PROMPT-BASED PERSONALIZED FEDERATED LEARNING FOR MEDICAL VISUAL QUESTION ANSWERING

He Zhu[†], Ren Togo[††], Takahiro Ogawa[††] and Miki Haseyama[††]

[†]Graduate School of Information Science and Technology, Hokkaido University, Japan
[††]Faculty of Information Science and Technology, Hokkaido University, Japan
E-mail: {zhu, togo, ogawa, mhaseyama}@lmd.ist.hokudai.ac.jp

ABSTRACT

We present a novel prompt-based personalized federated learning (pFL) method to address data heterogeneity and privacy concerns in traditional medical visual question answering (VQA) methods. Specifically, we regard medical datasets from different organs as clients and use pFL to train personalized transformer-based VQA models for each client. To address the high computational complexity of client-to-client communication in previous pFL methods, we propose a succinct information sharing system by introducing prompts that are small learnable parameters. In addition, the proposed method introduces a reliability parameter to prevent the negative effects of low performance and irrelevant clients. Finally, extensive evaluations on various heterogeneous medical datasets attest to the effectiveness of our proposed method.

Index Terms— visual question answering, prompt learning, personalized federated learning, medical image.

1. INTRODUCTION

Rising poverty levels and an increasing aging population are exacerbating the global shortage of healthcare resources. Medical AI is now seen as a pivotal solution to address these problems in the foreseeable future [1]. The general tasks in medical AI, such as classification [2] and segmentation [3], lack adaptability to rapidly changing clinical situations and often neglect patient involvement in medical diagnoses. The recent rise of chatGPT [4] demonstrates capabilities that closely resemble human conversation and analytical skills, showing the potential of the language for medical applications. As a vision-language task, medical visual question answering (VQA) [5] can interpret questions framed in natural language and provide accurate, user-friendly responses. This is expected to alleviate the workload on physicians by facilitating communication with patients.

While previous research has primarily focused on improving model accuracy [6], two significant challenges remain in the medical domain. First, data distributions of the different patients are difficult to learn by a single model due to the heterogeneity such as feature distribution drift and label distribution skew [7]. The other is that due to patient privacy concerns, the volume of data available for training remains constrained [8]. In the general domain, the federated learning (FL) [9] and the personalized federated learning (pFL) [10] are proposed for training personalized models separately using private data. FL utilizes a central server to manage the global model, enabling certain clients to participate. On the contrary, pFL

has been developed to learn a customized model for each client that performs better on private data while benefiting from collaborative training. Essentially, each pFL client model is uniquely calibrated to its respective user, and the individual clients do not communicate any data during the training process. However, due to the limitations of the previous pFL framework, it is challenging to introduce it directly into medical VQA.

The traditional pFL framework [11] allows for information sharing by communicating the model weights of each client [12], which leads to many parameters requiring updates. As a result of computational limitations, pFL applications have been mostly confined to Deep Neural Networks (DNNs) that handle simpler datasets such as MNIST [13] and CIFAR10/100 [14]. However, the medical VQA task has textual information as an additional modality and typically deals with informative images. Furthermore, the transformer model [15], known for its state-of-the-art performance in the VQA task, inherently possesses more parameters than conventional DNNs. Although Li et al. [16] proposed the personalized simple transformer with few blocks [15], it follows the previous consumptive parameter communication, which brings a massive consumption of computing resources. Inspired by Wang et al. [17, 18], we introduce prompt learning [19] into the pFL method to reduce the number of parameters involved in client communication, which can reduce the computational complexity by replacing model weights with prompts.

In this paper, we propose prompt-based personalized federated learning for medical VQA to address the heterogeneity and privacy problem. We establish clients for heterogeneous medical data from different organs and simultaneously train personalized VQA models for each client locally, where each model can be regarded as learning a different data distribution. To achieve privacy protection, each client can only access its own private data during training, and information is shared among clients through model parameters. Specifically, we introduce the prompt, a small learnable parameter with high dimensionality, as an additional input to the residual attention blocks in the transformer encoder. We only update the prompt during training so that it obtains the information in the training data. Since a client's prompts are accessible to other clients, we facilitate client-to-client communication by integrating the prompts to generate shared information. Moreover, we propose a reliability parameter that determines the weight of each client in the communication process. Our experimental findings indicate that parameter communication between clients outperforms individualized training. Finally, we summarize our contributions of this paper as follows.

- We propose a novel prompt-based pFL method for personalized medical VQA addressing the problems of the heterogeneity and privacy preservation in medical domain.
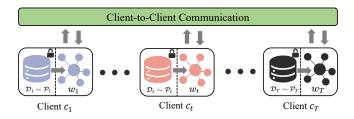
Fig. 1. An overview of the proposed prompt-based pFL method. A client c_t has private data $\mathcal{D}_t$ following the distribution $\mathcal{P}_t$, and a personalized client VQA model with the weight $\boldsymbol{w}_t$. The client uploads the local information and obtains shared information through client-to-client communication.

- We provide a simple method to represent model information using prompts, which also reduces the computational complexity of the client-to-client communication process.
- We propose a reliability parameter for controlling the weight of each client in the communication process to avoid negative effects.

2. PROMPT-BASED PERSONALIZED FEDERATED LEARNING FOR MEDICAL VQA

We provide an overview of our method in Fig. 1. We first formalize the proposed prompt-based pFL problem and then present the pFL medical VQA method.

2.1. Problem Formulation

The proposed method aims to train personalized medical VQA models for T clients collaboratively. Each client has its own private data and can communicate with others without sharing these data. The t-th client denoted as c_t is characterized by its data distribution $\mathcal{P}_t : \mathcal{X} \times \mathcal{A}$, where $\mathcal{X}$ represents the input data space and $\mathcal{A}$ denotes the answer space. Assume that client c_t has access to N_t independent and identically distributed (IID) samples from $\mathcal{P}_t$, denoted by $\mathcal{D}_t = \{(I_t^j, Q_t^j, A_t^j)\}_{j=1}^{N_t}$, where I_t^j, Q_t^j and A_t^j respectively represent the input images, questions and the ground truth answers.
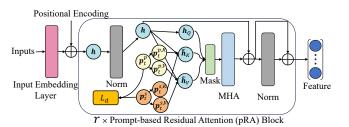
Let $\mathcal{L}_t : \mathcal{X} \times \mathcal{A} \to \mathbb{R}+$ represent the loss function associated with client c_t and the data sampled from the distribution $\mathcal{P}_t$. The goal of the proposed method is to optimize

$$\mathcal{L}(\boldsymbol{W}) = \arg\min_{\boldsymbol{w}_t} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{(\mathcal{X}, \mathcal{A}) \sim \mathcal{P}_t} [\mathcal{L}_t(\boldsymbol{w}_t; \mathcal{S}_t, \mathcal{D}_t)], \quad (1)$$

where $\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_T\}$ is the parameter set of the client models, and $\mathcal{S}_t$ represents the shared information from all clients except c_t. However, as the true underlying distribution is often inaccessible, the goal is generally achieved through empirical risk minimization. Consequently, the training objective can be computed as follows:

$$\arg\min_{\boldsymbol{w}_t} \sum_{t=1}^{T} [\mathcal{L}_{\text{CE}} + \mathcal{R} + \mathcal{L}_{\text{d}}], \quad (2)$$

where $\mathcal{L}_{\text{CE}}$ is a local average loss over personal training data, specifically, cross-entropy loss, $\mathcal{R}$ serves as a regularization term applied during the training process to mitigate overfitting among clients, and $\mathcal{L}_d$ represents a distance loss to ensure the shared information can be utilized efficiently. The proposed method enables the simultaneous training of personalized VQA models tailored for diverse data dis-

Fig. 2. The proposed transformer encoder that introduces the prompt-based self-attention block. Local prompt $\boldsymbol{p}_t^p$ and shared prompt $\boldsymbol{p}_t^s$ are integrated into the process of self-attention, and the distance between them is controlled by the distance loss $\mathcal{L}_d$.

tributions while sharing information instead of private data, thereby ensuring privacy and facilitating mutual benefits among clients.

2.2. Prompt-based Visual Question Answering Model

We describe the proposed prompt-based VQA client model with a general client c_t as an example. We employ a specific variational auto-encoder structure [20] that integrates an image encoder, a text encoder, and an answering decoder, where the two encoders have the same transformer architecture and different parameters. The original transformer encoder consists of an input embedding layer and a set of multi-head self-attention (MHA) based residual attention (RA) blocks. The RA block uses self-attention mechanisms and feedforward neural networks, combined with residual connections and normalization, to effectively process data sequences.

Directly using the original architecture can lead to excessive load in the communication process, significantly reducing the efficiency of federated learning. Therefore, we propose a novel transformer encoder that introduces the prompt-based residual attention (pRA) block as shown in Fig. 2. The pRA block uses a set of learnable prompts as an additional input of the MHA layer. Specifically, for client c_t, we introduce local prompt $\boldsymbol{p}_t^p \in \mathbb{R}^{r \times L \times D}$ trained locally and shared prompts $\boldsymbol{p}_t^s \in \mathbb{R}^{r \times L \times D}$ obtained from other clients, where r is the number of the pRA blocks, and L is the length of the prompts. In contrast to the original RA block, the proposed pRA block introduces the prompts into the MHA layer using Prefix-Tuning (Pre-T) [21]. Pre-T splits $\boldsymbol{p}_t^p$ and $\boldsymbol{p}_t^s$ into $\{\boldsymbol{p}_t^{p,K}, \boldsymbol{p}_t^{p,V}\} \in \mathbb{R}^{1 \times L/2 \times D}$ and $\{\boldsymbol{p}_t^{s,K}, \boldsymbol{p}_t^{s,V}\} \in \mathbb{R}^{1 \times L/2 \times D}$ for each of the r blocks, respectively.

Each input image or question is embedded by the input embedding layer to $\boldsymbol{h} = \{\boldsymbol{h}_Q, \boldsymbol{h}_K, \boldsymbol{h}_V\} \in \mathbb{R}^D$, where the positional encoding has been added. Next, h is fed into the MHA layer, which is calculated as follows:

$$\text{MHA}(\boldsymbol{h}_Q, \boldsymbol{h}_K, \boldsymbol{h}_V, \boldsymbol{p}_t^p, \boldsymbol{p}_t^s) = \text{Concat}(\hat{h}_1, ..., \hat{h}_i, ..., \hat{h}_m)\boldsymbol{W}^O, \quad (3)$$

$$\hat{h}_i = \text{Attention}(\boldsymbol{h}_Q \boldsymbol{W}_i^Q, \hat{\boldsymbol{h}}_K \boldsymbol{W}_i^K, \hat{\boldsymbol{h}}_V \boldsymbol{W}_i^V), \quad (4)$$

$$\hat{\boldsymbol{h}}_K = \left[\boldsymbol{p}_t^{p,K} \oplus \boldsymbol{p}_t^{s,K}; \boldsymbol{h}_K\right], \hat{\boldsymbol{h}}_V = \left[\boldsymbol{p}_t^{p,V} \oplus \boldsymbol{p}_t^{s,V}; \boldsymbol{h}_V\right]. \quad (5)$$

where $\boldsymbol{W}^O, \boldsymbol{W}_i^Q, \boldsymbol{W}_i^K$ and $\boldsymbol{W}_i^K$ are the parameter matrices of the transformer, and $\text{Attention}(\cdot)$ is the self-attention. Moreover, we merge the original mask of the MHA layer with a all-ones matrix as the new mask to ensure that each block can refer to all the prompts during the self-attention process. In particular, we do not add the location information for the prompts because they are introduced directly as parameters.

The outputs of the last pRA block of the image encoder and the text encoder are the image and language features of the input
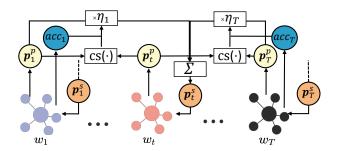
Fig. 3. The proposed client-to-client communication process. Each client communicates to update $\boldsymbol{p}_t^p$ and get $\boldsymbol{p}_t^s$ from the other clients. The reliability parameter η controls the weights of different clients in the communication process, which is computed from the accuracy acc and the similarity function $\mathrm{cs}(\cdot)$.

image I_t^j and question Q_t^j. Next, the extracted features are joined and input into the personalized answer layer consisting of an MLP to output the probabilities of the predicted answer $\hat{A}_t^j$. We employ the cross-entropy loss to train the model, which can be calculated as

$$\mathcal{L}_{CE} = -\sum_{j=1}^{N_t} A_t^j \log(\hat{A}_t^j). \tag{6}$$

Moreover, to prevent heterogeneity between local and shared prompts leading to degradation of the results, we also introduce the following loss to control the distance between the two prompts:

$$\mathcal{L}_d = 1 - \frac{\boldsymbol{p}_t^p \cdot \boldsymbol{p}_t^s}{\|\boldsymbol{p}_t^p\|_2 \times \|\boldsymbol{p}_t^s\|_2}, \tag{7}$$

where $\|\boldsymbol{p}_t^p\|_2$ and $\|\boldsymbol{p}_t^s\|_2$ are the L2 norms of the $\boldsymbol{p}_t^p$ and $\boldsymbol{p}_t^s$, respectively. The final loss can be calculated as follows:

$$\mathcal{L}_{\text{client}} = \mathcal{L}_{CE} + \alpha \mathcal{L}_d + \beta \mathcal{R}, \tag{8}$$

where $\mathcal{R} = \|\boldsymbol{w}_t^{\text{mlp}}\|_2^2$ is the L2 regularization constraint on the weights of the MLP layer to prevent overfitting.

All clients use parameter matrices with the same weights while training their respective prompts during the training process. By introducing the pRA block, the prompts can be used for client-to-client communication, which reduces the computational complexity by several orders of magnitude.

2.3. Client-to-Client Communication

We decouple the VQA model of client c_t into the shared feature extractors and the personalized final answer layer, which are parameterized by $\boldsymbol{p}_t^p$ and $\boldsymbol{w}_t^{\text{mlp}}$, respectively. Each client has a task similar to multitask learning, where the number of the answer list is completely different between clients. The clients share the information through the shared prompt $\boldsymbol{p}_t^s$, and Eq. (1) can be further presented as follows:

$$\mathcal{L}(W) = \arg \min_{\boldsymbol{p}_t^p, \boldsymbol{w}_t^{\text{mlp}}} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{(\mathcal{X},\mathcal{A}) \sim \mathcal{P}_t}[\mathcal{L}_t(\boldsymbol{p}_t^p, \boldsymbol{w}_t^{\text{mlp}}; \boldsymbol{p}_t^s, \mathcal{D}_t)]. \tag{9}$$

Since there are differences in the performance and tasks of clients, there should be a tendency to access shared information on the client. Therefore, we introduce a set of client's reliability parameter $\{\eta_t\}_{t=1}^T$ obtained from the performance and task similarity. For

---

Algorithm 1 Prompt-based pFL method for medical VQA

Require: I-Number of iterations, I_L-Number of local iterations, α — learning rate, T-Number of clients
1: Initialization: client set $C = \{\boldsymbol{p}_t^p, \boldsymbol{p}_t^s, \mathbf{w}_t^{\text{mlp}}\}_{t=1}^T$
2: **for** $i = 1$ to I **do**
3:     **for** $t = 1$ to T **do**
4:         **for** $j = 1$ to I_L **do**
5:             $\boldsymbol{p}_t^p = \boldsymbol{p}_t^p - \alpha \nabla_{\boldsymbol{p}_t^p}(\mathcal{L}_{\text{client}} + \mathcal{L}_d)$
6:             $\boldsymbol{w}_t^{\text{mlp}} = \boldsymbol{w}_t^{\text{mlp}} - \alpha \nabla_{\boldsymbol{w}_t^{\text{mlp}}} \mathcal{L}_{\text{client}} - \beta \nabla_{\boldsymbol{w}_t^{\text{mlp}}} \|\boldsymbol{w}_t^{\text{mlp}}\|_2^2$
7:         **end for**
8:         $\boldsymbol{p}_t^s = \sum_{i=1,\, i \neq t}^T \eta_i \boldsymbol{p}_i^p$
9:     **end for**
10: **end for**
Ensure: $C = \{\boldsymbol{p}_t^p, \boldsymbol{p}_t^s, \boldsymbol{w}_t^{\text{mlp}}\}_{t=1}^T$

client c_t, the shared prompt $\boldsymbol{p}_t^s$ can be calculated as follows:

$$\boldsymbol{p}_t^s = \sum_{i=1, i \neq t}^{T} \eta_i \boldsymbol{p}_i^p, \tag{10}$$

$$\eta_i = \frac{acc_i \times \mathrm{cs}(\boldsymbol{p}_i^p, \boldsymbol{p}_t^p)}{\sum_{j=1, j \neq t}^{T} acc_j \times \mathrm{cs}(\boldsymbol{p}_j^p, \boldsymbol{p}_t^p)}, \tag{11}$$

where acc_i denotes the current accuracy of the client c_i in answering the question, and $\mathrm{cs}(\cdot)$ denotes cosine similarity. It can be seen from Eq. (11) that clients with high accuracy and higher similarity to the target client c_t have higher reliability in the communication process. Therefore, introducing the reliability parameter makes clients refer more to the information from clients with high accuracy, similar to their tasks, preventing the interference of invalid information. We provide an overview of the client-to-client communication in Fig. 3.

After several local iterations, the client c_t interacts with other clients to update the shared information before making local updates. By sharing parameters between clients, the client c_t can access additional information without changing the end-to-end training process that only accesses the private data. The main training procedure is summarized in Algorithm 1.
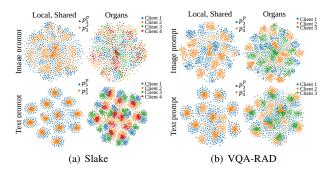
3. EXPERIMENTS

3.1. Settings

Dataset and Client. In our experiments, we leveraged medical images and closed-type QA pairs from both VQA-RAD [22] and Slake [23] datasets. These datasets were further partitioned into sub-datasets based on the anatomical focus of the images. We split the datasets into sub-datasets depending on the organ of the image taken and set up client models for each sub-dataset. Finally, we set four clients for Slake including "Lung", "Abdomen", "Brain" and "Others", which contain 734, 829, 461, and 335 question pairs, respectively. We set three clients for VQA-RAD including "Chest", "Abdomen" and "Brain", which contain 511, 439, and 269 question pairs, respectively. We adhered to the original dataset division ratios as 70% for training, 15% for validation and 15% for testing in Slake, and 85% for training and 15% for testing in VQA-RAD.

Our architecture employs the Contrastive Language–Image Pre-training (CLIP) [24] model as the backbone for both the image and text encoders. The answer layer is implemented as a 2-layer MLP with a hidden size of 512 and a dropout rate of 0.2. We standardized the length of all prompts to four. The weight α of the loss was set to 0.5, and β was set to 0.001. During the training, we set the global

Table 1. The quantitative results of each client of Slake.

	Brain	Abdomen	Others	Lung
VGG+SAN [23]	76.92	72.34	74.41	75.89
PMCLIP [26]	79.12	**80.85**	76.74	82.98
M2I2 [27]	79.12	68.09	74.42	81.56
AS1	79.12	70.92	<u>81.40</u>	<u>83.69</u>
AS2	**81.32**	72.34	79.07	82.98
AS3	<u>80.22</u>	<u>73.76</u>	**81.84**	82.27
AS4	<u>80.22</u>	70.92	76.74	**84.40**
PM	76.92	72.34	<u>81.40</u>	**84.40**

Table 2. The quantitative results of each client of VQA-RAD.

	Brain	Abdomen	Chest
MEVF [25]	68.63	66.67	60.61
PMCLIP [26]	<u>73.53</u>	<u>70.00</u>	62.12
M2I2 [27]	64.38	**71.11**	<u>63.64</u>
AS1	**74.24**	67.78	61.76
AS2	68.18	63.33	61.76
AS3	**74.24**	67.78	61.76
AS4	69.70	67.78	60.78
PM	**74.24**	68.89	**64.71**

(a) Slake (b) VQA-RAD

Fig. 4. The tSNE visualization of the prompt embedding results on the Slake and VQA-RAD datasets. The "Local, Shared" represents the local and shared prompts in client 3 that is trained in the sub-dataset of the Abdomen. The "Ograns" represents the local prompt of each client.

epoch to 30 and each local iteration to 5. Based on the experimental results, we used a uniform learning rate of 0.01 for all client models, which is multiplied by 0.1 every ten epochs. After adjusting for the block number r, we found that the model had the best performance when r was set to 12.

Comparison Methods. We compared the proposed method with several state-of-the-art medical VQA models. We used the DNNs-based methods, VGG+SAN [23] and MEVF [25] for Slake and VQA-RAD, respectively. And we used the transformer-based methods, PMCLIP [26] and M2I2 [27] for both of them. To ensure consistent experimental conditions, we trained the comparison separately using split sub-datasets and did not use pre-training.

Ablation Studies. We used ablation experiments to demonstrate the effectiveness of the proposed method. We separately experimented with the proposed method with the following settings. AS1: w/o prompt and pFL, AS2: w/o pFL, AS3: w/o image prompt, AS4: w/o text prompt.
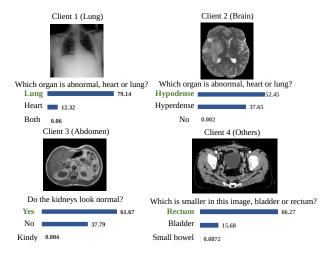
3.2. Results

We conducted experiments on the accuracy of different client models in answering questions, which are shown in Tables 1 and 2. The high accuracy of PM, as well as AS3 and AS4, which incorporated pFL, attest to the efficacy of the pFL method. Compared to AS1, AS2 does not exhibit significant improvements, demonstrating that incorporating prompts offers limited enhancements to the model's performance. Moreover, for Slake, the text prompt had a more positive impact than the image prompt, and for the VQA-RAD, both the image and text prompts improved the accuracy of the model.

We also visualized the prompts with tSNE [28]. From the visualization results in Fig. 4(a), consistent with the quantitative results, we see that the text prompt of the Slake is more accurately mapped

Fig. 5. Samples of the generated answers on Slake. The results show the four clients we set up, with the corresponding sub-datasets in parentheses.

to different regions, while the image prompt is relatively cluttered. Furthermore, as shown in both Figs. 4(a) and 4(b), the 12 regions in the figures correspond to the 12 pRA blocks in the proposed method, which can prove that each block contains different potential information. The clear separation into several parts within each region in the "Organ" can prove that the prompts trained under different data distributions contain entirely different information. The "Local, Shared" results demonstrate the effectiveness of the proposed distance loss, where the corresponding local and shared prompts in each block are mapped to a similar space. We also qualitatively evaluated the model answering the questions as shown in Fig. 5. It can be seen that our model can correctly answer the questions corresponding to the medical images. In the experiments, the parameters to be updated of the proposed method were only 0.05% of that of previous pFL methods.

4. CONCLUSION

The proposed prompt-based personalized federated learning method for medical VQA can train personalized models for sub-datasets of medical data with different distributions, which overlooked problems of medical data heterogeneity and privacy protection. By using the prompt-based self-attention block, the client model circumvented the computational intricacies of traditional model weight sharing. Both qualitative and quantitative results show that our method allows the model to gain in communication. Moreover, the proposed method is expected to be used in any transformer-based VQA baseline models.

5. REFERENCES

[1] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[2] Eka Miranda, Mediana Aryuni, and E. Irwansyah, "A survey of medical image classification techniques," in *Proceedings of International Conference on Information Management and Technology*, 2016, pp. 56–61.

[3] Saleha Masood, Muhammad Sharif, Afifa Masood, Mussarat Yasmin, and Mudassar Raza, "A survey on medical image segmentation," *Current Medical Imaging*, pp. 3–14, 2015.

[4] OpenAI, "GPT-3.5 (ChatGPT) – A large language model by OpenAI," https://openai.com, 2023, Accessed: Sep. 6, 2023.

[5] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge, "Medical visual question answering: A survey," *Artificial Intelligence in Medicine*, vol. 143, pp. 102611, 2023.

[6] Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li, "Cross-modal self-attention with multi-task pre-training for medical visual question answering," in *Proceedings of the International Conference on Multimedia Retrieval*, 2021, pp. 456–460.

[7] Nicholas J Schork, "Artificial intelligence and personalized medicine," *Precision medicine in Cancer therapy*, pp. 265–283, 2019.

[8] Qiong Wu, Kaiwen He, and Xu Chen, "Personalized federated learning for intelligent iot applications: A cloud-edge based framework," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 35–44, 2020.

[9] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, pp. 1–19, 2021.

[10] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–17, 2022.

[11] Junyi Zhu, Xingchen Ma, and Matthew B Blaschko, "Confidence-aware personalized federated learning via variational expectation maximization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24542–24551.

[12] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant, "Survey of personalization techniques for federated learning," in *Proceedings of World Conference on Smart Trends in Systems, Security and Sustainability*, 2020, pp. 794–797.

[13] Li Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[14] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," *University of Toronto*, 2009.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[16] Hongxia Li, Zhongyi Cai, Jingya Wang, Jiangnan Tang, Weiping Ding, Chin-Teng Lin, and Ye Shi, "Fedtp: Federated learning by transformer personalization," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.

[17] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister, "Learning to prompt for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 139–149.

[18] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al., "Dualprompt: Complementary prompting for rehearsal-free continual learning," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 631–648.

[19] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[20] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[21] Xiang Lisa Li and Percy Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[22] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific Data*, vol. 5, pp. 1–10, 2018.

[23] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu, "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *Proceedings of the IEEE International Symposium on Biomedical Imaging*, 2021, pp. 1650–1654.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 8748–8763.

[25] Binh D. Nguyen, Thanh-Toan Do, Binh X. Nguyen, Tuong Do, Erman Tjiputra, and Quang D. Tran, "Overcoming data limitation in medical visual question answering," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention*, 2019, pp. 522–530.

[26] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo, "Pubmedclip: How much does clip benefit visual question answering in the medical domain?," in *Findings of the Association for Computational Linguistics: EACL*, 2023, pp. 1151–1163.

[27] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong, "Self-supervised vision-language pretraining for medial visual question answering," in *Proceedings of the IEEE 20th International Symposium on Biomedical Imaging*, 2023, pp. 1–5.

[28] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, 2008.