MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
IEEE COMMUNICATIONS SOCIETY
http://mmc.committees.comsoc.org/

MMTC Communications – Review

IEEE COMMUNICATIONS SOCIETY

Vol. 15, No. 1, April 2024

TABLE OF CONTENTS

Message from the Review Board Directors

Welcome to the April 2024 issue of the IEEE ComSoc MMTC Communications – Review.

This issue comprises five reviews that cover multiple facets of multimedia communication research including remote visual monitoring, multimodal emotion analysis, ambient backscatter communication, and vision transformer model compression. These reviews are briefly introduced below.

The first paper, published in IEEE Transactions on Wireless Communications and edited by Dr. Qing Wang, considers the privacy leakage issue that arises during decentralized inference in graph neural networks (GNNs) and proposes solutions for privacy-preserving decentralized reasoning using GNNs in wireless networks.

The second paper, published in IEEE Transactions on Multimedia and edited by Dr. Yujiao Hu, introduces the typical resources in Computing and Network Convergence (CNC), which include heterogeneous computing nodes distributed across various locations.

The third paper, edited by Dr. Qifan Liu, was also published in IEEE Transactions on Circuits and Systems for Video Technology. This paper proposes a two-stage background suppression and foreground alignment framework for few-shot fine-grained recognition, which can be trained in an end-to-end manner.

The fourth paper, edited by Dr. Wenming Cao, was published in IEEE Transactions on Medical Imaging. The authors introduce a Recursive Deformable Pyramid (RDP) network that leverages a pure convolutional pyramid structure to achieve efficient and accurate registration without the need for high-weight attentions or transformers.

The fifth paper, published in the Proceedings of the IEEE/CVF International Conference on Computer Vision 2023, and edited by Dr. Jianqi Zhong, proposes a new framework called HumanMAC, which diverges from traditional encoding-decoding paradigms by using a masked motion completion method.

All the authors, reviewers, editors, and others who contribute to the release of this issue deserve appreciation with thanks.

IEEE ComSoc MMTC Communications – Review Directors

Yao Liu
Rutgers University, USA
Email: yao.liu@rutgers.edu

Wenming Cao
Shenzhen University, China
Email: wmcao@szu.edu.cn

Phoenix Fang
California Polytechnic State University, USA
Email: dofang@calpoly.edu

Ye Liu
Macau University of Science and Technology, Macau, China
Email: liuye@must.edu.mo

Privacy-Preserving Decentralized Inference with Graph Neural Networks

A short review for "Privacy-Preserving Decentralized Inference With Graph Neural Networks in Wireless Networks"

Edited by Qin Wang

M. Lee, G. Yu and H. Dai, "Privacy-Preserving Decentralized Inference With Graph Neural Networks in Wireless Networks," in IEEE Transactions on Wireless Communications, vol. 23, no. 1, pp. 543-558, Jan. 2024.

In recent years, wireless technologies based on artificial intelligence (AI) have attracted more and more attention. AI is a potential tool for developing breakthrough wireless technologies for the next generation of wireless communications [1]. However, there are several problems when applying AI in wireless communications. First, neural networks usually require a large number of training samples. It is difficult to obtain labeled training samples in general wireless communication systems. Meanwhile, too many training samples can make the training process both memory and time-consuming. Second, as far as most existing studies are concerned, the structure of neural network models is highly dependent on the system size, lacks generalization ability, and cannot be used when the system size changes. Finally, most existing AI-based wireless algorithms are centralized and suffer from high signaling costs, traffic congestion, low flexibility, and limited computational power.

Graph neural network (GNN), as a neural network model especially proposed for graph data, has achieved good performance in various graph-related applications [2]. Wireless networks have characteristics of high-dimensional features, active interactions between nodes, and potential time-varying structures. GNN can effectively integrate graph structures into the neural network's architecture, model node attributes, and relationships between nodes, thus mining the hidden features in the graph structure data [3]. Therefore, GNN not only has good scalability for large-scale graphs and generalization ability for dynamic graph structures, but also can obtain near-optimal performance through more efficient training [4]. As a neural network specialized for graph-structured data, GNN enriches the features of each node by extracting neighborhood information and propagating these features to the graph according to its structure, thus obtaining near-optimal learning performance with good scalability and generalization ability. Based on the topology of the graph, applications of GNN can cover resource management problems in device-to-device (D2D), cellular, cell-free, distributed networks and other signal processing domains.

In real systems, users are becoming more aware of personal information protection. During the process of exchanging information between neighboring nodes, private information and important symptoms that users do not want to share with others may be inferred or detected by other users, including honest but curious neighbors or anonymous hackers. Therefore, in order to prevent privacy leakage in wireless transmission, Privacy-Preserving decentralized inference with GNN in wireless networks is an important issue that cannot be ignored.

For the first time, the authors study privacy-preserving decentralized reasoning using GNNs in wireless networks. First, the authors propose the utilization of local differential privacy as a privacy-preserving metric and novel privacy-preserving wireless signaling. Artificial noise and channel noise are utilized to achieve privacy-preserving decentralized inference using GNNs in wireless networks. Based on the optimal parameter solutions for privacy-preserving signaling, the authors define the signal-to-noise ratio-privacy trade-off function and theoretically analyze the performance upper bound of the proposed wireless signaling.

Introducing noise leads to degradation of estimation accuracy, which in turn leads to degradation of inference performance. However, traditional training algorithms do not take the wireless environment into account and cannot overcome the above problems. Therefore, the authors design a privacy-assured training algorithm based on privacy-preserving wireless signaling to mitigate the effects of noise and fading and to improve the inference performance of decentralized GNNs with certain privacy guarantees. The basic idea is to utilize noise and fading for training. Compared with traditional GNN training algorithms, the proposed privacy-assured training algorithm is improved in the following three points: (1) Enhanced training samples: Wireless transmission parameters and privacy protection are also included in the training samples. (2) Preprocessing of training samples: Before training, the privacy-preserving signaling parameters of each training sample are computed. (3) Corrected forward pass: According to the idea of adversarial training, noise and fading are injected into the forward pass process.

The proposed GNN-based training algorithms for optimal wireless signal design, performance cap analysis and improved privacy-preserving decentralized inference are generic and applicable not only to different wireless applications but also to different communication models. The authors have found that the use of over-the-air computing (Aircomp) techniques for decentralized reasoning for GNNs in wireless networks can improve communication and computational efficiency as well as privacy preservation. This is because, in the absence of the Aircomp technique, all signals are transmitted in an orthogonal manner and the privacy of each neighbor can only be protected by artificial noise and channel noise. However, in a system with Aircomp technology, the signals from all neighbors are mixed. Thus, each neighbor's privacy is protected not only by artificial noise and channel noise, but also by the transmitted signals from other users. In this way, it is more difficult for an honest but curious neighbor or adversary to successfully detect the personal characteristics of each user. Thus, the Aircomp technique not only improves communication and computational efficiency, but also improves the performance and privacy of decentralized reasoning in GNNs.

Facilitating decentralized reasoning for GNNs is crucial for taking GNN-based wireless technologies from theory to practice, and the authors are among the first to investigate decentralized reasoning for privacy-preserving using GNNs over noisy and fading wireless channels. The authors first design a new privacy-preserving wireless signaling for GNN-based privacy-preserving distributed reasoning. Based on this, a privacy-guaranteed training algorithm is proposed to further improve the performance of distributed reasoning with GNNs in wireless networks. In addition, the authors theoretically demonstrate that the Aircomp technique can simultaneously improve communication and computational efficiency as well as privacy preservation.

References:

[1] W. Saad, M. Bennis and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," in IEEE Network, vol. 34, no. 3, pp. 134-142, May/June 2020.

[2] S. Munikoti et al., "Challenges and Opportunities in Deep Reinforcement Learning With Graph Neural Networks: A Comprehensive Review of Algorithms and Applications," in IEEE Transactions on Neural Networks and Learning Systems, 2023.

[3] Y. Shi et al., "Machine Learning for Large-Scale Optimization in 6G Wireless Networks," in IEEE Communications Surveys & Tutorials, vol. 25, no. 4, pp. 2088-2132, Fourthquarter 2023.

[4] J. Suárez-Varela et al., "Graph Neural Networks for Communication Networks: Context, Use Cases and Opportunities," in IEEE Network, vol. 37, no. 3, pp. 146-153, May/June 2023.

Qin Wang, Ph.D, is an Associate Professor at Nanjing University of Posts and Telecommunications (NJUPT), China. She received B.S. and Ph.D degrees from NJUPT, in 2011 and 2016. Prior to joining NJUPT, she was with the New York Institute of Technology (NYIT) between Feb. 2017 and Aug. 2020. From July 2018 to June 2020, she was a Postdoctoral Research Fellow at NJUPT. From 2015 to 2016, she was a visiting scholar at San Diego State University, USA. Her research interests include multimedia communications, multimedia pricing, resource allocation in 6G, and Internet of Things. She has published papers in prestigious journals such as IEEE Transactions on Vehicular Technology and IEEE Communications Magazine, in prestigious conferences such as IEEE INFOCOM.

Scheduling in Computing and Network Convergence

A short review for "A Survey on Scheduling Techniques in Computing and Network Convergence"

Edited by Yujiao Hu

Tang S, Yu Y, Wang H, et al. A Survey on Scheduling Techniques in Computing and Network Convergence[J]. IEEE Communications Surveys & Tutorials, 2023.

Recent advancements in Computing and Network Convergence (CNC), also known as Computing Force Network or Computing Power Network [1-5], have facilitated the interconnection of various computing resources distributed across ubiquitous locations through the network. This advancement is anticipated to enable services to be as instantly available as water and electricity. Scheduling serves as the core technology of the CNC control plane, playing a pivotal role in achieving ultra-low-latency and on-demand computing [6]. It determines the placement of tasks on computing nodes and the allocation of resources.

CNC schedulers are tasked with coordinating diverse computing entities such as geographically distributed data centers, clouds, edge devices, and terminals, interconnected by networks. These schedulers assign the execution location of tasks, impacting data and storage scheduling, network scheduling, computing performance, transmission delay, and cost. The speed of task processing is primarily determined by the availability of computing and network resources. However, the previous studies address either cloud, edge, geo-distributed DCs, or specific scheduling issues only [7][8], while none of them focuses on the hierarchical scheduling issue with both the network and computing resources in CNC in a comprehensive way. This paper, "A Survey on Scheduling Techniques in Computing and Network Convergence", succeeds in achieving overview of the scheduling approaches in computing and network convergence.

In this paper, the authors introduce the typical resources in Computing and Network Convergence (CNC), which include heterogeneous computing nodes distributed across various locations. CNC schedules a multitude of services to these nodes on demand through the unified coordination of multi-dimensional resources. For example, when running Metaverses for agriculture across several regions to provide control decisions, multiple resources are utilized. These resources include GPUs, CPUs, FPGAs, and TPUs, each serving different purposes within the CNC architecture. Schedulers play a crucial role in analyzing tasks and selecting appropriate resources for execution. In the context of this survey, more attention is paid to GPU scheduling, In Network Computing (INC), as well as schedulers for other heterogeneous computing resources such as FPGAs and TPUs.

Secondly, the authors highlight that Computing and Network Convergence (CNC) deals with a wide range of tasks with diverse objectives. In light of this, they review schedulers with different scheduling objectives, focusing on several key aspects. Given the extensive usage of computing power, the initial focus should prioritize environmental responsibility. Additionally, Quality of Service (QoS) considerations such as meeting deadlines, managing costs, and ensuring trustworthiness must be addressed. Furthermore, there is potential in exploring AI for developing hands-free, flexible schedulers that can achieve higher performance.

Following that, the authors summarize various scheduling tasks in CNC, including scheduling in real-time tasks, AI training and inference tasks, big data applications with geographically distributed DCs, and High-Performance Computing (HPC) tasks. The tasks are motivated by the practical applications. Real-time needs are sometimes vital. Content Delivery Networks, edge computing, and other technologies provide real-time infrastructure. Smart factories, autonomous driving, and Metaverse are classic real-time scenarios. Despite sufficient resources, the tasks can conflict and cause missed deadlines. Many real-time tasks exist in parallel. The hard real-time task has strict time constraints, and a delay always leads to a significant impact or even a catastrophe, while soft real-time tasks could bear the violation to some degree and the delay does not necessarily cause performance degradation. For the typical AI training and inference tasks in CNC, the training workloads may last hours or even months, whereas the inference workloads usually pursue a real-time performance by using the trained model, pre-cached services, and historical results. The scheduler could better achieve this by balancing placement sensitivity and queueing delay, utilization rate and mutual interference, job characteristics and priorities. Also the big data applications and HPC tasks are arisen from practical scenarios. The scheduling solutions for these scenarios with different characteristics are crucial to the realization of the CNC. The authors presenting the detailed scheduling techniques in the paper depicts a clear view to understand the technical progresses.

Finally, the authors consider the directions and challenges in scheduling to achieve CNC, including intelligent computing, distributed scheduling, real-time computing scheduling, green CNC, digital twins, credibility, and security concerns in case of possible data and job exposure. Each listed challenge is formidable and will bring about significant performance improvements to schedulers.

In summary, the paper comprehensively reviews the literature on scheduling in various scenarios, covering the scheduling problem of Computing and Network Convergence from heterogeneous resources, multiple-objective optimization, and diverse tasks. The authors also point out important challenges for future work, which can facilitate the development of domain studies.

References:

[1] Yukun S, Bo L, Juniin L, et al. Computing power network: A survey[J]. China Communications, 2024.

[2] Li J, Lv H, Lei B, et al. A computing power resource modeling approach for computing power network[C]//2022 International Conference on Computer Communications and Networks (ICCCN). IEEE, 2022: 1-2.

[3] Tang X, Cao C, Wang Y, et al. Computing power network: The architecture of convergence of computing and networking towards 6G requirement[J]. China communications, 2021, 18(2): 175-185.

[4] Duan Q, Yan Y, Vasilakos A V. A survey on service-oriented network virtualization toward convergence of networking and cloud computing[J]. IEEE Transactions on Network and Service Management, 2012, 9(4): 373-392.

[5] Bouras M A, Farha F, Ning H. Convergence of computing, communication, and caching in Internet of Things[J]. Intelligent and Converged Networks, 2020, 1(1): 18-36.

[6] Kumar M, Sharma S C, Goel A, et al. A comprehensive survey for scheduling techniques in cloud computing[J]. Journal of Network and Computer Applications, 2019, 143: 1-33.

[7] Luo Q, Hu S, Li C, et al. Resource scheduling in edge computing: A survey[J]. IEEE Communications Surveys & Tutorials, 2021, 23(4): 2131-2165.

[8] Goudarzi M, Palaniswami M, Buyya R. Scheduling IoT applications in edge and fog computing environments: a taxonomy and future directions[J]. ACM Computing Surveys, 2022, 55(7): 1-41.

Yujiao Hu, received her Bachelor and PhD degrees from the Department of Computer Science of Northwestern Polytechnical University, Xi'an, China, in 2016 and 2021 respectively. From Nov. 2018 to March 2020, she was a visiting PhD student in National University of Singapore. Currently, she is a faculty member in Purple Mountain Laboratories. She focuses on deep learning, edge computing, multi-agent cooperation problems and time sensitive networks.

Background Suppression and Foreground Alignment for Few-Shot Fine-Grained Image Classification

A short review for "Boosting Few-Shot Fine-Grained Recognition With Background Suppression and Foreground Alignment"

Edited by Qifan Liu

Z. Zha, H. Tang, Y. Sun and J. Tang, "Boosting Few-Shot Fine-Grained Recognition With Background Suppression and Foreground Alignment," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 8, pp. 3947-3961, Aug. 2023, doi: 10.1109/TCSVT.2023.3236636..

Few-shot learning (FSL) [1] has received widespread attention in the computer vision and multimedia fields because it mimics the ability of humans to learn new concepts with few available examples. Fine-grained recognition (FGR) [2] a popular and challenging problem, aims to recognize images of multiple sub-categories belonging to a super-category (e.g., birds, dogs, cars). Considering the manual annotation for fine-grained images requires domain-specific knowledge, it is labor- and time-consuming to collect high quality and fully labeled large-scale datasets thus FGR is a suitable application scenario of FSL. In this paper, the authors study a more challenging and practical task, namely few-shot fine-grained recognition (FS-FGR), which aims to recognize fine-grained objects under few-shot settings.

In order to migrate the model from general datasets to fine-grained datasets, current FS-FGR models generally attempt to capture key distinctions, but can easily ignore the negative effects of background. Therefore, removing the influence of the background (e.g., by manually annotating bounding boxes) is significant to boost current few-shot methods. However, achieving better performance with human-annotated bounding boxes runs counter to the original aspiration of FSL to free people from burdensome and boring annotation tasks. Thus, aligning semantically relevant local regions remains non-trivial for the FS-FGR task, which could effectively reduce the negative effects of feature inconsistency caused by the changes of object pose or viewing angle. Therefore, the authors aim to remove cluttered backgrounds and align semantically relevant foregrounds when only image-level labels and limited training images are available.

In this paper, to tackle these problems with only image-level labels available, the authors propose a two-stage background suppression and foreground alignment framework for few-shot fine-grained recognition, which can be trained in an end-to-end manner. The proposed method mainly consists of four modules: a feature extractor, a Background Activation Suppression (BAS) module, a Foreground Object Alignment (FOA) module, and a Local to Local (L2L) similarity metric. Firstly, a feature extractor is used to extract the features for subsequent matching and localization learning. The BAS aims to generate a foreground mask map for localization based on the activation maps since the position in the activation map with a higher value are often where the interesting parts are located. Specifically, without adding extra trainable parameters, the authors only process the feature map to generate object location coordinates, and the BAS is supervised by the global classification loss. With the help of the generated bounding box information, the authors further obtained the finer scale of the object image by cropping and zooming in to remove the cluttered background. Different from the conventional method, the authors incorporate both the original images and the refined ones obtained by BAS into the model for learning, which can effectively generate fine-grained tailored representations for few-shot recognition.

Traditional fine-grained recognition is a challenging problem due to the small inter-category variations and the large intra-category variations that exist in the fine-grained dataset. The visual patterns of different sub-categories are similar to some extent, and the backgrounds in the same sub-categories may be different but appear slightly similar among different sub-categories. Therefore, the image background plays a negative role in the fine-grained recognition task, especially in the few-shot scenario. Inspired by the weakly-supervised localization methods [3] for the fine-grained task, the proposed method attempts to eliminate the negative effect of background by the

BAS module, without additional annotations except for image-level annotation. BAS directly generates class-agnostic activation maps to disentangle the foreground object and background in an image.

Although global embedding generated by global average pooling or global max pooling operations is capable of capturing discriminative information in general few-shot learning tasks, it might suppress some local discriminative characteristics while overemphasizing the irrelevant background features for the fine-grained object [4], hurting the FG-FSR performance. To this end, their proposed method attempts to exploit dense local features to calculate the similarity between a given pair of samples, which can further capture spatial structure information of the foreground for FG-FSR. In addition, the lower intra-category visual consistency of fine-grained objects usually causes different similarities in both global appearance and various local regions. Therefore, an ideal solution to FS-FGR should not only be sensitive to the subtle discrepancy among instances from different sub-categories but also be invariant to the arbitrary poses, scales, and appearances of the same subcategory. In light of this, their proposed method goes two steps further to compute the similarity of each support-query feature pair: (1) aligning semantic features between a query image and various support images according to their correlations; (2) modeling the similarity measurement between each aligned support-query pair as a local-to-local (patch-to-patch) matching process. Considering that the semantic features aligned by the FOA module are no longer affected by the position deviation, they further propose a patch-level similarity metric, indicated by L2L , which reformulates similarity measurement as a local-to-local matching process.

The authors conduct experiments on three widely used benchmarks for FG-FSR, including CUB-200-2011, Stanford Dogs, and Stanford Cars. They follow the same evaluation protocol as the previous work [5], and all the input images are resized to 84 × 84.

In summary, the authors show that background suppression and foreground alignment matter for few-shot fine-grained recognition (FS-FGR). They propose a novel two-stage weakly-supervised framework for FS-FGR, where the background suppression and foreground alignment are jointly learned in an end-to-end manner. In particular, a background activation suppression module is introduced in the raw stage to weaken background disturbance and enhance dominative foreground objects for the refined stage. Extensive experiments on three fine-grained benchmarks demonstrate that their proposed model can be trained in an end-to-end manner with only the image-level label, and achieve state-of-the-art performance.

References:

[1] J. Snell, K. Swersky and R. Zemel, "Prototypical networks for few-shot learning", Proc. Adv. Neural Inf. Process. Syst., pp. 4077-4087, 2017.

[2] T.-Y. Lin, A. RoyChowdhury and S. Maji, "Bilinear CNN models for fine-grained visual recognition", Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 1449-1457, Dec. 2015.

[3] F. Zhang, M. Li, G. Zhai and Y. Liu, "Multi-branch and multi-scale attention learning for fine-grained visual categorization", Proc. Int. Conf. Multimedia Modeling, vol. 12572, pp. 136-147, 2021.

[4] Yan S, Tang H, Zhang L, et al. Image-specific information suppression and implicit local alignment for text-based person search[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023.

[5] Y. Wu et al., "Object-aware long-short-range spatial alignment for few-shot fine-grained image classification", Proc. ACM Multimedia, pp. 107-115, 2021.

Qifan Liu, received his PhD degrees from College of Electronics and Information Engineering, Shenzhen University, Guangdong, China, in 2023. Currently, he is a Postdoctor in the Department of Electronics and Electrical Engineering, Southern University of Science and Technology, Guangdong, China. He focuses on deep learning, few-shot learning and large language model.

Recursive Deformable Pyramid Network for Medical Image Registration

A short review for "Recursive Deformable Pyramid Network for Unsupervised Medical Image Registration"

Edited by Wenming Cao

H. Wang, D. Ni and Y. Wang, "Recursive Deformable Pyramid Network for Unsupervised Medical Image Registration," in IEEE Transactions on Medical Imaging, doi: 10.1109/TMI.2024.3362968.

Deformable medical image registration (DMIR) has been widely deployed in various medical image applications, such as image diagnostics, surgical navigation, disease monitoring [1]. DMIR enables clinical comparison of images captured from different devices, patients, or times. The anatomical correspondence is established to predict a deformation field that warps a moving image and aligns it with the other fixed one [2].

Compared to existing methods, large deformations in non-rigid registration tasks present unique challenges due to the complexity of underlying tissue dynamics and the computational demands of precise alignment. These challenges necessitate advanced models that can capture intricate anatomical variations while operating efficiently. While networks like VoxelMorph [3] have made strides in unsupervised registration by utilizing deep learning for efficient displacement field estimation, they often fall short when dealing with significant tissue deformations. This is primarily because single-stage networks may not capture the intricate details required for high-fidelity registration, particularly in cases involving large spatial discrepancies between the images. Recent research has shown that the pyramid deformation approach is designed to overcome these limitations by leveraging a multi-resolution strategy that predicts a sequence of increasingly refined deformation fields [4,5]. However, one major drawback of traditional pyramid approaches is their computational complexity and the potential loss of semantic information across different resolution levels.

In this paper, the authors introduce a Recursive Deformable Pyramid (RDP) network that leverages a pure convolutional pyramid structure to achieve efficient and accurate registration without the need for high-weight attentions or transformers. The network employs a step-by-step recursive strategy, integrating high-level semantic information to predict the deformation field from coarse to fine scales, while ensuring the rationality of the deformation field. Furthermore, due to the recursive pyramid strategy, the network can effectively achieve deformable registration without the need for separate affine pre-alignment.

To be specific, the proposed network includes two key steps. In the first step, called feature encoding with ResNet design, the network involves a dual-stream encoder architecture that separately processes the fixed and moving images to extract hierarchical features. This step adopts the ResNet design, whose encoder is composed of a 4-layer convolutional block that progressively doubles the number of channels with each layer, allowing for the capture of features at multiple scales. The use of ResNet enhances the network's ability to learn complex transformations by leveraging its deep architectural design. The encoded feature maps from the fixed and moving images are then prepared for the subsequent deformation field prediction and refinement stages. The choice of the ResNet design is important as it lays the groundwork for accurate registration by providing a rich set of features that capture the nuances of the medical images.

In the second step, the proposed recursive deformable pyramid strategy involves multiple decoding layers that iteratively refine the deformation field. At each level of the pyramid, the network performs a series of recursive operations that integrate high-level semantic information, leading to a progressively detailed and accurate deformation field. The the network effectively decomposes complex deformations and enhances the registration accuracy. At the core of each recursion is the deformation estimation block, which predicts the residual deformation field. This block consists of a convolutional layer for generating a velocity field and a diffeomorphism layer to ensure the generated deformation field is physically plausible and smooth. The newly estimated residual deformation field is then fused with the previous deformation field to generate a refined deformation field. The final output is a semantically rich deformation field that aligns the moving image with the fixed image in a physically plausible manner.

The authors compare the RDP network with several state-of-the-art registration methods using three public brain MRI datasets: LPBA, Mindboggle, and IXI. The experimental results demonstrate that the RDP network consistently outperforms existing methods across various metrics, including Dice score, average symmetric surface distance, Hausdorff distance, and Jacobian. Notably, the network maintains high performance even without affine pre-alignment, showcasing its robustness in handling large deformations. The results suggest that the proposed network is highly effective in accurately registering brain MRI images and handling large deformations.

In summary, the RDP network offers a promising solution for unsupervised non-rigid medical image registration with its efficient recursive strategy and pure convolutional pyramid design. It adeptly tackles the challenge of large deformations through a hierarchical, recursive approach that iteratively refines the deformation field. The initial stage employs a dual-stream encoder to meticulously extract features from both the fixed and moving images, setting the foundation for subsequent deformation field predictions. The network then iteratively refines these predictions through a series of recursive steps, each building upon the previous to achieve a highly accurate registration. The paper contributes significantly to the field of medical image registration, providing a robust and practical approach that could be readily integrated into clinical workflows.

References:

[1] X. Deng, E. Liu, S. Li, Y. Duan and M. Xu, "Interpretable Multi-Modal Image Registration Network Based on Disentangled Convolutional Sparse Coding," in IEEE Transactions on Image Processing, vol. 32, pp. 1078-1091, 2023.

[2] A. Sotiras, C. Davatzikos and N. Paragios, "Deformable Medical Image Registration: A Survey," in IEEE Transactions on Medical Imaging, vol. 32, no. 7, pp. 1153-1190, July 2013.

[3] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag and A. V. Dalca, "VoxelMorph: A Learning Framework for Deformable Medical Image Registration," in IEEE Transactions on Medical Imaging, vol. 38, no. 8, pp. 1788-1800, Aug. 2019.

[4] M. Kang, X. Hu, W. Huang, M. R. Scott, and M. Reyes, “ Dual-stream pyramid registration network, ” Medical Image Analysis, vol. 78, p.102379, 2022.

[5] Z. Chen, Y. Zheng and J. C. Gee, "TransMatch: A Transformer-Based Multilevel Dual-Stream Feature Matching Network for Unsupervised Deformable Image Registration," in IEEE Transactions on Medical Imaging, vol. 43, no. 1, pp. 15-27, Jan. 2024.

Wenming Cao, received the M.S. degree from the System Science Institute, China Science Academy, Beijing, China, in 1991, and the Ph.D. degree from the School of Automation, Southeast University, Nanjing, China, in 2003. From 2005 to 2007, he was a PostDoctoral Researcher with the Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China. He is currently a Professor with Shenzhen University, Shen zhen, China. He is also a foreign academician of the Russian Academy of Natural Sciences. He has authored or coauthored over 80 publications in toptier conferences and journals. His research interests include pattern recognition, image processing, and visual tracking.

Denoising Diffusion Model for 3D Human motion prediction

A short review for "HumanMAC: Masked Motion Completion for Human Motion Prediction"

Edited by Jianqi Zhong

Chen L H, Zhang J, Li Y, et al., "HumanMAC: Masked Motion Completion for Human Motion Prediction," in Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 9544-9555.

3D skeleton-based human motion prediction aims to forecast future human motions based on past observations, which has a wide range of applications, such as human-machine interaction [1] and autonomous driving [2].

Prior state-of-the-art methods work in an encoding-decoding fashion to tackle the HMP problem, which conditions on previous motion frames and predicts unobserved motions [3, 4, 5]. Technically, these methods first encode the previous motion frames to latent representations explicitly and then decode the latent representations into prediction results [6].

Although these methods enjoy the good performance in some scenarios, they are still unsatisfactory in practice. We detail the issues from three aspects. (1) Most state-of-the-art methods rely on multiple loss constraints for high-quality prediction results. e.g., the average pairwise distance, final displacement error [7], and adversarial loss [8].

Consequently, they need carefully designed hyper-parameters to balance different loss constraints, which makes it laborious for method applications. (2) Previous state-of-the-art methods need multistage training [9,10]. That is to say, the learning of the encoder/decoder and sampling in the latent space is performed in different stages. To make matters worse, complex pipelines always require an additional stage of engineering tuning. (3) For the methods, it is hard to realize the switch of different categories of motions, e.g., switch from WalkDog to Sitting, which is pivotal for result diversity. The reason is that these methods are largely limited to observed motion sequences for prediction, which include few such switches.

This paper proposes a new framework called HumanMAC, which diverges from traditional encoding-decoding paradigms by using a masked motion completion method. This framework aims to address the limitations of existing methods, such as complex loss functions, multi-stage training, and difficulty in switching between different motion categories. The introduction of a masked motion completion framework for HMP is a novel idea. By leveraging diffusion models and a masked completion technique, the authors provide a fresh perspective on motion prediction. Firstly, the proposed method simplifies the training process by using a single loss function and an end-to-end training process. This simplicity could make the approach more accessible and easier to implement compared to previous methods that require multiple loss constraints and complex hyper-parameter tuning. Secondly, HumanMAC demonstrates the ability to effectively switch between different motion categories, which is essential for applications like animation. This is achieved by holistically modeling the entire sequence of observed and predicted motions. The contributions are clear. The authors clearly outline their contributions, providing a solid foundation for further research. They also offer detailed discussions on the limitations of existing paradigms, which can inspire future work in this area.

Numerical results show that HumanMAC outperforms state-of-the-art methods in both qualitative and quantitative metrics. In both qualitative and visualization comparisons with the state-of-the-art methods, our method achieves superior performance, which creates a simple and strong baseline for future research. Comprehensive ablation studies and discussions are also provided.

In summary, "HumanMAC: Masked Motion Completion for Human Motion Prediction" is a significant contribution to the field of human motion prediction. It introduces a novel, simplified, and effective method for predicting human motions, addressing key issues in existing approaches. While there are areas for further exploration and validation, the proposed framework holds great potential for advancing the state of the art in motion prediction tasks.

References:

[1] Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, José MF Moura, and Manuela Veloso. Teaching robots to predict human motion. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 562–567. IEEE, 2018. 1

[2] Siheng Chen, Baoan Liu, Chen Feng, Carlos VallespiGonzalez, and Carl Wellington. 3d point cloud processing and learning for autonomous driving: Impacting map creation, localization, and perception. IEEE Signal Processing Magazine, 38(1):68–86, 2020. 1.

[3] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Bjorn Ommer. Behavior-driven synthesis of human dynamics. In CVPR, pages 12236–12246, 2021. 1, 2

[4] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In ACM MM, pages 5162–5171, 2022. 1, 2, 5, 15

[5] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatialtemporal anchors. In ECCV, 2022. 1, 2

[6] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In ICCV, 2021. 1

[7] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In CVPR, pages 336–345, 2017. 1, 2

[8] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In CVPR Workshops, pages 1418–1427, 2018. 1

[9] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. arXiv preprint arXiv:2211.14304, 2022. 1, 2, 3, 5, 15

[10] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In ECCV, 2020. 1.

Jianqi Zhong, received the M.S. in Electronic Information Engineering from Shenzhen University, China, in 2017. He is currently pursuing a doctor's degree in information and communication engineering with Shenzhen University, Shenzhen, China. His research interests include computer vision and deep learning.

MMTC Communications – Review Editorial Board

Multimedia Communications Technical Committee Officers

MMTC examines systems, applications, services and techniques in which two or more media are used in the same session. These media include, but are not restricted to, voice, video, image, music, data, and executable code. The scope of the committee includes conversational, presentational, and transactional applications and the underlying networking systems to support them.