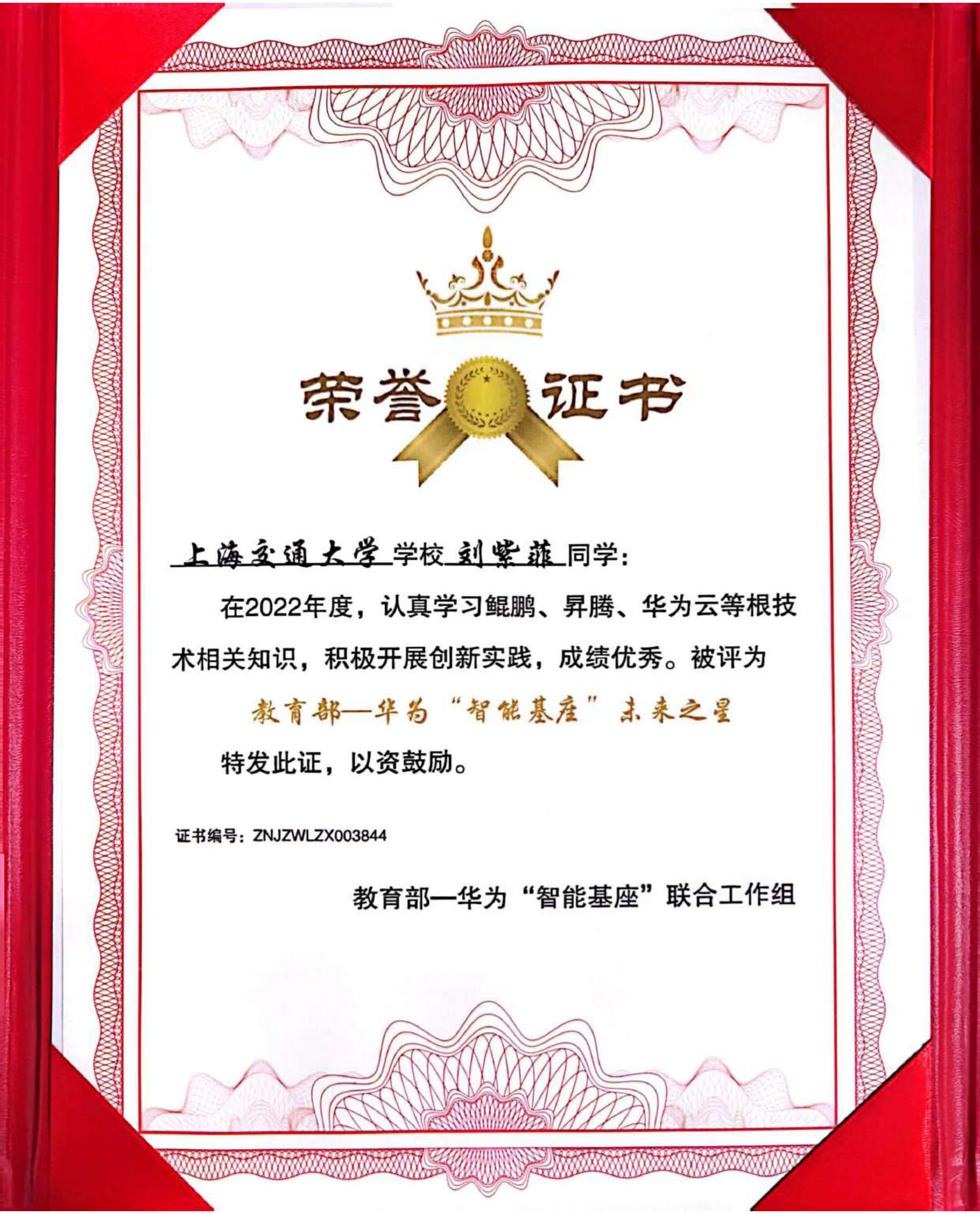
[26] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv preprint arXiv:1701.06538.

[27] Pawłowski, Maciej, Anna Wróblewska, and Sylwia Sysko-Romańczuk. "Effective techniques for multimodal data fusion: A comparative analysis." Sensors 23.5 (2023): 2381.

[28] You, Zhao, et al. "Speechmoe: Scaling to large acoustic models with dynamic routing mixture of experts." arxiv preprint arxiv:2105.03036 (2021).

[29] Huang, Quzhe, et al. "Harder Tasks Need More Experts: Dynamic Routing in MoE Models." arxiv preprint arxiv:2403.07652 (2024).

[30] Zoph, Barret et al. “ST-MoE: Designing Stable and Transferable Sparse Expert Models.” (2022).



荣誉证书

上海交通大学学校刘紫菲同学：

在2022年度，认真学习鲲鹏、昇腾、华为云等根技术相关知识，积极开展创新实践，成绩优秀。被评为

教育部—华为“智能基座”未来之星

特发此证，以资鼓励。

证书编号：ZNJZWLZX003844

教育部—华为“智能基座”联合工作组

荣誉证书

Certificate of Excellence

刘紫菲 同学:

荣获2020年度上海交通大学

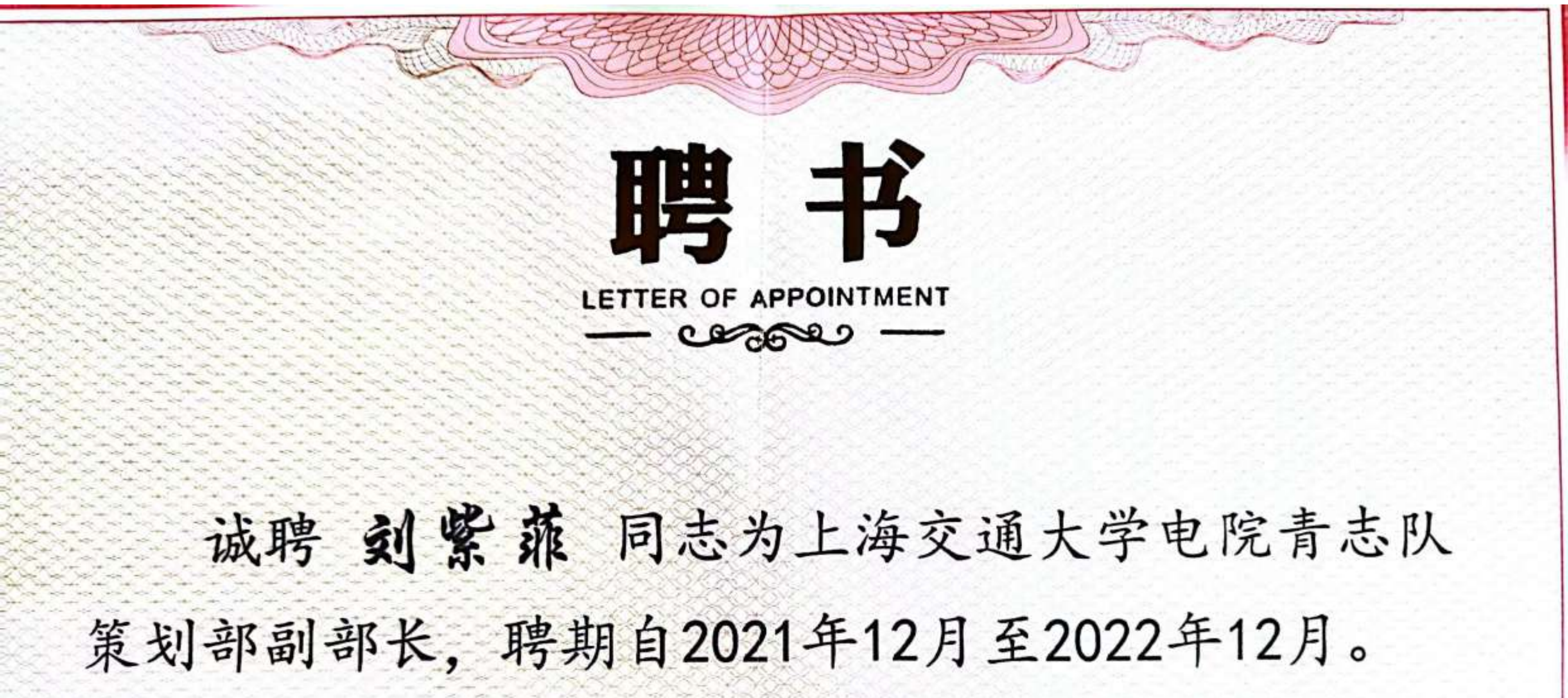
致远荣誉奖学金

特颁此证。

This is to certify that Liu Zifei is awarded the 2020 Zhiyuan Honors Scholarship of Shanghai Jiao Tong University.

上海交通大学

二〇二〇年十二月



聘书

LETTER OF APPOINTMENT

诚聘 **刘紫菲** 同志为上海交通大学电院青志队策划部副部长，聘期自2021年12月至2022年12月。

上海交通大学电院青志队
二〇二一年十二月十九日

applied sciences

MDPI

Article

A Combined Multi-Classification Network Intrusion Detection System Based on Feature Selection and Neural Network Improvement

Yunhui Wang [1,2,†], Zifei Liu [3,†], Weichu Zheng [3], Jinyan Wang [1,2], Hongjian Shi [3,*] and Mingyu Gu [4]

[1] National Key Laboratory of Science and Technology on Avionics System Integration, Shanghai 200233, China; wang.yh@outlook.com (Y.W.); wangjy121@avic.com (J.W.)
[2] China National Aeronautical Radio Electronics Research Institute, Shanghai 200233, China
[3] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; liuzifei@sjtu.edu.cn (Z.L.); sjtu_zwc0518@sjtu.edu.cn (W.Z.)
[4] Sino-European School of Technology, Shanghai University, Shanghai 200244, China; 22124558@shu.edu.cn
* Correspondence: shhjwu5@sjtu.edu.cn; Tel.: +86-137-1795-9365
† These authors contributed equally to this work.

Featured Application: Increased parallelism in edge network security issues, Feature loss handling.

Abstract: Feature loss in IoT scenarios is a common problem. This situation poses a greater challenge in terms of real-time and accuracy for the security of intelligent edge computing systems, which also includes network security intrusion detection systems (NIDS). Losing some packet information can easily confuse NIDS and cause an oversight of security systems. We propose a novel network intrusion detection framework based on an improved neural network. The new framework uses 23 subframes and a mixer for multi-classification work, which improves the parallelism of NIDS and is more adaptable to edge networks. We also incorporate the K-Nearest Neighbors (KNN) algorithm and Genetic Algorithm (GA) for feature selection, reducing parameters, communication, and memory overhead. We named the above system as Combinatorial Multi-Classification-NIDS (CM-NIDS). Experiments demonstrate that our framework can be more flexible in terms of the parameters of binary classification, has a fairly high accuracy in multi-classification, and is less affected by feature loss.

Keywords: feature loss; network intrusion detection systems; multi-classification; attack-type identification

check for updates

Citation: Wang, Y.; Liu, Z.; Zheng, W.; Wang, J.; Shi, H.; Gu, M. A Combined Multi-Classification Network Intrusion Detection System Based on Feature Selection and Neural Network Improvement. *Appl. Sci.* **2023**, *13*, 8307. https://doi.org/10.3390/app13148307

Academic Editor: Christos Bouras

Received: 12 June 2023
Revised: 4 July 2023
Accepted: 14 July 2023
Published: 18 July 2023

1. Introduction

Smart edge computing systems are distributed computing architectures that bring computing and data processing power closer to the network edge from traditional centralized cloud computing environments [1]. The design goal of smart edge computing systems is to place computing resources as close as possible to the data generation source or data usage endpoint to provide a lower latency, reduce network bandwidth pressure, and enhance user experience.

Smart edge computing systems face important challenges. Security and privacy have become more complex and critical in distributed environments [2,3]. Protecting systems from the threat of cyber attacks, data breaches, and malicious behavior is critical. At the same time, ensuring the security and privacy of data during transmission and storage, as well as effective authentication, access control, and vulnerability patching mechanisms, are key measures to ensure the security and privacy of smart edge computing systems.

One security measure in smart edge computing systems is NIDS (network intrusion detection system), which is used to monitor and detect potential intrusions in network traffic. NIDS detects intrusion attacks before they cause harm to the system and uses the

Article

A Federated Network Intrusion Detection System with Multi-Branch Network and Vertical Blocking Aggregation

Yunhui Wang [1,2,†], Weichu Zheng [3,†], Zifei Liu [3], Jinyan Wang [1,2], Hongjian Shi [3,*], Mingyu Gu [4] and Yicheng Di [5]

[1] National Key Laboratory of Science and Technology on Avionics System Integration, Shanghai 200233, China; wang.yh@outlook.com (Y.W.); wangjy121@avic.com (J.W.)
[2] China National Aeronautical Radio Electronics Research Institute, Shanghai 200233, China
[3] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; sjtu_zwc0518@sjtu.edu.cn (W.Z.); liuzifei@sjtu.edu.cn (Z.L.)
[4] Sino-European School of Technology, Shanghai University, Shanghai 200444, China; 22124558@shu.edu.cn
[5] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China; diyicheng@stu.jiangnan.edu.cn
* Correspondence: shhjwu5@sjtu.edu.cn; Tel.: +86-13717959365
† These authors contributed equally to this work.

Abstract: The rapid development of cloud–fog–edge computing and mobile devices has led to massive amounts of data being generated. Also, artificial intelligence technology, like machine learning and deep learning, is widely used to mine the value of the data. Specifically, detecting attacks on the cloud–fog–edge computing system using mobile devices is essential. External attacks on network press organizations led to anomaly flow in network traffic. The network intrusion detection system (NIDS) has been an effective method for detecting anomaly flow. However, the NIDS is hard to deploy in distributed networks because network flow data are kept private. Existing methods cannot obtain an accurate NIDS under such a federated scenario. To construct an NIDS while preserving data privacy, we propose a combined model that integrates binary classifiers into a whole network based on simple classifier networks to specify the type of attack on anomalous data and offer instruction to other security system components. We also introduce federated learning (FL) methods into our system and design a new aggregation algorithm named vertical blocking aggregation (FedVB) according to our model structure. Our experiments demonstrate that our system can be more effective than simple multi-classifiers in terms of accuracy and significantly reduce communication and computation overhead when applying FedVB.

Keywords: cloud–fog–edge computing; mobile device; artificial intelligence; network intrusion detection system; federated learning

Citation: Wang, Y.; Zheng, W.; Liu, Z.; Wang, J.; Shi, H.; Gu, M.; Di, Y. A Federated Network Intrusion Detection System with Multi-Branch Network and Vertical Blocking Aggregation. *Electronics* **2023**, *12*, 4049. https://doi.org/10.3390/electronics12194049

Academic Editor: Aryya Gangopadhyay

Received: 25 August 2023
Revised: 9 September 2023
Accepted: 22 September 2023
Published: 27 September 2023

1. Introduction

Massive amounts of data have been collected for various applications against the background of the rapid development of cloud–fog–edge computing and mobile devices. Such data contain valuable information, but that information is hard to mine. Artificial intelligence (AI) methods, like machine learning (ML) and deep learning (DL), are widely used in this context to extract practical knowledge from these data. DL has led to significant breakthroughs in distributed networks like the Internet of Things (IoT) [1,2] and cloud computing (CC). Specifically, DL has also been used to handle privacy and security problems, like network intrusion detection, in distributed machine learning (DML) or cloud–fog–edge computing (CFEC) frameworks.

In CFEC frameworks, the increasing number of threats to network traffic can come from various sources, causing a higher likelihood of organizations being exposed to intruders. Security mechanisms, especially network intrusion detection systems (NIDSs) [3], are

Automatic Pipeline Parallelism: A Parallel Inference Framework for Deep Learning Applications in 6G Mobile Communication Systems

Hongjian Shi, Weichu Zheng, Zifei Liu, Ruhui Ma, *Member, IEEE*, and Haibing Guan

Abstract— **With the rapid development of wireless communication, achieving the neXt generation Ultra-Reliable and Low-Latency Communications (xURLLC) in 6G mobile communication systems has become a critical problem. Among many applications in xURLLC, deep learning model inference requires improvement over its efficiency. Due to the heterogeneous hardware environment in 6G, parallel schedules from distributed machine learning and edge computing has been borrowed to tackle the efficiency problem. However, traditional parallel schedules suffer from high latency, low throughput, and low device utility. In this paper, we propose Automatic Pipeline Parallelism (AP^2), a parallel inference framework for deep learning applications in 6G mobile communication systems, to improve the model inference efficiency while maintaining reliability. AP^2 contains three sub-modules. A task-device affinity predictor predicts a task's expected execution time on a given device. The parallel inference arrangement optimizer finds the most suitable device for each task. The parallel inference scheduler converts the arrangement to a schedule that can be directly executed in the system. The experimental results show that AP^2 can achieve better latency, throughput, reliability, and device utility than other parallel schedules. Also, the priority of the sub-module designs has been approved through the experiments.**

Index Terms— **Distributed learning, system heterogeneity, parallel inference, hardware profiling, task scheduling.**

I. Introduction

WITH the increasing demand for wireless communication, Ultra-Reliable and Low-Latency Communications (URLLC) [1], [2], [3] has become a critical communication scenario in the 5th Generation (5G) mobile communication systems [4], [5], [6], [7]. 5G mobile communication systems cannot fulfill many Key Performance Indicators (KPIs) like high spectrum efficiency, throughput, network availability, etc., for different applications, leading to the development of the 6th Generation (6G) mobile communication systems [8], [9], [10] over the neXt generation URLLC (xURLLC) [11]. 6G mobile communication systems mainly focus on application-specific KPIs. Different from the development of the previous generations, 6G does not only improve the physical communication capability in the systems and improves the logical communication pattern based on application requirements. The logical communication pattern focuses on the arrangement of the communication flow and the distribution of the applications. In such a scenario, the latency and reliability metrics in URLLC are not adequate to measure communication performance. For such reason, we provide two new definitions of the metrics. The **latency** of the system refers to the time used to complete a specific deep learning application, which includes the communication and the computation times. The **reliability** of the system refers to the percentage of devices that can continue operating when device malfunction happens.

More specifically, deep learning [12] has become a critical application in 6G mobile communication systems with the increasing volume of data and the demand for data analysis. Deep learning can handle different missions, including image classification [13], image segmentation [14], natural language processing [15], etc. In addition, more and more users tend to accept deep learning models as a tool, which means that model inference in deep learning is more common than model training. Thus, integrating deep learning model inference into 6G mobile communication systems to reduce the inference latency, improve the inference throughput, and maintain reliability has become a crucial problem [16]. Developing strategies to organize the execution of the applications on the devices can increase resource utility, which is an efficient way to meet the target. As a result, we focus on achieving xURLLC for deep learning applications in 6G. The latency is represented by the inference latency of the deep learning applications, and the applicability of the inference process on the given hardware environment represents the reliability.

To achieve xURLLC in 6G mobile communication systems, researchers borrow ideas and techniques from two major fields: distributed machine learning, cloud computing, or edge computing. Distributed machine learning uses multiple devices to train the model or predict the result [17], [18]. In such a way, the system can handle more data and larger models while improving efficiency. There are different structures of distributed machine learning, including centralized structure [19], [20], semi-centralized structure [21], decentralized

Manuscript received 22 October 2022; revised 31 January 2023; accepted 7 March 2023. Date of publication 29 May 2023; date of current version 6 July 2023. This work was supported in part by the National NSF of China under Grant 61872234 and Grant 61732010; in part by the Shanghai Key Laboratory of Scalable Computing and Systems, Innovative Research Foundation of Ship General Performance, under Grant 25622114; in part by the Key Laboratory of Phytium-Kylin (PK) System Technologies Research of Hainan; in part by the Internet of Things (IoT) Special Subject Program; in part by the China Institute of the IoT (Wuxi); and in part by the Wuxi IoT Innovation Promotion Center. *(Corresponding author: Ruhui Ma.)*

The authors are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shhjwu5@sjtu.edu.cn; sjtu_zwc0518@sjtu.edu.cn; liuzifei@sjtu.edu.cn; ruhuima@sjtu.edu.cn; hbguan@sjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2023.3280970.

Digital Object Identifier 10.1109/JSAC.2023.3280970