Visual-Attribute Prompt Learning for Progressive Mild Cognitive Impairment Prediction

Luoyao Kang[1,2], Haifan Gong[1,2], Xiang Wan[1], and Haofeng Li[1(✉)]

[1] Shenzhen Research Institute of Big Data, Shenzhen, China
lhaof@sribd.cn
[2] The Chinese University of Hong Kong, Shenzhen, China

Abstract. Deep learning (DL) has been used in the automatic diagnosis of Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) with brain imaging data. However, previous methods have not fully exploited the relation between brain image and clinical information that is widely adopted by experts in practice. To exploit the heterogeneous features from imaging and tabular data simultaneously, we propose the Visual-Attribute Prompt Learning-based Transformer (VAP-Former), a transformer-based network that efficiently extracts and fuses the multi-modal features with prompt fine-tuning. Furthermore, we propose a Prompt fine-Tuning (PT) scheme to transfer the knowledge from AD prediction task for progressive MCI (pMCI) diagnosis. In details, we first pre-train the VAP-Former without prompts on the AD diagnosis task and then fine-tune the model on the pMCI detection task with PT, which only needs to optimize a small amount of parameters while keeping the backbone frozen. Next, we propose a novel global prompt token for the visual prompts to provide global guidance to the multi-modal representations. Extensive experiments not only show the superiority of our method compared with the state-of-the-art methods in pMCI prediction but also demonstrate that the global prompt can make the prompt learning process more effective and stable. Interestingly, the proposed prompt learning model even outperforms the fully fine-tuning baseline on transferring the knowledge from AD to pMCI.

Keywords: Alzheimer's disease · Prompt learning · Magnetic resonance imaging · Multi-modal classification · Transformer · Attention modeling

1   Introduction

Alzheimer's disease (AD) is one of the most common neurological diseases in elderly people, accounting for 50–70% of dementia cases [31]. The progression of AD triggers memory deterioration, impairment of cognition, irreversible loss of neurons, and further genetically complex disorders as well. Mild Cognitive Impairment (MCI), the prodromal stage of AD, has been shown as the optimal stage to be treated to prevent the MCI-to-AD conversion [29]. Progressive MCI (pMCI) group denotes those MCI patients who progress to AD within 3 years, while stable MCI (sMCI) patients remained stable over the same time period. pMCI is an important group to study longitudinal changes associated with the development of AD [26]. By predicting the pMCI progress of the patients, we can intervene and delay the progress of AD. Thus, it is valuable to distinguish patients with pMCI from those with sMCI in the early stage with a computer-aided diagnosis system.

In recent years, deep learning (DL) based methods [2,8,11,13] have been widely used to identify pMCI or AD based on brain MRI data. Several works [9,24,29] take both brain MRI and clinical tabular data into account, using convolutional neural networks (CNNs) and multi-layer perceptions as the feature encoder. Due to the limited data in pMCI diagnosis, some works [19,29] resort to transfer learning to fine-tune the model on the pMCI-related task, by pre-training weights on the AD detection task. Previous CNN-based approaches [9,24,29] may fail on the lack of modeling long-range relationship, while the models based on two-stage fine-tuning [19,29] are inefficient and may cause networks to forget learned knowledge [32].

Inspired by the advance in transformers [6,17,28] and prompt learning [4,15, 25], we tailor-design an effective multi-modal transformer-based framework based on prompt learning for pMCI detection. The proposed framework is composed of a transformer-based visual encoder, a transformer-based attribute encoder, a multi-modal fusion module, and a prompt learning scheme. Clinical attributes and brain MR images are sent to the attribute encoder and visual encoder with the prompt tokens, respectively. Then the high-level visual and tabular features are aggregated and sent to fully-connected layers for final classification. For the proposed prompt tuning scheme, we first pre-train the neural network on the AD classification task. After that, we fine-tune the neural network by introducing and updating only a few trainable parameters. Importantly, we observe that the number of image patches and input tokens have a gap between 2D natural images and 3D MR images, due to the different dimensions. To complement local interactions between prompt and other tokens, we develop a global prompt token to strengthen the global guidance and make the visual feature extraction process more efficient and stable.

Our contributions have three folds: (1) We propose a visual-attribute prompt learning framework based on transformers (VAP-Former) for pMCI detection. (2) We design a global prompt to adapt to high-dimension MRI data and build a prompt learning framework for transferring knowledge from AD diagnosis to pMCI detection. (3) Experiments not only show that our VAP-Former obtains

state-of-the-art results on pMCI prediction task by exceeding the full fine-tuning methods, but also verify the global prompt can make the training more efficient and stable.
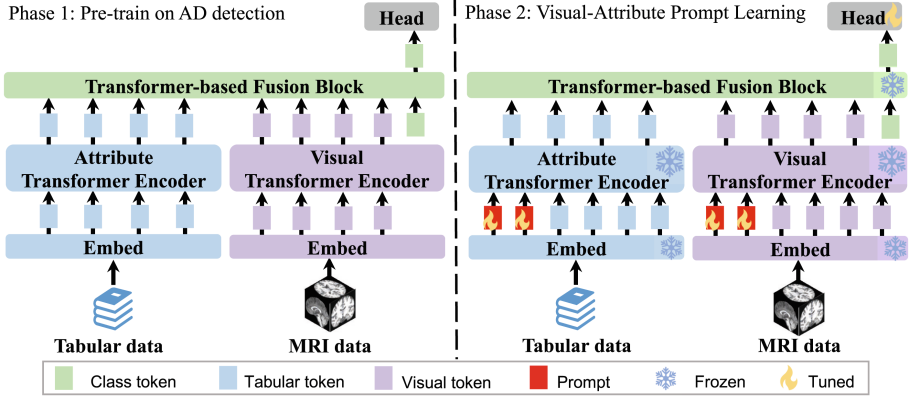
Fig. 1. The proposed Visual-Attribute Prompt learning transformer framework. The left part shows the architecture of VA-Former without using prompts. The right part shows the proposed framework with the prompt tuning strategy (VAP-Former) that only updates the learnable prompts.

2   Methodology

We aim to predict if an MCI patient will remain stable or develop Alzheimer's Disease, and formulate the problem as a binary classification task based on the brain MR image and the corresponding tabular attribute information of an MCI patient from baseline examination. As Fig. 1 shows, we adopt the model weights learned from AD identification to initialize the prediction model of MCI conversion. In the prompt fine-tuning stage, we keep all encoders frozen, only optimizing the prompts concatenated with feature tokens.

2.1   Network Architecture

We propose a transformer-based framework for pMCI prediction (VAP-Former) based on visual and attribute data, which is shown in Fig. 1. VAP-Former is mainly composed of three parts: a visual encoder for processing MRI data, an attribute encoder for processing attribute data, and a transformer-based fusion block for combining the multi-modal feature. Considering that capturing the long-range relationship of MRI is important, we employ the encoder of 3D UNETR++ [28] as our visual encoder. For the attribute encoder, since the clinical variables have no order or position, we embed the tabular data with the transformer blocks [30]. Followed by [6,7], we prepend a class token for dual-modality before the transformer-based fusion block. A class token is a learnable vector concatenated with dual-modal feature vectors. The class token of dual-modal is further processed by fully-connected layers for the final classification.
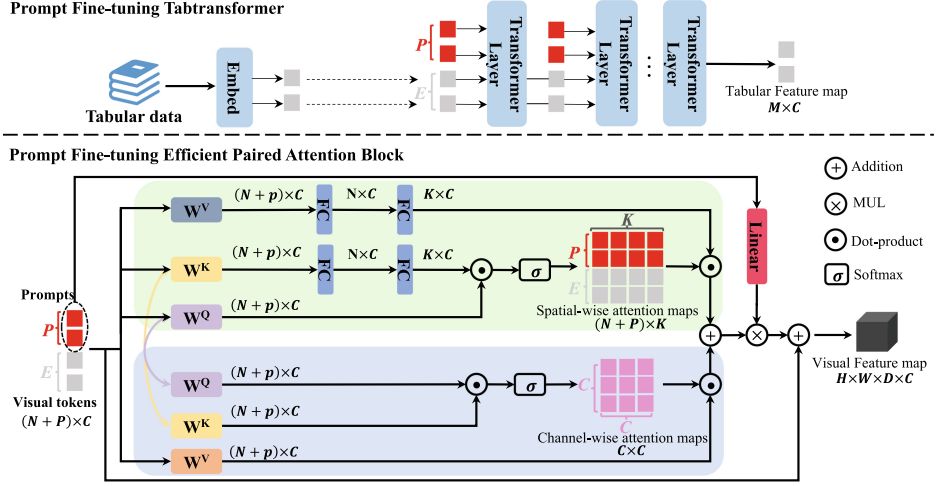
Fig. 2. The proposed global prompt-based learning strategy. The strategy replaces the attribute encoder with the Prompt Fine-tuning Tab-transformer (upper part) and replaces the Efficient Paired Attention (EPA) block with the prompt fine-tuning EPA block (lower part).

2.2 Knowledge Transfer with Multi-modal Prompt Learning

To effectively transfer the knowledge learned from AD prediction task to the pMCI prediction task, we propose a multi-modal prompt fine-tuning strategy that adds a small number of learnable parameters (i.e., prompt tokens) to the input of the transformer layer and keeps the backbone frozen. The overall pipeline is shown in Fig. 2, where the upper part indicates adding the prompt tokens to the attribute transformer, while the lower part denotes sending the prompt tokens to the visual transformer.

Tabular Context Prompt Learning. In the attribute encoder, we insert prompts into every transformer layer's input [15]. For the (i)-th Layer L_i of SA transformer block, we denote collection of p prompts as $P = \{p^k \in \mathbb{R}^C \| k \in \mathbb{N}, 1 \leq k \leq p\}$, where k is the number of the tabular prompt. The prompt fine-tuning Tabtransformer can be formulated as:

$$\begin{aligned} [\_\_, X_i] &= L_i([P_{i-1}, X_{i-1}]), \\ [\_\_, X_{i+1}] &= L_{i+1}([P_i, X_i]), \end{aligned} \tag{1}$$

where $X_i \in \mathbb{R}^{M \times C}$ denotes the attribute embedding at L_i's output space and P_i denotes the attribute prompt at L_{i+1}'s input space concatenated with X_i.

Visual Context Prompt Learning with Paired Attention. For the visual prompt learning part, we concatenate a small number of prompts with visual embedding to take part in the spatial-wise and channel-wise attention block [12,16,18] denoted as the prompt fine-tuning efficient paired attention

block in Fig. 2. Within the prompt fine-tuning efficient paired attention block, we insert shared prompts into the spatial-wise attention module (SWA) and the channel-wise attention module (CWA), respectively. With a shared keys-queries scheme, queries, keys, and values are noted as Q_{shared}^p, K_{shared}^p, $V_{spatial}^p$, and $V_{channel}^p$. Let $[\cdot, \cdot]$ be the concatenation operation, the SWA and CWA can be formulated as:

$$
\begin{aligned}
[P_{spatial}, I_S] &= SWA(Q_{shared}^p, K_{shared}^p, V_{spatial}^p), \\
[P_{channel}, I_C] &= CWA(Q_{shared}^p, K_{shared}^p, V_{channel}^p),
\end{aligned}
\tag{2}
$$

where $I_S \in \mathbb{R}^{N \times C}$ and $P_{spatial} \in \mathbb{R}^{\frac{P}{2} \times C}$ are spatial-wise visual embedding and prompts, and $I_C \in \mathbb{R}^{N \times C}$ and $P_{channel} \in \mathbb{R}^{\frac{P}{2} \times C}$ are channel-wise visual embedding and prompts, and P is the number of visual prompt. After that, the initial feature map I is added to the attention feature map using a skip connection. Let $+$ be the element-wise addition, this process is formulated as $I = I + (I_S + I_C)$.

2.3   Global Prompt for Better Visual Prompt Learning

Compared to natural images with relatively low dimensionality (e.g., shaped $224 \times 224 \times 3$) and the salient region usually locates in a small part of the image, brain MRIs for diagnosis of Alzheimer's are usually high dimensional (e.g., shaped $144 \times 144 \times 144$) and the cues to diagnosis disease (e.g., cortical atrophy and beta protein deposition) can occupy a large area of the image. Therefore, vanilla prompt learning [32] methods that are designed for natural images may not be directly and effectively applied to MRI's Recognition of Alzheimer's disease. We consider that the above differences lead to the following two problems: (1) vanilla prompt token often focuses on local visual information features; (2) the interaction between vanilla prompt token and visual feature is insufficient. Therefore, we consider that a sophisticated prompt learning module should be able to address the above-mentioned issues with the following feature: (1) the prompt token can influence the global feature; (2) the prompt token can effectively interact with the visual input features. A simple approach to achieve the second goal is to increase the number of prompt tokens so that they can better interact with other input features. However, the experiment proves that this is not feasible. We think it is because too many randomly initialized prompt tokens (i.e., unknown information) will make the model hard to train.

Thus, we tailor-design a global prompt token g to achieve the above two goals. Specifically, we apply a linear transformation T to the input prompt tokens P to obtain the global prompt token g (i.e., a vector), which further multiplies the global feature map. Since the vector is directly multiplied with the global feature, we can better find the global feature responses in each layer of the visual network, which enables the model to better focus on some important global features for pMCI diagnosis, such as cortical atrophy. Since this linear transformation, T operation is learnable in the prompt training stage, the original prompt token can better interact with other features through the global token. To embed the

global prompt token into our framework, we rewrite Eq. 3 as:

$$I = I + (I_S + I_C) \times T([P_{spatial}, P_{channel}]), \qquad (3)$$

where T denotes the linear transformation layer with $P \times C$ elements as input and $1 \times C$ element as output. Experiments not only demonstrate the effectiveness of the above method but also make the training stage more stable.

3   Experiments and Results

3.1   Datasets and Implementation Details

The datasets are from Alzheimer's Disease Neuroimaging Initiative (ADNI) [14], including ADNI-1 and ADNI-2. Following [19,22], we adopt ADNI-1/ADNI-2 as the train/test set. The subjects that exist in both datasets are excluded from ADNI-2. There are 1340 baseline T1-weighted structure MRI scans of 4 categories including Alzheimer's disease (AD), normal control (NC), stable mild cognitive impairment (sMCI), and progressive mild cognitive impairment (pMCI). 707 subjects (158 AD, 193 NC, 130 pMCI, 226 sMCI) are from ADNI-1, while 633 subjects (137 AD, 159 NC, 78 pMCI, 259 sMCI) are from ADNI-2. We pre-process the MR image as [22]. All MRI scans are pre-processed via 4 steps: (1) motion correction, (2) intensity normalization, (3) skull stripping, and (4) spatial normalization to a template. We use the Computational Anatomy Toolbox (CAT12)[1] via Statistical Parametric Mapping software (SPM12)[2] to perform the above procedures. Then all images are re-sampled as the size of $113 \times 137 \times 113$ and the resolution of $1 \times 1 \times 1\,mm^3$. For tabular clinical data, we select 7 attributes including age, gender, education, ApoE4, P-tau181, T-tau, and a summary measure (FDG) derived from 18F-fluorodeoxyglucose PET imaging. For tabular clinical data, we apply one-hot encoding to the categorical variables and min-max normalization to the numerical variables. We adopt the model weights learned from AD identification to initialize the prediction model of MCI conversion. Our model is trained using 1 NVIDIA V100 GPU of 32GB via AdamW optimizer [20] with 30 epochs for AD identification and 20 epochs for pMCI detection. The batch size is set to 4. We adopt the ReduceLROnPlateau [1] learning rate decay strategy with an initial learning rate of 1e-5. The loss function is binary cross-entropy loss. The number of visual and tabular prompts is 10 and 5, respectively. We take F1-score [23], class balanced accuracy (BACC) [3], and the area under the receiver operating characteristic curve (AUC) [23] as evaluation metrics.

3.2   Comparison with the State-of-the-Art Methods

To validate the proposed VAP-Former and prompt fine-tuning (PT) strategy, we compare our model with three unimodal baselines: 1) UNETR++, which denotes

---

[1] https://neuro-jena.github.io/cat/.
[2] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/.

Table 1. Comparison with existing pMCI classification methods. 'FT' represents full fine-tuning without prompts and 'PT' represents prompt fine-tuning. '# params' denotes the number of parameters being tuned in the fine-tuning stage.

Method	Modal		Fine-tuning	#params (M)	BACC	F1	AUC
	Vis	Tab					
UNETR++ [28]	✓	✗	FT	64.59	$65.94_{\pm1.19}$	$47.42_{\pm1.20}$	$70.46_{\pm1.67}$
Tabformer [21]	✗	✓	FT	27.13	$78.05_{\pm0.52}$	$60.29_{\pm0.27}$	$84.26_{\pm0.17}$
4-Way Classifier [27]	✓	✗	FT	12.27	$71.86_{\pm2.07}$	$53.49_{\pm3.44}$	$75.81_{\pm1.01}$
DFAF [10]	✓	✓	FT	69.32	$77.14_{\pm0.35}$	$59.66_{\pm0.85}$	$84.41_{\pm0.49}$
HAMT [5]	✓	✓	FT	71.83	$75.07_{\pm0.91}$	$55.58_{\pm1.11}$	$84.23_{\pm0.31}$
DAFT [24]	✓	✓	FT	67.89	$78.06_{\pm0.55}$	$62.92_{\pm0.71}$	$85.33_{\pm0.41}$
VA-Former	✓	✓	FT	70.19	$78.29_{\pm0.52}$	$62.93_{\pm0.29}$	$84.77_{\pm0.35}$
VAP-Former	✓	✓	PT	0.59	$\mathbf{79.22_{\pm0.58}}$	$\mathbf{63.13_{\pm0.11}}$	$\mathbf{86.31_{\pm0.25}}$

the encoder of UNETR++ [28] only using MRI data as input, 2) Tabformer [21], which only uses tabular data and is similar to the attribute encoder work in our model. 3) 4-Way Classifier [27] which used 3D DenseNet as the backbone to construct Alzheimer's disease diagnosis model only using MRI data as input. To further evaluate the efficiency of our model and the proposed PT strategy, we integrate the DAFT [24], HAMT [5], and DFAF [10] into the same attribute and visual encoder for a fair comparison. And we fine-tune the proposed model with two strategies including full fine-tuning (FT) and prompt tuning (PT).

In Table 1, the proposed model with PT strategy achieves 79.22% BACC, 63.13% F1, and 86.31% AUC. VAP-Former and VA-Former outperform all unimodal baselines in all metrics, indicating that our model can effectively exploit the relation between MRI data and tabular data to improve the prediction of pMCI. By using the PT strategy, the VAP-Former outperforms the second-best model, DAFT, by 1.16% BACC, 0.21% F1, and 0.98% AUC, demonstrating that the proposed PT strategy can efficiently adapt learned knowledge to pMCI detection and even achieves better classification results. Besides that, VAP-Former achieves the best results by tuning the minimum number of parameters.

3.3 Ablation Study and Investigation of Hyper-parameters

In this section, we study the internal settings and mechanism of the proposed PT strategy. First, we investigate how the prompts affect the VAP-Former performance. So we remove the visual prompts from the VAP-Former (denoted as TabPrompt in Table 2) and tabular prompts (denoted as VisPrompt), respectively. Compared with VA-Former, VisPrompt outperforms it by 0.07% AUC and TabPrompt degrades the performance. However, VAP-Former, which combines tabular prompts and visual prompts, significantly outperforms VA-Former by 1.54% AUC, indicating that introducing both types of prompts into the model simultaneously results in a more robust pMCI classification model. We further validate the importance of the global prompt module in the VAP-Former by

Table 2. Ablation study for prompt fine-tuning. 'Vis-GP' denotes the visual global prompt. The best results are in **bold**.

Method	Prompt		Vis-GP	BACC	F1	AUC
	Visual	Tabular				
VA-Former	✗	✗	✗	$78.29_{\pm0.52}$	$62.93_{\pm0.29}$	$84.77_{\pm0.35}$
VisPrompt	✓	✗	✗	$77.95_{\pm0.66}$	$60.15_{\pm2.51}$	$84.84_{\pm1.28}$
TabPrompt	✗	✓	✗	$77.39_{\pm0.80}$	$60.49_{\pm0.60}$	$84.70_{\pm0.81}$
Vis-TabPrompt	✓	✓	✗	$78.19_{\pm0.77}$	$63.11_{\pm1.45}$	$85.23_{\pm0.53}$
VAP-Former	✓	✓	✓	$\mathbf{79.22_{\pm0.58}}$	$\mathbf{63.13_{\pm0.11}}$	$\mathbf{86.31_{\pm0.25}}$
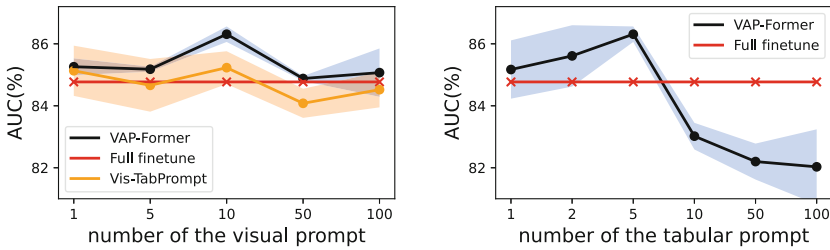
Fig. 3. Investigation of the number of prompt tokens. The line in red, black, and orange denote the performance of fully-finetuned VA-Former, VAP-Former, and Vis-TabPrompt, respectively. The light blue and orange area represents the error interval obtained by 3 different running seeds. The left figure shows that our global prompt improves the AUC and alleviates the performance fluctuations [15] in prompt learning. (Color figure online)

removing it from the model (denoted as Vis-TabPrompt). As shown in Table 2, removing the global prompt results in degraded performance by 1.08% AUC.

To investigate the impact of the number of prompts on performance, we evaluate VAP-Former with varying numbers of prompts. Given that the number of visual tokens exceeds that of tabular tokens. We fix the number of tabular prompts at 5 (left plot in Fig. 3) and fix the number of visual prompts at 10 (right plot), respectively while changing the other one. We hypothesize that the interaction between prompts and feature tokens is insufficient, so we gradually increase the number of tokens. As shown in Fig. 3, we observe that the model's performance ceases to increase after a certain number of prompts, confirming our assumption in Sect. 3.2 that too many randomly initialized prompt tokens make the model difficult to train. Conversely, when the number of prompts is small, the prompts can not learn enough information for the task and the interaction between prompt and feature tokens is not sufficient. Furthermore, as depicted in Fig. 3, the light blue area is smaller than the orange area, suggesting that the global prompt module makes VAP-Former more robust by helping the model to focus on important global features for pMCI detection.

4   Conclusion

To detect pMCI with visual and attribute data, we propose a simple but effective transformer-based model, VA-Former, to learn multi-modal representations. Besides, we propose a global prompt-based tuning strategy, which is integrated with the VA-Former to obtain our overall framework VAP-Former. The proposed framework can efficiently transfer the learned knowledge from AD classification to the pMCI prediction task. The experimental results not only show that the VAP-Former performs better than uni-modal models, but also suggest that the VAP-Former with the proposed prompt tuning strategy even surpasses the full fine-tuning while dramatically reducing the tuned parameters.

References

1. Al-Kababji, A., Bensaali, F., Dakua, S.P.: Scheduling techniques for liver segmentation: Reducelronplateau vs OneCycleLR. In: Bennour, A., Ensari, T., Kessentini, Y., Eom, S. (eds.) ISPR 2022, vol. 1589, pp. 204–212. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-08277-1_17
2. Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D.: Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage **145**, 137–165 (2017)
3. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: 20th International Conference on Pattern Recognition, pp. 3121–3124. IEEE (2010)
4. Brown, T., et al.: Language models are few-shot learners. Adv. Neural Inf. Process. Syst. **33**, 1877–1901 (2020)
5. Chen, S., Guhur, P.L., Schmid, C., Laptev, I.: History aware multimodal transformer for vision-and-language navigation. Adv. Neural Inf. Process. Syst. **34**, 5834–5847 (2021)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
7. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=YicbFdNTTy
8. Ebrahimighahnavieh, M.A., Luo, S., Chiong, R.: Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. Comput. Methods Prog. Biomed. **187**, 105242 (2020)
9. El-Sappagh, S., Abuhmed, T., Islam, S.R., Kwak, K.S.: Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. Neurocomputing **412**, 197–215 (2020)
10. Gao, P., et al.: Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6639–6648 (2019)
11. Gong, H., Chen, G., Mao, M., Li, Z., Li, G.: Vqamix: conditional triplet mixup for medical visual question answering. IEEE Trans. Med. Imaging **41**(11), 3332–3343 (2022)

12. He, X., Yang, S., Li, G., Li, H., Chang, H., Yu, Y.: Non-local context encoder: robust biomedical image segmentation against adversarial attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8417–8424 (2019)

13. Huang, J., Li, H., Li, G., Wan, X.: Attentive symmetric autoencoder for brain MRI segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, pp. 203–213. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-16443-9_20

14. Jack, C.R., Jr., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging **27**(4), 685–691 (2008)

15. Jia, M., et al.: Visual prompt tuning. In: Avidan, S., Brostow, G., Cisse, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, pp. 709–727. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-19827-4_41

16. Li, H., Chen, G., Li, G., Yu, Y.: Motion guided attention for video salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7274–7283 (2019)

17. Li, H., et al.: View-disentangled transformer for brain lesion detection. In: IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2022)

18. Li, H., Li, G., Yang, B., Chen, G., Lin, L., Yu, Y.: Depthwise nonlocal module for fast salient object detection using a single thread. IEEE Trans. Cybern. **51**(12), 6188–6199 (2020)

19. Lian, C., Liu, M., Zhang, J., Shen, D.: Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural MRI. IEEE Trans. Pattern Anal. Mach. Intell. **42**(4), 880–893 (2020)

20. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018). https://openreview.net/forum?id=rk6qdGgCZ

21. Padhi, I., et al.: Tabular transformers for modeling multivariate time series. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3565–3569 (2021)

22. Pan, Y., Chen, Y., Shen, D., Xia, Y.: Collaborative image synthesis and disease diagnosis for classification of neurodegenerative disorders with incomplete multimodal neuroimages. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 480–489. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_46

23. Pan, Y., Liu, M., Lian, C., Xia, Y., Shen, D.: Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages. IEEE Trans. Med. Imaging **39**(9), 2965–2975 (2020)

24. Pölsterl, S., Wolf, T.N., Wachinger, C.: Combining 3D image and tabular data via the dynamic affine feature map transform. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 688–698. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_66

25. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)

26. Risacher, S.L., Saykin, A.J., Wes, J.D., Shen, L., Firpi, H.A., McDonald, B.C.: Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Curr. Alzheimer Res. **6**(4), 347–361 (2009)

27. Ruiz, J., Mahmud, M., Modasshir, Md., Shamim Kaiser, M.: 3D DenseNet ensemble in 4-way classification of alzheimer's disease. In: Mahmud, M., Vassanelli, S., Kaiser, M.S., Zhong, N. (eds.) BI 2020. LNCS (LNAI), vol. 12241, pp. 85–96. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59277-6_8

28. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: Unetr++: delving into efficient and accurate 3d medical image segmentation. arXiv preprint arXiv:2212.04497 (2022)
29. Spasov, S., Passamonti, L., Duggento, A., Lio, P., Toschi, N., Initiative, A.D.N., et al.: A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. Neuroimage **189**, 276–287 (2019)
30. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
31. Winblad, B., et al.: Defeating Alzheimer's disease and other dementias: a priority for European science and society. Lancet Neurol. **15**(5), 455–532 (2016)
32. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vision **130**(9), 2337–2348 (2022)