Letter

A Local-Global Attention Fusion Framework With Tensor Decomposition for Medical Diagnosis

Peishu Wu, Han Li, Liwei Hu, Jirong Ge, and Nianyin Zeng

Dear Editor,

In this letter, a novel hierarchical fusion framework is proposed to address the imperfect data property in complex medical image analysis (MIA) scenes. In particular, by combining the strengths of convolutional neural networks (CNNs) and transformers, the enhanced feature extraction, spatial modeling, and sequential context learning are realized to provide comprehensive insights on the complex data patterns. Integration of information in different level is enabled via a multi-attention fusion mechanism, and the tensor decomposition methods are adopted so that compact and distinctive representation of the underlying and high-dimensional medical image features can be accomplished [1]. It is shown from the evaluation results that the proposed framework is competitive and superior as compared with some other advanced algorithms, which effectively handles the imperfect property of inter-class similarity and intra-class differences in diseases, and meanwhile, the model complexity is reduced within an acceptable level, which benefits the deployment in clinic practice.

MIA has assumed a pivotal role in numerous critical clinical scenarios, where sophisticated image analysis techniques have proven instrumental in augmenting medical decision-making, facilitating individualized therapeutic interventions, and enhancing patient prognostication [2]−[4]. In this regard, it is of vital significance to develop the sophisticated and robust intelligent computational methodologies, which are expected to facilitate quantitative assessments of anatomical structures and pathological changes via extracting informative features and recognizing intricate patterns from the medical images.

The prominent challenges in the analysis of clinical data stems from its intricate nature and diverse characteristics. To be specific, during the image acquisition, the inherent variability in the equipment, environment and patient conditions makes it hard to achieve consistent and robust results [5], [6], where the presence of noise, artifacts, and anatomical variations even further raises the difficulty for accurate analysis. Then, the data-driven methods are always restricted by the scarcity of annotated data, especially for the rare conditions in a specialized domain, and it is also a challenging task to develop a generalized model that can adapt to variations among different data. Moreover, the computational cost is another important concern for practical deployment, where too much attention on the computation efficiency may potentially sacrifice the accuracy. To handle above challenging issues, there is an urgent need to realize a seamless integration of domain knowledge, feature fusion strategy, and advanced model compression techniques [7].

Corresponding author: Nianyin Zeng.

Citation: P. Wu, H. Li, L. Hu, J. Ge, and N. Zeng, "A local-global attention fusion framework with tensor decomposition for medical diagnosis," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 6, pp. 1536–1538, Jun. 2024.

P. Wu, H. Li, L. Hu, and N. Zeng are with the Department of Instrumental and Electrical Engineering, Xiamen University, Fujian 361005, China (e-mail: wupeishu@stu.xmu.edu.cn; hanlix@stu.xmu.edu.cn; huliwei@stu.xmu.edu.cn; zny@xmu.edu.cn).

J. Ge is with the Key Research Laboratory of Osteoporosis Syndrome Genomics, Fujian Academy of Chinese Medical Sciences, Fuzhou 350003, China (e-mail: fjszyykxy@fjtcm.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2023.124167

In this letter, a novel global-local attention fusion framework with tensor decomposition (GLA-TD) is developed, which consists of a dual-attention mechanism-based fusion branch, a CNN-based local and a transformer-based global feature extraction branch. By capturing anatomical structures and subtle pathological changes at multiple scales, the feature representation can be effectively enhanced, which facilitates improving the generalization and robustness of the model. Moreover, the tensor decomposition technique is incorporated in the GLA-TD framework, which benefits realizing compact model structure and, by doing so, deployment in complex healthcare scenes with limited computation resources is enabled.

This study makes significant contributions in the following three key aspects: 1) An MIA-oriented framework GLA-TD is proposed, which leverages attention mechanisms and tensor decomposition to simultaneously capture both local and global features. 2) Advantages of CNN and transformer are sufficiently utilized to extract the semantic and detailed information. 3) Structural redundancy is effectively reduced via the adopted tensor decomposition method, which enhances the clinical practicality of the proposed framework.

Preliminaries: In the field of MIA, the combination of CNN and transformer architectures has emerged as a compelling paradigm [8], which integrates the advantages of CNN in capturing spatial hierarchies in images and the merits of transformer in modeling intricate contextual information in sequences. Consequently, it has paved a novel feasible path to simultaneously exploit the complex spatial characteristics and intricate global contextual dependencies in medical imaging data.

Given a sequence X, the self-attention mechanism is adopted in transformer to capture the global contextual relationships as

$$Attention(X) = Softmax\left(\frac{XW_Q(XW_K)^T}{\sqrt{d_k}}\right)XW_V \tag{1}$$

where XW_Q, XW_K, and XW_V are linear transformations, and $\sqrt{d_k}$ ensures the stability of gradient. In CNN, convolution and pooling operators are adopted to extract local features as

$$C(I) = I * K \tag{2}$$

where I is the input image tensor and K refers to the convolution kernel. To further enhance the computational efficiency and interpretability, the tensor decomposition method is employed to factorize above convolution kernel as

$$K = \sum_{i=1}^{r} U_i \otimes V_i \otimes W_i \tag{3}$$

where U_i, U_i and W_i are the factorized tensors, and r denotes the rank of decomposition. In addition, the weight matrix W in the fully connected layers decomposed as

$$W = C \times_1 A \times_2 B \tag{4}$$

where C represents the core tensor, A and B are projection matrices. In essence, the tensor decomposition applies higher-order representations to capture intricate relationships within data [9], which is adopted in the proposed GLA-TD to encapsulate the underlying data structure so as to enable efficient learning and generalization of the network.

Method: For a clear view, the overall structure of GLA-TD is illustrated in Fig. 1, and it is shown that three distinct branches are contained, namely the local, global, and fusion branch, respectively.

Firstly, the CNN-based local branch consists of 4 stages, including in total 16 large kernel attention (LKA) blocks with the patch embed operation. In LKA block, a macroscopic kernel is dissected into three components of depth-wise, dilation, and channel convolution, which encompasses local structural information and long-range dependencies, and moreover, the adaptability in channel dimension is also taken into consideration [10]. As a result, the merits of both convolu-

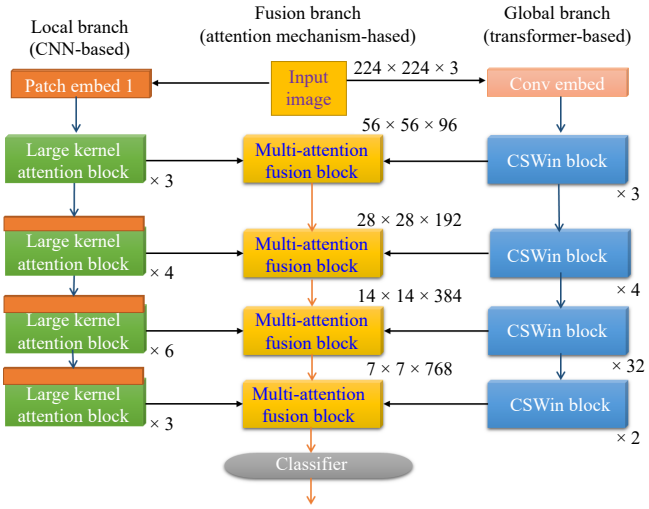

Fig. 1. Illustration of the proposed GLA-TD framework, where the orange rectangle represents the patch embed operation.

tion and self-attention are effectively integrated in the local branch.

Secondly, the transformer-based global branch is composed of series of the CSwin blocks [11], which is constructed by stacking two kinds of modules. One module involves the application of layer-norm and cross-shaped window self-attention, which are followed by a shortcut connection; the other one employs the layer-normalization (LN) and multi-layer perceptron (MLP) with residual connection. By embedding the CSwin blocks into global branch, feature extraction ability can be greatly improved without much sacrifice on the computations.

Last but not the least, as is shown in Fig. 2, the attention-based fusion branch applies a hierarchical structure to dynamically integrate features originated from different stages, including distinct local attributes L_i, comprehensive global features G_i, and semantic information derived from the previous layer fusion F_{i-1}.

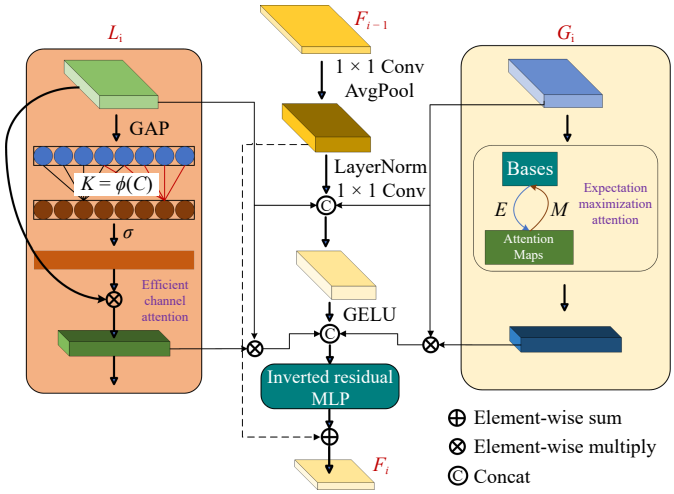

Fig. 2. Structure of the designed attention-based fusion branch.

In the fusion branch, advantages of diverse self-attention mechanism in capturing spatial and temporal information are fully exploited. By utilizing the efficient channel attention (ECA) mechanism [12], feature representation can be enhanced from different semantic aspects; subsequently, the local features are integrated into an expectation-maximization attention (EMA) mechanism [13] for further refinement, where influences of irrelevant regions will be alleviated. In the final feature fusion stage, a residual inverted MLP (RIMLP) is employed to capture subtle context-aware details related to the inter-class similarity and intra-class differences, which benefits extracting the highly discriminative features. Principles of above mentioned mechanisms are presented as follows:

$$ECA(x) = x \otimes \{\sigma[C1D_k(GAP(x))]\}, \text{ where } k = \frac{\log_3 C}{2} \tag{5}$$

$$EMA(x) = \text{Relu}\{x \oplus C2D_1(\text{Relu}(z \otimes k))\} \tag{6}$$

$$RIMLP(x) = C2D_1(C2D_1(C2D_3(LN(x)) + LN(x))) \tag{7}$$

where CmD_n represents the $n \times n$ convolution with the dimension of m, GAP refers to the global average pooling. σ is the sigmoid function, and LN is the layer-normalization operation. z and k indicate the outputs of expectation and maximization, respectively.

In addition, the block term (BT) method [14] is employed within the GLA-TD framework to optimize both the efficiency and interpretability of convolutional operations. This method entails decomposing convolution filters into low-rank block term components. Consequently, it significantly diminishes computational complexity while faithfully preserving crucial spatial relationships. Particularly, the BT convolutional layers carry out the Tucker decomposition [15] to decompose the convolution kernel into various modes as

$$\mathcal{Y}_{j_1,\ldots,j_M} = \sum_{i_1,\ldots,i_N=1}^{I_1,\ldots,I_N} \sum_{c}^{C} \sum_{r_1,\ldots,r_N=1}^{R_1,\ldots,R_N} g_{r1},\ldots,r_N a^{(1)}_{i_1,c,r_1} \ldots a^{(N)}_{i_N,c,r_N} a^{(N+1)}_{j_1,c,r_{N+1}} \ldots a^{(N+M)}_{j_M,c,r_{N+M}} x_{i1,i2},\ldots,iN \tag{8}$$

where x and $\mathcal{Y}$ are the input and output tensors, respectively, C represents the CP-rank (short for Candecomp/Parafac), and $R_1, R_2, \ldots, R_N$ are the Tucker-ranks, which can take different values.

Results and discussions: In this section, the proposed GLA-TD is comprehensively evaluated on some challenging MIA tasks, and the results are compared with other state-of-the-art deep learning models to further validate the competitiveness of our method, which are named ConvNext [16], T2T-ViT [17], VGG-19 [18], Conformer [19], HiFuse [20], and ShuffleNet V2 [21]. For fairness, experiments are carried out under the same conditions, where the MIA datasets Kvasir (https://datasets.simula.no/kvasir/) and PALM (https://palm.grand-challenge.org/) are adopted for model evaluation.

In MIA tasks, multiple metrics are employed to assess the comprehensive performance of the proposed GLA-TD model. Specifically, *Accuracy* gauges the overall accuracy of predictions, *Precision* quantifies the fraction of true positive predictions among all positive instances, *Recall* evaluates the ratio of true positive samples among all actual positives, and the $F1$ offers a balanced measure by calculating the harmonic mean of precision and recall to appraise the classification performance.

Benchmark evaluation results on Kvasir dataset are reported in Table 1, and in this task, the models are required to analyze data of 4000 endoscopic gastrointestinal diseases in eight classes. As is shown, the proposed GLA-TD outperforms other advanced models on all metrics, whose *Accuracy*, *Precision*, *Recall* and $F1$ reach 90.63%, 89.58%, 91.24% and 90.76%, respectively. The superior performance demonstrated in the aforementioned outcomes can be attributed primarily to the innovative global-local feature extraction architecture and the fusion method employed within GLA-TD, rendering it a formidable contender in the field of MIA.

In Table 2, the benchmark evaluation results on PALM dataset are displayed, where the models are expected to distinguish the ocular fundus images with high myopia character from those normal ones, and in total 800 images are included. According to the results, the proposed GLA-TD model presents significant advantages in recognizing the myopia samples, which yields the *Accuracy*, *Precision*, *Recall*, and $F1$ of 98.75%, 97.26%, 99.38%, and 98.31%, respectively. It is noticeable that a remarkable precision of 97.26% is achieved by our GLA-TD, indicating a minimal incidence of misdiagnosis, thereby effectively mitigating the risk of unwarranted treatments for individuals with mild myopia and enhancing patient care. Moreover, the recall of 99.38% shows that the proposed model also yields low false negative rate, which implies that our GLA-TD can almost identify all cases of high myopia so as to provide valuable diagnostic reference to the clinicians.

In addition, an ablation study is carried out to investigate the influ-

Table 1. Benchmark Evaluation Results on the Kvasir Dataset

	Accuracy	*Precision*	*Recall*	*F*1
ConvNext [16]	74.60	74.78	74.64	74.61
T2T-ViT [17]	76.90	77.60	76.91	76.78
VGG-19 [18]	77.75	77.86	77.83	77.75
Conformer [19]	84.25	84.45	84.37	84.27
HiFuse [20]	85.00	85.08	85.00	84.96
ShuffleNet V2 [21]	87.12	87.38	87.26	87.32
GLA-TD (ours)	**90.63**	**89.58**	**91.24**	**90.76**

Table 2. Benchmark Evaluation Results on the Palm Dataset

	Accuracy	*Precision*	*Recall*	*F*1
ConvNext [16]	92.83	91.96	94.15	93.04
T2T-ViT [17]	94.48	95.28	94.54	94.91
VGG-19 [18]	93.56	94.25	95.73	94.98
Conformer [19]	95.12	93.73	94.86	94.29
HiFuse [20]	95.32	94.69	93.88	94.28
ShuffleNet V2 [21]	96.50	96.19	97.25	96.72
GLA-TD (ours)	**98.75**	**97.26**	**99.38**	**98.31**

ences of the adopted tensor decomposition method, and the results are presented in Table 3, where *K* and *P* stand for the Kvasir and PALM dataset, respectively. Based on Table 3, it is found that when tensor decomposition method is introduced, there is a significant reduction of the parameter number, which declines from 467.27 M to 135.04 M, and the floating-point operations per second (FLOPs) even declines by nearly 70.58%. Simultaneously, a marginal sacrifice in accuracy, averaging just 0.22% across both datasets is observed, which underscores the effectiveness of the employed tensor decomposition method in harmonizing model complexity and accuracy. Consequently, deploying the proposed GLA-TD in resource-constrained environments becomes feasible.

Table 3. The Performance of Tensor Decomposition Methods

	Accuracy K	*Accuracy P*	*Params*	*FLOPs*
GLA	90.63	98.80	467.27 M	73.89 G
GLA-TD	90.25	98.75	135.04 M	21.74 G

Conclusion: In this letter, a local-global attention fusion framework with TD has been proposed for MIA tasks. Based on the CNN-based local and transformer-based global branches, the model is able to learn from rich detailed and semantic information. An efficient local-global fusion branch with attention mechanism has been designed to enhance the focus on key regions and channels. Moreover, the tensor decomposition technique has been incorporated in the convolution and linear operations to balance the model accuracy and computational cost. Evaluation results have shown that the proposed GLA-TD can effectively deal with the complex MIA tasks with considerable performance, which can provide valuable diagnostic references in clinic. In future work, we aim to 1) Apply the GLA-TD to industrial defect detection; 2) Investigate multi-modal image analysis; 3) Further optimize the model performance with neural architecture search.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China (62073271), the Fundamental Research Funds for the Central Universities of China (20720220076), and the Natural Science Foundation for Distinguished Young Scholars of the Fujian Province of China (2023 J06010).

References

[1] L. Sun, Z. Liu, X. Sun, L. Liu, R. Lan, and X. Luo, "Lightweight image super-resolution via weighted multi-scale residual network," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 7, pp. 1271–1280, 2021.

[2] H. Li, N. Zeng, P. Wu, and K. Clawson, "Cov-Net: A computer-aided diagnosis method for recognizing COVID-19 from chest X-ray images via machine vision," *Expert Systems Applications*, vol. 207, p. 118029, 2022.

[3] H. Li, P. Wu, Z. Wang, J. Mao, F.-E. Alsaadi, and N. Zeng, "A generalized framework of feature learning enhanced convolutional neural network for pathology-image-oriented cancer diagnosis," *Computers in Biology and Medicine*, vol. 151, p. 106265, 2023.

[4] P. Wu, Z. Wang, B. Zheng, H. Li, F.-E. Alsaadi, and N. Zeng, "AGGN: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion," *Computers in Biology and Medicine*, vol. 152, p. 106457, 2023.

[5] Y. Lei, H. Zhu, J. Zhang, and H. Shan, "Meta ordinal regression forest for medical image classification with ordinal labels," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1233–1247, 2022.

[6] D. Wu and X. Luo, "Robust latent factor analysis for precise representation of high-dimensional and sparse data," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 4, pp. 796–805, 2021.

[7] X. Luo, Y. Yuan, S. Chen, N. Zeng, and Z. Wang, "Position-transitional particle swarm optimization-incorporated latent factor analysis," *IEEE Trans. Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3958–3970, 2022.

[8] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, and X. Liu, "A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection," *IEEE Trans. Instrumentation and Measurement*, vol. 71, p. 3507014, 2022.

[9] L. Chen and X. Luo, "Tensor distribution regression based on the 3D conventional neural networks," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 7, pp. 1628–1630, 2023.

[10] Q. Cheng, Y. Zhou, H. Huang, and Z. Wang, "Multi-attention fusion and fine-grained alignment for bidirectional image-sentence retrieval in remote sensing," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 8, pp. 1532–1535, 2022.

[11] M. Guo, C. Lu, Z. Liu, M. Cheng, and S. Hu, "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. 35th IEEE/CVF Computer Vision and Pattern Recognition Conf.*, 2022, pp. 1–11.

[12] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. 33rd IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 11531–11539.

[13] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation", in *Proc. 23rd Int. Conf. Computer Vision*, 2019.

[14] G. Xu, X. Wang, and X. Xu, "Single image enhancement in sandstorm weather via tensor least square," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 6, pp. 1649–1661, 2020.

[15] F. Bi, X. Luo, B. Shen, H. Dong, and Z. Wang, "Proximal alternating-direction-method-of-multipliers-incorporated nonnegative latent factor analysis," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 6, pp. 1388–1406, 2023.

[16] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. 35th IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2022, pp. 11966–11976.

[17] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. Tay, J. Feng, and S. Yan, "Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet," in *Proc. 25th IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 538–547.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv: 1409.1556, 2015.

[19] Z. Peng, Z. Guo, W. Huang, Y. Wang, L. Xie, J. Jiao, Q. Tian, and Q. Ye, "Conformer: Local features coupling global representations for recognition and detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9454–9468, 2023.

[20] X. Huo, G. Sun, S. Tian, Y. Wang, L. Yu, J. Long, W. Zhang, and A. Li, "HiFuse: Hierarchical multi-scale feature fusion network for medical image classification," arXiv preprint arXiv: 2209.10218, 2022.

[21] N. Ma, X. Zhang, H. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. 15th European Conf. Computer Vision*, 2018, vol. 11218, pp. 122–138.