Stealing Knowledge from Pre-trained Language Models for Federated Classifier Debiasing

Meilu Zhu[1], Qiushi Yang[2], Zhifan Gao[3], Jun Liu[1(✉)], and Yixuan Yuan[4(✉)]

[1] Department of Mechanical Engineering, City University of Hong Kong
meiluzhu2-c@my.cityu.edu.hk, Jun.Liu@cityu.edu.hk
[2] Department of Electrical Engineering, City University of Hong Kong
[3] School of Biomedical Engineering, Sun Yat-sen University
[4] Department of Electronic Engineering, Chinese University of Hong Kong
yxyuan@ee.cuhk.edu.hk

Abstract. Federated learning (FL) has shown great potential in medical image computing since it provides a decentralized learning paradigm that allows multiple clients to train a model collaboratively without privacy leakage. However, current studies have shown that heterogeneous data of clients causes biased classifiers of local models during training, leading to the performance degradation of a federation system. In experiments, we surprisingly found that continuously freezing local classifiers can significantly improve the performance of the baseline FL method (FedAvg) for heterogeneous data. This observation motivates us to pre-construct a high-quality initial classifier for local models and freeze it during local training to avoid classifier biases. With this insight, we propose a novel approach named Federated Classifier deBiasing (FedCB) to solve the classifier biases problem in heterogeneous federated learning. The core idea behind FedCB is to exploit linguistic knowledge from pre-trained language models (PLMs) to construct high-quality local classifiers. Specifically, FedCB first collects the class concepts from clients and then uses a set of prompts to contextualize them, yielding language descriptions of these concepts. These descriptions are fed into a pre-trained language model to obtain their text embeddings. The generated embeddings are sent to clients to estimate the distribution of each category in the semantic space. Regarding these distributions as the local classifiers, we perform the alignment between the image representations and the corresponding semantic distribution by minimizing an upper bound of the expected cross-entropy loss. Extensive experiments on public datasets demonstrate the superior performance of FedCB compared to state-of-the-art methods. The source code is available at https://github.com/CUHK-AIM-Group/FedCB.

Keywords: Federated learning · Medical Image Classification · Pre-trained Language Model.

1 Introduction

The excellent performance of deep learning in medical image analysis [17,21,24] depends to the availability of large-scale medical image datasets. In real-world

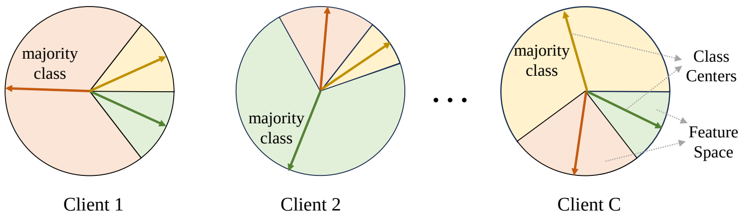

Fig. 1. Data heterogeneity causes biased classifiers of local clients (Best viewed in color).

Table 1. The performance of different methods on OCT-C8 dataset under different heterogeneity. β represents the Dirichlet coefficient which is selected from $[0.05, 0.1]$.Client number is 12 and the backbone network is ResNet-18 [5].

Methods	$\beta = 0.05$		$\beta = 0.1$	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
FedAvg [14]	74.82±5.98	72.34±6.87	78.64±5.44	76.25±7.23
Random	76.11±6.31	73.07±7.57	82.67±3.40	82.03±3.80

medical scenarios, data may be distributed in different hospitals or institutions, leading to the infeasibility of constructing large datasets due to growing privacy concerns or legal restrictions [25]. Driven by such realistic needs, federated learning (FL) [14,10] has become an emerging research topic, allowing to learn a global model across different clients (hospitals) without exchanging their private data under the orchestration of a cloud server.

Unfortunately, data from clients may be heterogeneous and seriously result in the performance degradation of a federation system [11,10,6]. One primary reason behind this dilemma is that data heterogeneity causes biased classifiers of local models during training [13,20], which are highly divergent and tend to focus on majority classes of local clients, as shown in Figure 1. Some recent studies have attempted to mitigate classifier biases, which can be roughly divided into two categories. The first branch [10,7,9,1] is to modify the local objectives of the clients, so that the local classifiers are consistent with the global classifier to a certain degree. Another route is to share private data information (including feature representation [20,6], feature distribution [13], and data subset [23,4]) to help construct a more balanced data distribution on the client or on the server. However, these approaches still fail to solve the classifier biases problem, since the global classifier is usually obtained by aggregating local biased classifiers and thus is also biased, while sharing private data information easily incurs data privacy leakage and high bandwidth cost.

Differing from these methods [10,7,9,1,20,6,13,23,4], we consider a straightforward idea to avoid the classifier biases — **Pre-constructing a high-quality classifier for clients and freezing it during training**. To preliminarily verify the feasibility of this idea, we conduct a pilot experiment on FedAvg [14]. Specifically, the server **randomly** initializes the global classifier and broadcasts it to all clients. During training, we **freeze** local classifiers, and only train and upload feature extractors, investigating the final performance of the global model. As shown in Table 1, this simple strategy surprisingly outperforms the vanilla FedAvg when the data of clients are heterogeneous. The results indicate that sharing a fixed classifier across clients is a feasible path to alleviate the classifier biases problem. Intuitively, random initialization is not the optimal strategy to build the fixed classifier, since it does not consider intra-class semantic information and inter-class distance relation. Therefore, how to construct a high-quality classifier for clients becomes a vital step to mitigate the classifier biases.

Motivated by recent language-to-vision models [16,22], natural language descriptions (such as diagnosis reports) carry rich semantic information and can represent images or clinical scans of different categories. This provides us with a new perspective to construct a good classifier by borrowing linguistic knowledge from pre-trained language models (PLMs). Based on this insight, we propose a new framework called Federated Classifier deBiasing (FedCB) to solve the classifier biases problem in heterogeneous federated learning. Specifically, to describe images of different categories more comprehensively, the server side first collects the class concepts from clients, and then uses a set of commonly-used prompts to contextualize them, yielding language descriptions of these concepts. These descriptions are fed into a pre-trained language model to obtain their text embeddings. We further exploit the generated embeddings to estimate the distribution of each category in the semantic space. Regarding these distributions as the local classifiers, we perform the alignment between the image representations of each class and the corresponding semantic distribution by minimizing an upper bound of the expected cross-entropy (CE) loss. We conduct extensive experiments on public datasets to evaluate the proposed framework. The results demonstrate the superior performance of FedCB against state-of-the-art methods.

2 Methodology

2.1 Motivation and Overview

Let's consider a typical federated learning scenario for medical image classification with C distributed clients and a server. Each client has a local cohort $\mathcal{D}^c = \{(\mathbf{x}_i^c, \mathbf{y}_i^c)\}_{i=1}^{N_c}$ with K classes, where N_c is the data amount of $\mathcal{D}^c$, and $\mathbf{x}_i^c$ is a training instance with the label $\mathbf{y}_i^c$. FL aims to unite these clients to train a global model f, where f contains a feature extractor and a classifier. The overall training process proceeds through communication between clients and the server for multiple rounds. In each round, the c-th client downloads the global model from the server to initialize the parameters of the local model f^c. After local training, local models are uploaded to the server to update the global model via model aggregation. However, the data of clients tend to be not independent and identically distributed (Non-IID) and lead to biased local classifiers [13,20]. Aggregating these biased classifiers will incur the performance degradation of the global model. To mitigate the classifier biases problem, the proposed FedCB attempts to steal knowledge from pre-trained language models to pre-construct

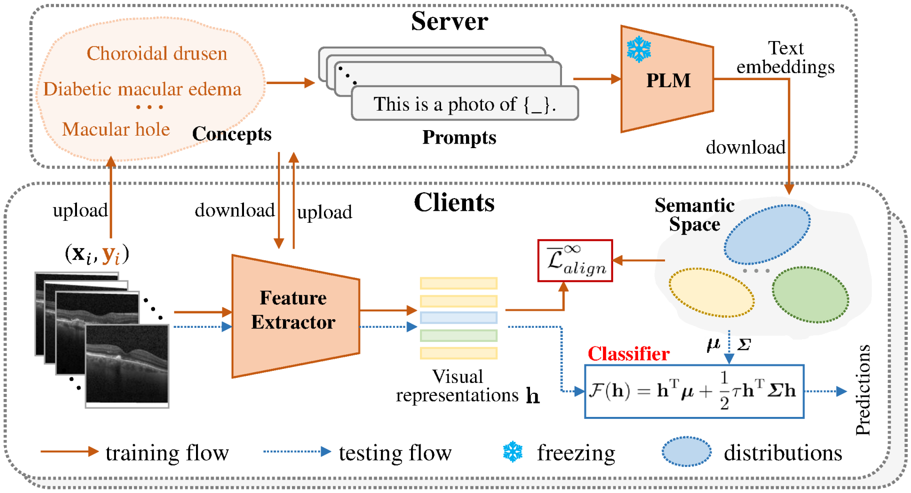
Server
Choroidal drusen
Diabetic macular edema
Macular hole
Concepts
This is a photo of {_}.
Prompts
PLM
Text embeddings
download
Clients
upload
download
upload
$(\mathbf{x}_i, \mathbf{y}_i)$
Feature Extractor
Visual representations $\mathbf{h}$
$\mathcal{L}^{\infty}_{align}$
Semantic Space
$\boldsymbol{\mu}$ $\boldsymbol{\Sigma}$
Classifier
$\mathcal{F}(\mathbf{h}) = \mathbf{h}^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\tau\mathbf{h}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{h}$
Predictions
training flow
testing flow
freezing
distributions

Fig. 2. The overview of the proposed FedCB framework. FedCB first collects a concept set from clients to obtain text embeddings via a pre-trained language model (PLM). These text embeddings are sent to clients for building local classifiers and training local feature extractors.

a high-quality classifier for all clients and then fix it at the client side during the overall training process. The overall framework is shown in Figure 2.

2.2 FedCB: Federated Classifier deBaising

The core challenge of pre-constructing a high-quality classifier is to guarantee intra-class semantic information and inter-class distance relation. Recently, FedETF [12] utilizes orthogonal initialization to construct the classifier. However, this method lacks the semantic interpretability. In addition, classifier vectors are not necessarily strictly orthogonal. To tackle this challenge, we propose to introduce pre-trained language models and natural language descriptions of category concepts to construct the classifier.

As shown in Figure 2, the server first collects a concept set $\{P_k\}_{k=1}^{K}$ from clients, where P_k is the category name of the k-th class and K is the total category number. A set of M predetermined prompts (such as "This is an image of {concept}" and "The image shows {concept}." and so on) is used to contextualize the concepts. We input the contextualized concepts into a pre-trained language model (PLM) (such as the text encoder of BiomedCLIP [22]) to obtain a set of text embeddings $\boldsymbol{E}$, where $\boldsymbol{E} = \cup_{k=1}^{K}\{\boldsymbol{e}_1^{(k)}, \boldsymbol{e}_2^{(k)}, ..., \boldsymbol{e}_M^{(k)}\}$. PLM is trained on large-scale datasets based on contrastive learning and demonstrates strong feature transferability. Therefore, the obtained text embeddings in $\boldsymbol{E}$ contain rich semantics, which have two favorable properties: (i) distance relationship between concepts can be reflected through their similarities, (ii) text embeddings in the semantic space are more domain-agnostic. Afterwards, the server sends the text embeddings to all clients to build local classifiers. Noticeably, this process is only conducted once at the 1-st round, and thus does not incur the high communication and computation overhead.

After clients receive the text embeddings from the server, a naive method to build the local classifier is to average the text embeddings $\{\boldsymbol{e}_1^{(k)}, \boldsymbol{e}_2^{(k)}, ..., \boldsymbol{e}_M^{(k)}\}$ of each category, and then conduct feature alignment between image representations and the averaged embeddings for training the feature extractor, which can be formulated as minimizing the following contrastive loss:

$$\mathcal{L}_{align} = \frac{1}{N_c}\sum_{i=1}^{N_c} \mathbb{E}_{\boldsymbol{e}^{(\mathbf{y}_i)} \in \boldsymbol{\Omega}^{(\mathbf{y}_i)}} \left(-\log \frac{e^{\tau \mathbf{h}_i^{\mathrm{T}} \boldsymbol{e}^{(\mathbf{y}_i)}}}{e^{\tau \mathbf{h}_i^{\mathrm{T}} \boldsymbol{e}^{(\mathbf{y}_i)}} + \sum_{k \neq \mathbf{y}_i}^{K} \mathbb{E}_{\boldsymbol{e}^{(k)} \in \boldsymbol{\Theta}^{(\mathbf{y}_i)}} e^{\tau \mathbf{h}_i^{\mathrm{T}} \boldsymbol{e}^{(k)}}} \right), \quad (1)$$

where $\mathbf{h}_i = \frac{f^c(\mathbf{x}_i)}{\|f^c(\mathbf{x}_i)\|_2}$ is the normalized representation of a sample $\mathbf{x}_i$ of the client c. We add a fully-connected layer on top of the feature extractor to align the dimension of $\mathbf{h}_i$ and $\boldsymbol{e}_m^{(\mathbf{y}_i)}$. $\boldsymbol{\Omega}^{(y_i)}$ is the positive embedding set of the class $\mathbf{y}_i$ and contains the embeddings $\{\boldsymbol{e}_1^{(\mathbf{y}_i)}, \boldsymbol{e}_2^{(\mathbf{y}_i)}, ..., \boldsymbol{e}_M^{(\mathbf{y}_i)}\}$. $\boldsymbol{\Theta}^{(\mathbf{y}_i)}$ is the negative embedding set and contains the text embeddings of the other categories. τ is the temperature coefficient. Generally, more diverse prompts can obtain rich text embeddings (corresponding to language descriptions) to comprehensively describe one category. Aligning image representations to these text embeddings can force the model to learn exhaustive visual details. Hence, the performance of the model f^c highly depends on the number M of prompts under the supervision of Eq. (1). However, it is difficult to obtain all prompts for a specific task via prompt engineering. Besides, some inappropriate prompts should be removed to avoid misguiding feature learning. Considering these issues, we propose to further generalize Eq. (1) to the infinite space, namely, aligning the image representations and the text embedding distribution of each class.

Assume that the text embeddings $\{\boldsymbol{e}_1^{(k)}, \boldsymbol{e}_2^{(k)}, ..., \boldsymbol{e}_M^{(k)}\}$ of the k-th class are sampled from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, we compute the mean $\boldsymbol{\mu}_k$ and the variance $\boldsymbol{\Sigma}_k$ as follows:

$$\boldsymbol{\mu}_k = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{e}_m^{(k)}, \qquad \boldsymbol{\Sigma}_k = \frac{1}{M-1}\sum_{m=1}^{M} (\boldsymbol{e}_m^{(k)} - \boldsymbol{\mu}_k)(\boldsymbol{e}_m^{(k)} - \boldsymbol{\mu}_k)^{\mathrm{T}}. \quad (2)$$

After estimating the distributions $\{\mathcal{N}^{(k)}\}_{k=1}^{K}$ of all classes, we can sample infinite text embeddings, which correspond to instances with different characteristics in the image space. In the context, Eq. (1) can be reformulated as

$$\mathcal{L}_{align}^{\infty} = \frac{1}{N_c}\sum_{i=1}^{N_c} \mathbb{E}_{\boldsymbol{e}^{(\mathbf{y}_i)} \sim \mathcal{N}^{(\mathbf{y}_i)}} \left(-\log \frac{e^{\tau \mathbf{h}_i^{\mathrm{T}} \boldsymbol{e}^{(\mathbf{y}_i)}}}{e^{\tau \mathbf{h}_i^{\mathrm{T}} \boldsymbol{e}^{(\mathbf{y}_i)}} + \sum_{k \neq \mathbf{y}_i}^{K} \mathbb{E}_{\boldsymbol{e}^{(k)} \sim \mathcal{N}^{(k)}} e^{\tau \mathbf{h}_i^{\mathrm{T}} \boldsymbol{e}^{(k)}}} \right). \quad (3)$$

$\mathcal{L}_{align}^{\infty}$ is difficult to compute its exact form when the sampled text embeddings are infinite. Here, we can derive its upper bound based on [19] and find a surrogate loss $\overline{\mathcal{L}}_{align}^{\infty}$:

$$\mathcal{L}_{align}^{\infty} \leq \overline{\mathcal{L}}_{align}^{\infty} = \frac{1}{N^c}\sum_{i=1}^{N_c} \left(-\log \frac{e^{\mathcal{F}(\mathbf{h}_i, \mathbf{y}_i)}}{\sum_{k=1}^{K} e^{\mathcal{F}(\mathbf{h}_i, k)}} + \frac{\tau^2}{2} \mathbf{h}_i^{\mathrm{T}} \boldsymbol{\Sigma}_{(\mathbf{y}_i)} \mathbf{h}_i \right), \quad (4)$$

where $\mathcal{F}(\mathbf{h}_i, k) = \mathbf{h}_i^{\mathrm{T}} \boldsymbol{\mu}_{(k)} + \frac{1}{2}\tau \mathbf{h}_i^{\mathrm{T}} \boldsymbol{\Sigma}_{(k)} \mathbf{h}_i$. The detailed derivation is shown in the supplementary. By minimizing the loss $\overline{\mathcal{L}}_{align}^{\infty}$, we can implement the alignment between the image representations and the text embedding distributions. It can be observed that $\overline{\mathcal{L}}_{align}^{\infty}$ is a softmax-based cross-entropy loss over $\mathcal{F}(\mathbf{h}_i, k)$, with a constraint on variance of features. Therefore, we redefine the local classifier as

$$\mathcal{F}(\mathbf{h}) = \mathbf{h}^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\tau\mathbf{h}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{h}. \tag{5}$$

Compared with the naive method of averaging the text embeddings as the local classifier, the classifier $\mathcal{F}(\mathbf{h})$ in Eq. (5) considers the variance of the text embeddings and thus are more robust to match the semantic diversification of image representations, thereby achieving more accurate classification.

3 Experiment

3.1 Dataset and Implementation Details

Dataset To investigate the effectiveness of our FedCB framework, we evaluate it on two public medical datasets, OCT-C8 [18] and Kvasir-v2 [15]. **OCT-C8** [18] consists of 24,000 retinal OCT images and is divided into eight categories: age-related macular degeneration (AMD), choroidal neovascularisation (CNV), diabetic macular edema (DME), drusen, macular hole (MH), diabetic retinopathy (DR), central serous retinopathy (CSR) and one for healthy classes. Based on the official division, 18400 images are used for training, 2800 for validation, and 2800 for testing. **Kvasir-v2** [15] contains 8000 endoscopic images of the gastrointestinal tract, which belong to 8 categories: esophagitis, cecum, pylorus, Z-line, polyps, ulcerative colitis, dyed lifted polyp, dyed resection margin. We randomly partition all samples into training, validation, and test sets with a ratio of 7 : 1 : 2. The prompts of two datasets are shown in supplementary.

Implementation Details The proposed FedCB and comparison methods are implemented with PyTorch library. We adopt the ResNet-18 [5] as the backbone network of all methods. The number of clients is set to 12 and 10 for OCT-C8 and Kvasir-v2 datasets, respectively. For two datasets, we utilize the Adam [8] optimizer with the initial learning rate of 1×10^{-2}. The batch size is set to 8 and the learning rate decays at a rate of 0.99 per epoch. The numbers of local epochs and communication rounds are 2 and 200, respectively. The default client sampling ratio is 0.5. Similar to existing FL works [13,12], we use Dirichlet distribution on label ratios to simulate the Non-IID data distribution among clients. We set the Dirichlet parameter β as 0.05 and 0.1 to ensure the high data heterogeneity. Two commonly-used metrics, accuracy, and F1 score, are used to measure the classification performance. In all the experiments, we conduct three trials for each setting and present the mean and the standard deviation.

3.2 Comparison with State-of-the-art Methods

To evaluate the performance of our FedCB framework, we perform a comprehensive comparison with the state-of-the-art FL methods on OCT-C8 and Kvasir-v2 datasets, including FedAvg [14], FedDYN [1], FedPROX [10], FedREP [3], FedROD [2] and FedETF [12].

Table 2. The performance comparison of the proposed method and existing methods on OCT-C8 dataset.

Methods	$\beta = 0.05$		$\beta = 0.1$	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
FedAvg [14]	74.82±5.98	72.34±6.87	78.64±5.44	76.25±7.23
FedDYN [1]	70.79±1.94	65.88±5.12	73.46±5.49	69.86±7.96
FedPROX [10]	76.60±5.43	74.49±6.12	78.37±5.76	75.64±7.68
FedREP [3]	43.87±7.24	32.98±8.83	59.37±12.81	55.20±12.56
FedROD [2]	70.20±4.17	64.24±4.21	79.11±5.62	77.65±7.17
FedETF [12]	77.79±5.17	74.47±8.64	82.81±3.76	81.67±5.31
FedCB	**79.14±3.77**	**77.13±5.10**	**85.00±2.66**	**84.56±2.99**

Table 3. The performance comparison of the proposed method and existing methods on Kvasir-v2 dataset.

Methods	$\beta = 0.05$		$\beta = 0.1$	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
FedAvg [14]	60.10±5.76	54.07±9.58	67.02±1.72	63.93±2.49
FedDYN [1]	55.20±3.03	49.84±3.86	63.18±2.19	60.98±1.41
FedPROX [10]	59.68±2.02	53.11±3.26	68.77±1.45	66.81±2.53
FedREP [3]	33.06±15.71	23.88±13.72	48.95±0.62	39.71±2.50
FedROD [2]	61.79±2.72	58.83±3.40	70.10±3.70	68.01±5.71
FedETF [12]	63.77±3.73	60.12±6.34	69.70±3.56	67.15±6.64
FedCB	**65.00±4.36**	**61.67±6.61**	**70.90±1.97**	**68.73±3.36**

On OCT-C8 dataset, our method achieves the best performance under different Non-IID settings as illustrated in Table 2, with the overwhelming average F1 of 77.13% ($\beta = 0.05$) and 84.56% ($\beta = 0.1$). Noticeably, our framework exceeds the baseline FL method, FedAvg [14], by a large margin, e.g., 4.79% in average F1 ($\beta = 0.05$) and 6.36% in average accuracy. This advantage confirms that the proposed FedCB can alleviate the classifier biases problem. Moreover, compared with FedETF [12] that employs orthogonal initialization to build local classifiers, our method obtains superior performance with a remarkable increase of 2.66% ($\beta = 0.05$) and 2.89% ($\beta = 0.1$) in average F1. On Kvasir-v2 dataset, FedCB also outperforms existing FL methods as shown in Table 3, obtaining the best performance, with the average accuracy of 65.00% ($\beta = 0.05$) and 70.90% ($\beta = 0.1$). Although FedROD [2] yields the similar performance to the proposed FedCB in the setting of $\beta = 0.05$, it undergoes a serious performance degradation when the data are more heterogeneous, with a remarkable decrease of 8.31% in average accuracy. By comparison, FedCB is more robust against data heterogeneity and only suffers from a decrease of 5.9% in average accuracy. These experimental

Table 4. The performance of the proposed FedCB framework with different proportions of prompts. FedCB(25%) indicates that only 25% of prompts are used.

Methods	$\beta = 0.05$		$\beta = 0.1$	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
FedCB(25%)	76.64±6.26	76.05±6.48	83.71±2.83	83.45±3.05
FedCB(50%)	76.82±5.41	76.54±4.88	82.40±3.17	82.23±3.30
FedCB(75%)	77.70±3.53	75.71±4.51	84.12±4.18	83.39±4.79
FedCB(100%)	79.14±3.77	77.13±5.10	85.00±2.66	84.56±2.99

Table 5. The performance of different methods on OCT-C8 dataset.

Methods	$\beta = 0.05$		$\beta = 0.1$	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Baseline	74.82±5.98	72.34±6.87	78.64±5.44	76.25±7.23
Embed. Average	76.75±5.00	76.34±4.38	84.22±2.38	83.81±2.64
Embed. Distribution	79.14±3.77	77.13±5.10	85.00±2.66	84.56±2.99

results on two datasets demonstrate the performance advantage of our method over state-of-the-art FL methods under different Non-IID settings.

3.3 Ablation Study

The Impact of the Prompt Number To study the impact of the prompt number, we compare the performance of FedCB with different proportions of prompts in Table 4. With the proportion of prompts increases, all metrics generally show an increasing trend. When all prompts are used to construct local classifiers, FedCB achieves the highest performance. The experiment results demonstrate the importance of the number of prompts.

Embedding Average VS **Embedding Distribution** Instead of averaging text embeddings as local classifiers, FedCB uses text embeddings distribution of each category. We compare the performance of these two strategies and the baseline (FedAvg) on OCT-C8 dataset. As shown in Table 5, both these two strategies outperform the baseline by a large margin, highlighting the effectiveness of using text embedding as local classifiers to mitigate classifier biases. Meanwhile, embedding distribution is superior to embedding average in different Non-IID settings. This is because using embedding distribution as local classifiers can help the model to capture the semantic diversification of image representations.

4 Conclusion

In this paper, we propose a novel framework, called Federated Classifier deBiasing (FedCB), to solve the classifier biases problem in heterogeneous federated learning. In FedCB, the server side first collects the class concepts from clients. Then, a set of prompts are employed to contextualize these concepts to generate the corresponding language descriptions. These descriptions are input into a pre-trained language model to obtain the text embeddings. The generated embeddings are used to estimate the distribution of each category in the semantic space. Regarding these distributions as the local classifiers, we perform the alignment between the images representations of each class and the corresponding semantic distribution. The experimental results on two public datasets show the superior performance of FedCB in contrast to state-of-the-art methods under different Non-IID settings.

Acknowledgments. This work is supported by the Science and Technology Innovation Committee of Shenzhen under Grant Type-C SGDX20210823104001011, the Innovation and Technology Commission-Innovation and Technology Fund ITS/229/22, and the Research Grant Council (RGC) of Hong Kong under Grant 11217922, 11212321, and 21212720.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Acar, D.A.E., Zhao, Y., Navarro, R.M., Mattina, M., Whatmough, P.N., Saligrama, V.: Federated learning based on dynamic regularization. arXiv preprint arXiv:2111.04263 (2021)
2. Chen, H.Y., Chao, W.L.: On bridging generic and personalized federated learning for image classification. ICLR (2021)
3. Collins, L., Hassani, H., Mokhtari, A., Shakkottai, S.: Exploiting shared representations for personalized federated learning. In: ICML. pp. 2089–2099 (2021)
4. Guo, Yongxin, T.X., Lin, T.: Fedbr: Improving federated learning on heterogeneous data via local learning bias reduction. In: ICML (2023)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
6. Jeong, E., Oh, S., Kim, H., Park, J., Bennis, M., Kim, S.L.: Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. arXiv preprint arXiv:1811.11479 (2018)
7. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. In: ICML. pp. 5132–5143 (2020)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: CVPR. pp. 10713–10722 (2021)
10. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems. vol. 2, pp. 429–450 (2020)
11. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. In: ICLR (2021)

12. Li, Z., Shang, X., He, R., Lin, T., Wu, C.: No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In: ICCV. pp. 5319–5329 (October 2023)
13. Luo, M., Chen, F., Hu, D., Zhang, Y., Liang, J., Feng, J.: No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. NeurIPS **34**, 5972–5984 (2021)
14. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. pp. 1273–1282. PMLR (2017)
15. Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.T., Lux, M., Schmidt, P.T., et al.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: MMSys. pp. 164–169 (2017)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
17. Razzak, M.I., Naz, S., Zaib, A.: Deep learning for medical image processing: Overview, challenges and the future. Classification in BioApps: Automation of Decision Making pp. 323–350 (2018)
18. Subramanian, M., Shanmugavadivel, K., Naren, O.S., Premkumar, K., Rankish, K.: Classification of retinal oct images using deep learning. In: ICCCI. pp. 1–7. IEEE (2022)
19. Wang, Y., Pan, X., Song, S., Zhang, H., Huang, G., Wu, C.: Implicit semantic data augmentation for deep networks. NeurIPS **32** (2019)
20. Xu, J., Tong, X., Huang, S.L.: Personalized federated learning with feature alignment and classifier collaboration. arXiv preprint arXiv:2306.11867 (2023)
21. Yan, Z., Yang, X., Cheng, K.T.: Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation. IEEE Transactions on Biomedical Engineering **65**(9), 1912–1923 (2018)
22. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Lungren, M., Naumann, T., Poon, H.: Large-scale domain-specific pretraining for biomedical vision-language processing (2023). `https://doi.org/10.48550/ARXIV.2303.00915`
23. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 (2018)
24. Zhu, M., Chen, Z., Yuan, Y.: Dsi-net: Deep synergistic interaction network for joint classification and segmentation with endoscope images. IEEE Transactions on Medical Imaging **40**(12), 3315–3325 (2021)
25. Zhu, M., Liao, J., Liu, J., Yuan, Y.: Fedoss: Federated open set recognition via inter-client discrepancy and collaboration. IEEE Transactions on Medical Imaging **43**(1), 190–202 (2024)