SCIENCE CHINA
Information Sciences

• LETTER •

June 2025, Vol. 68, Iss. 6, 160110:1–160110:2
https://doi.org/10.1007/s11432-024-4399-9

Special Topic: Enabling Techniques and Cutting-Edge Applications of Foundation Models

Visual and text prompt learning for multi-modal brain disease diagnosis

Yumiao ZHAO[1], Bo JIANG[1,2*], Yuhe DING[1], Xixi WAN[1] & Jin TANG[1]

[1]*School of Computer Science and Technology, Anhui University, Hefei 230601, China*
[2]*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China*

Received 31 October 2024/Revised 9 February 2025/Accepted 17 April 2025/Published online 21 May 2025

Citation Zhao Y M, Jiang B, Ding Y H, et al. Visual and text prompt learning for multi-modal brain disease diagnosis. Sci China Inf Sci, 2025, 68(6): 160110, https://doi.org/10.1007/s11432-024-4399-9

Recently, multi-modal neuroimaging data like positron emission tomography (PET) and magnetic resonance imaging (MRI) are widely employed in disease diagnosis tasks. PET can capture the functional information of the brain, while MRI can display the structure of the brain. These heterogeneous multi-modal data contain rich information that facilitates a more reliable brain disease diagnosis. To explore the complementary information between different modalities, some academics propose multi-modal fusion methods for brain disease diagnosis.

Overall, existing multi-modal fusion methods for brain disease diagnosis can be categorized into two types. The first type is image-level fusion, including spatial-level and frequency-level fusion. The second type is feature-level fusion. For example, Goel et al. [1] adapted the wavelet packet transform to reduce spatial distortion and achieved high-quality fusion between MRI and PET. Qiu et al. [2] concatenated modality-specific features from the last stage to fuse the multi-modal information. Lei et al. [3] adopted a feature induction learning technique to align multi-modal features within a unified feature space. Although there are several attempts to fuse data across different modalities for disease diagnosis, some unsolved challenges still exist for this task. (i) Failing to capture long-range relationships in 3D scanned images. Traditional methods typically use convolutional neural networks (CNNs) to extract the modality-specific features. However, due to the local nature of the convolutional operation, these methods fail to model long-range relationships among different slices of 3D scan images. (ii) Lack of integrated expert knowledge in brain image representation. Most existing methods overlook the importance of integrating expert knowledge, which provides valuable diagnostic cues in medical diagnosis tasks. Focusing more attention on these distinctive regions can enhance the performance of disease diagnosis, such as the hippocampus and temporal lobes for Alzheimer's disease. (iii) Challenges of missing modalities in clinical diagnosis. Modality missing frequently occurs in clinical diagnosis, making it challenging to apply existing multi-modal models directly for this task, as they typically require paired multi-modal data for inference.

Methods. To address the above challenges, in this study, we propose a novel multi-modal visual and text prompt learning framework (mVTPrompt) for brain disease diagnosis tasks. The overall framework is illustrated in Figure 1(a) and can be summarized as follows. (1) The Conv-Transformer3D network is used to fully model the long-range relationships among different slices of 3D scan images. (2) The region and text prompt learning module (RTPL) leverages expert knowledge to learn the semantic-aware local feature representation for disease diagnosis. (3) The cross-modality visual prompt learning module (CVPL) provides a general solution to fine-tune pre-trained unimodal models for multi-modal learning tasks, enabling the framework to support both unimodal and multi-modal inference.

Overview. For the raw data from different modalities (e.g., PET and MRI), mVTPrompt consists of three stages. In the first stage, we use unimodal datasets to train the Conv-Transformer3D network, obtaining pre-trained PET and MRI models. The extracted modality-specific global features are denoted as $\boldsymbol{f}_{\text{global}}^{\text{pet}}$ and $\boldsymbol{f}_{\text{global}}^{\text{mri}}$. To maximally retain the modality-specific feature representations and mitigate modality competition during fine-tuning, the pre-trained models are frozen. In the second stage, we use a multi-modal dataset and add a CVPL to fine-tune the pre-trained PET and MRI models for adapting the multi-modal task. The global multi-modal feature representation is represented as $\boldsymbol{f}_{\text{global}}^{\text{mu}}$. To extract semantic-aware local feature representations from expert knowledge, we integrate the RTPL into the multi-modal branch. The local feature representation is denoted as $\boldsymbol{F}_{\text{local}}^{\text{mu}}$. In the third stage, we add the RTPL into the pre-trained PET and MRI models to obtain the local feature representation of the unimodal denoted as $\boldsymbol{F}_{\text{local}}^{\text{pet}}$ and $\boldsymbol{F}_{\text{local}}^{\text{mri}}$. Moreover, a simple yet effective decision-level fusion method is employed to fuse the local and global features from each branch in the proposed framework, achieving a more comprehensive diagnosis of brain disease while supporting inputs from any modality.

Conv-Transformer3D. As shown in Figure 1(b), we propose Conv-Transformer3D for brain neuroimaging analysis.

* Corresponding author (email: jiangbo@ahu.edu.cn)

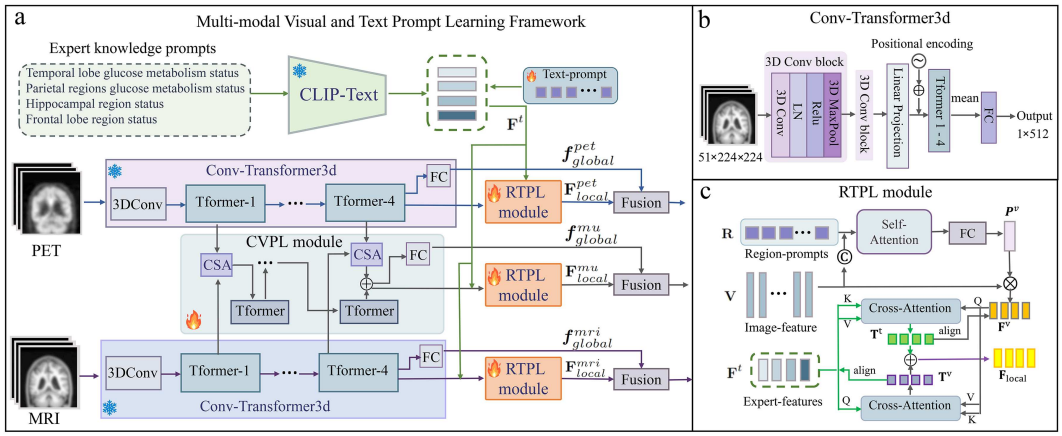

Figure 1 (Color online) (a) The details of the mVTPrompt, where "Tformer" refers to a transformer block; (b) Conv-Transformer3D backbone network to extract modality-specific features; (c) RTPL module incorporates expert knowledge to learn semantic-aware local feature representation.

Brain neuroimaging data often exhibits high spatial redundancy and noise, and a spatial smoothing operation is required during preprocessing. Inspired by this, we design Conv-Transformer3D, which integrates 3D convolution with transformer blocks for efficient feature extraction. Specifically, we first employ two 3D convolution blocks to extract low-level features and reduce the dimensionality of feature maps. This step not only enhances the robustness of feature representations by suppressing noise but also reduces the computational cost. Next, we utilize transformer blocks to capture the long-range relationship among all feature tokens from different slices, enabling more comprehensive modeling of spatial relationships within the brain. The details are presented in Appendix A.1.

CVPL. To effectively adapt pre-trained unimodal models to multi-modal tasks, we propose a CVPL, a lightweight cross-modal tuning module. A key component of CVPL is the consistency-constrained self-attention (CSA) block, which is designed to effectively exploit complementary information between pre-trained PET and MRI features. We add CSA at each pre-trained modality transformer block to progressively model stage-wise interaction between PET and MRI modalities. The details can be found in Appendix A.2.

RTPL. In brain disease diagnosis, only specific neuroimaging regions are strongly associated with the disease. Therefore, it is meaningful to obtain the feature representation of these distinctive regions. Expert knowledge usually provides critical information for brain disease diagnosis. Building upon this medical knowledge, we propose RTPL by introducing expert knowledge into the visual learning network to capture the semantic-aware local features representation for brain disease diagnosis. Specifically, we first utilize a CLIP model [4] to encode the hand-crafted expert knowledge prompts. Then, we introduce text and region prompts to capture the task-specific text and visual feature representation. Finally, bidirectional cross-attention is employed to enhance cross-modal feature interaction and alignment between text and visual features. The pipeline is shown in Figure 1(c), and details are provided in Appendix A.3.

Experiments. To evaluate the effectiveness and advantages of the proposed mVTPrompt, we conduct experiments under both complete and (modality) missing settings on multi-modal brain disease datasets, including ADNI and ABIDE. For quantitative comparison, we first evaluate the effectiveness and efficiency of the Conv-Transformer3D network on the ADNI dataset. Next, we compare our method with state-of-the-art (SOTA) multi-modal approaches under both complete and missing modality settings. These results show that the proposed method achieves superior performance on disease prediction tasks while effectively supporting flexible unimodal/multimodal inference. Detailed comparison results can be found in Appendix B.3. To further analyze the contribution of different components of the proposed method, we perform ablation experiments on the ADNI and ABIDE datasets. Additionally, to evaluate the efficiency of the RTPL, we use Grad-CAM to visualize the activation maps and analyze the effect of varying the number of expert knowledge prompts. The details of the comparison results can be found in Appendix B.4.

Conclusion. We propose a novel training-efficient mVTPrompt for brain disease diagnosis. The proposed method supports the input of complete modality and missing modality data. In our experiments, we adopt a simple method to fuse the global and local information. In the future, we aim to explore more efficient fusion methods for the local and global features, while also investigating a new method to address the challenge of missing modalities during the training and test phases.

Acknowledgements This work was supported by Natural Science Foundation of Anhui Province (Grant No. 2408085J037) and University Synergy Innovation Program of Anhui Province (Grant No. GXXT-2022-032).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Goel T, Sharma R, Tanveer M, et al. Multimodal neuroimaging based Alzheimer's disease diagnosis using evolutionary RVFL classifier. IEEE J Biomed Health Inform, 2024. doi: 10.1109/JBHI.2023.3242354
2 Qiu Z, Yang P, Xiao C, et al. 3D multimodal fusion network with disease-induced joint learning for early Alzheimer's disease diagnosis. IEEE Trans Med Imag, 2024, 43: 3161–3175
3 Lei B, Li Y, Fu W, et al. Alzheimer's disease diagnosis from multi-modal data via feature inductive learning and dual multilevel graph neural network. Med Image Anal, 2024, 97: 103213
4 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763