Attentive Modeling and Distillation for Out-of-Distribution Generalization of Federated Learning

Zhuang Qi†1, Weihao He†1, Xiangxu Meng† and Lei Meng‡1,2
1 School of Software, Shandong University, Jinan, China
2 Shandong Research Institute of Industrial Technology, Jinan, China
Email: z_qi@mail.sdu.edu.cn, rogerhe130@gmail.com, {mxx, lmeng}@sdu.edu.cn

Abstract—Out-of-distribution issues lead to different optimization directions between clients, which weakens collaborative modeling in federated learning. Existing methods aim to decouple invariant features in the latent space to mitigate attribute bias. However, their performance is limited by suboptimal decoupling capabilities in complex latent spaces. To address this problem, this paper presents a method, termed FedAKD, that adaptively identifies meaningful visual regions in images to guide the model in learning causal features. It includes two main modules, where the attentive modeling module adaptively locates critical regions to mitigate the negative impact of irrelevant elements, which are considered significant contributors to distribution heterogeneity. The attention-guided representation learning module leverages attentive knowledge to guide the local model to pay more attention to important regions, which acts as a soft attention regularizer to mitigate the trade-off between capturing category-relevant information and irrelevant contextual information in images. Experiments were conducted on four datasets, including performance comparison, ablation study, and case study. The results demonstrate that FedAKD can effectively enhance attention to causal features, which leads to superior performance compared with the state-of-the-art methods.

Index Terms—Federated Learning, Knowledge Distillation, Out-of-Distribution, Attentive modeling

I. INTRODUCTION

Federated learning has emerged as a promising distributed learning paradigm, enabling collaborative modeling with multiple data sources while preserving data privacy [1], garnering widespread attention across various fields [2]–[5]. It aggregates the parameters of local models trained on private data from multiple devices to obtain a global model, without involving data sharing [6]–[8]. Despite the advantages in preserving privacy that federated learning presents, it continues to confront significant challenges associated with data heterogeneity, such as out-of-distribution [1], [9], [10]. The huge attributes skew between data sources often harm the effectiveness of collaboration. This is primarily due to the disparate optimization directions among various local models, and the local model presents declining performance in other clients.

To mitigate the out-of-distribution issue, existing methods can roughly be divided into two groups: regularization-based representation alignment [11], [12] and representation decoupling of invariant attributes [13]–[16]. The former approaches aim to facilitate the learning of consistent knowledge between clients, to mitigate the adverse effects of inconsistent attributes on representation learning. They typically employ prototype-based representation alignment regularization to constrain the local training of clients. For example, FPL constructs unbiased prototypes and employs consistency regularization to align instances with the corresponding unbiased prototypes, which helps to alleviate feature heterogeneity between clients [12]. The latter methods focus on decoupling in the feature space to extract the invariant features, which contributes to the elimination of intervention from irrelevant contextual. For instance, DFL disentangles domain-specific and invariant attributes into two complementary branches, separating domain-specific attributes from model aggregation [14]. These methods offer insights into the efficacy of isolating domain-specific attributes locally to mitigate out-of-distribution issues. However, their performance is hampered by the limited ability to uncover causal relationships.

To address this issue, this paper presents a novel attentive knowledge distillation mechanism, termed FedAKD. As illustrated in Figure 1, compared with conventional methods, the proposed FedAKD effectively identifies the meaningful region in the image, which reduces the interference of irrelevant contextual noise. Specifically, FedAKD has two main modules, the attentive modeling (AM) module and the attention-guided representation learning (AGRL) module. To precisely identify key features and patterns of images, the AM module utilizes

† indicates equal contributions.
‡ indicates corresponding author.
self-learned geometric operators to adaptively locate image regions relevant to the class, which provides key guidance information for the model to learn causal features. Subsequently, the AGRRL module serves as a soft-attention regularizer to the local model in the task channel by aligning its visual features to the attentive features produced in the attentive channel, which effectively filters out irrelevant distractions. Notably, the AM module, as a plug-and-play component, can be easily integrated into various methods. As observed, FedAKD enhances the generalization across different domains for models.

Extensive experiments are conducted on four datasets in terms of performance comparison, ablation study of the key components, and case study for the effectiveness of key region extraction and recognition. The results verify that modeling meaningful attention in the input space can guide models to learn more robust features, which enhances their generalization capabilities. To summarize, this paper includes two main contributions:

- This paper presents a model-agnostic attentive knowledge distillation mechanism, termed FedAKD. To the best of our knowledge, it is the first method that modeling meaningful attention in the input space to alleviate the attributes skew problem between clients in federated learning.
- We propose a plug-and-play module, named the AM module, which can be easily integrated into various methods to enhance their performance, significantly improving the quality of representation learning.

II. RELATED WORK

A. Federated Learning with Non-IID Data

To tackle data heterogeneity in federated learning, commonly used methods are generally divided into two groups: one focuses on reducing local biases, while the other aims to improve aggregation efficiency. The former employs regularization or cross-training to help clients acquire comprehensive knowledge. Regularization comes in three types: weight-based [17], [18], feature-based [19], [20], and prediction-based [21], [22]. The latter approach believes that directly averaging local model parameters can harm performance. Instead, they either design better aggregation methods or fine-tune models on the server. For example, Elastic aggregation crops out the main object from the image. Inspired by Spatial Transformer Network (STN) [28], the AM module adaptively locates visual attentive regions based on classification loss. The localization network generates an affine transformation to generate a grid generator $G(\cdot, \cdot)$, a sampler.

The localization network generates an affine transformation matrix $\theta$ for each image $I$ to capture detailed regions in the original image. It can be formulated as follows:

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} = F_{\text{loc}}(I) \quad (1)$$

where $\{\theta_{11}, \theta_{22}\}, \{\theta_{12}\}, \{\theta_{21}\}$ and $\{\theta_{13}, \theta_{23}\}$ represent scaling, rotation, shearing and translation parameters, respectively. $F_{\text{loc}}(\cdot)$ is typically a lightweight network.

The grid generator $G(\cdot, \cdot)$ utilizes the affine transformation operator $T_{\theta}$ and transformation parameter $\theta$ to generate a transformed coordinate for image $I$, defined as

$$\begin{bmatrix} x'_{ji} \\ y'_{ji} \end{bmatrix} = G(T_{\theta}, x) = T_{\theta}(x) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{bmatrix} x_j \\ y_j \\ 1 \end{bmatrix} \quad (2)$$

where $(x_{ji}, y_{ji})$ and $(x'_{ji}, y'_{ji})$ denote coordinates of each pixel in the input image $I$ and transformed image $I_{\text{trans}}$. $j$ denotes the index of pixel.

III. METHODOLOGY

A. Overall Framework

The proposed attentive knowledge distillation scheme in federated learning (FedAKD), outlined in Figure 2, has two main phases: local training and global aggregation. FedAKD places particular emphasis on the local training within its dual-channel framework, where the Attentive Channel utilizes the attentive modeling (AM) module to dynamically extract class-aware region to provide the fine-grained knowledge. The Task Channel employs the attention-guided representation learning (AGRL) module to fully leverages attentive knowledge to help the model optimize the feature extraction and recognition process. Subsequently, FedAKD aggregates all local models to generate the global model in the server.

B. Attentive Modeling

The Attention Modeling (AM) module aims to transform the original image $I$ into one that is more focused on the meaningful object $I_{\text{trans}}$, i.e., $I_{\text{trans}} = AM(I)$. It can provide instructive knowledge to the local student model to mitigate the interference from backgrounds. An intuitive idea is to crop out the main object from the image. Inspired by Spatial Transformer Network (STN) [28], the AM module adaptively locates visual attentive regions based on classification loss. Specifically, it involves a localization network $F_{\text{loc}}(\cdot)$, a grid generator $G(\cdot, \cdot)$, and a sampler.
The sampler extracts pixels from the input image $I$, and produces the output image $I_{trans}$ through bilinear interpolation, enabling differentiable spatial transformations, defined by

$$I_{trans}(x'_j,y'_j) = \sum_{i} \sum_{n} I(n,m) \cdot \max(0,1 - |x'_j - n|) \cdot \max(0,1 - |y'_j - m|) \quad (3)$$

This shows that bilinear interpolation calculates the output pixel value $I_{trans}(x'_j,y'_j)$ at the transformed coordinates by taking a weighted average of the input image pixel values $I(n,m)$, where the weights decrease with the distance from the input image pixels to the transformed coordinate point.

Meanwhile, the classification loss is used to optimize the model of the attentive channel, i.e.,

$$L_{cls}^{att} = L_{CE}(\hat{y}_{trans}, y) \quad (4)$$

where $\hat{y}_{trans} = F_a(E_a(AM(I)))$ is the prediction, $F_a$ and $E_a$ are the classifier and the feature extractor in the attentive channel. $L_{CE}$ denotes the cross-entropy loss, $y$ is the label of the original image $I$ and the transformed image $I_{trans}$.

### C. Attention-Guided Representation Learning

The Attention-Guided Representation Learning (AGRL) module aims to use the attentive knowledge from the output of AM module as a soft regularizer to guide the training of the local student model. It effectively guides the student model to focus on key information by utilizing features from the most task-relevant regions in the image. Specifically, the AGRL module aligns the visual features output by the student model in the task channel with the fine-grained features generated by the teacher model in the attentive channel. To achieve this, the AGRL module utilizes KL Divergence (Kullback-Leibler Divergence) [29] as a measure to encourage the student model to adapt its feature representation to align more closely with that of the teacher model. Therefore, the alignment loss function can be defined as

$$L_{align} = D_{KL}(f_t || f_a) \quad (5)$$

where $f_t = E_t(I)$ and $f_a = E_a(I_{trans})$ denote the feature output of local student model and the teacher model, respectively. $E_t$ is a feature extractor in the task channel.

Subsequently, we use the empirical classification loss $L_{cls}^{task}$ to optimize the local student model, i.e.,

$$L_{cls}^{task} = L_{CE}(F_t(f_t), y) \quad (6)$$

where $F_t$ denotes the classifier in the task channel.

### D. Training Strategy of FedAKD

FedAKD focuses on optimizing the extraction of meaningful object regions in the attentive channel, its corresponding optimization objective is

$$L_{att} = E_{(x,y)\sim D_{local}}[L_{cls}^{att}] \quad (7)$$

where $x$ and $y$ denote the original image and the corresponding label in the local dataset $D_{local}$. Meanwhile, FedAKD aims to enhance the local student model’s ability to focus on important objects within the task channel, the corresponding objective is to minimize:

$$L_{task} = E_{(x,y)\sim D_{local}}[L_{cls}^{task} + \lambda_{align} \cdot L_{align}] \quad (8)$$

where $\lambda_{align}$ is a weight parameter.
A. Experiment Settings

1) Datasets: We validate the efficacy of the proposed framework through experiments conducted on two out-of-distribution generalization datasets, namely COLORMNIST [33] and NICO-Animal [34]. Additionally, we assess its performance on two well-established datasets commonly employed in federated learning, CIFAR10 and CIFAR100 [35]. The statistical details of these datasets are summarized in Table II.

2) Network Architecture: For a fair comparison, all methods share a common network architecture. And in the FedAKD, we maintain consistent architecture for both the student model in the task channel and the teacher model in the attention channel. For COLORMNIST, the architecture involves a convolutional layer serving as an image encoder and a 2-layer MLP as the classifier. Following previous works [19], [36], we employ ResNet-18 [37] as the network backbone for all other datasets. Notably, we adapt the first convolutional kernel size from 7 to 3 for CIFAR10 and CIFAR100, while keeping it at 7 for the NICO-Animal dataset. For the the model used in attentive modeling, it involves two convolutional layers for localisation and a fully connected layer to generate transformation parameter.

3) Hyper-parameter Settings: For all methods, we maintain consistency in hyperparameter settings across experiments. The local training epoch is fixed at 10 for each global round, with the number of clients set to 10 for CIFAR10, CIFAR100, COLORMNIST, and 7 for NICO-Animal, along with a sample fraction of 1.0. The local optimizer employed is the SGD algorithm, and the communication round is set to 100. During local training, we configure the weight decay to $1e-05$ and the batch size to 64. The learning rate is initialized at 0.01, and the Dirichlet parameter $\beta$ is set to 0.1 and 0.5 for CIFAR10, CIFAR100 and COLORMNIST. Furthermore, $\lambda_{align}$ is fine-tuned from the set $\{0.01, 0.05, 0.1, 0.5\}$. The remaining hyperparameters follow the specifications outlined in the corresponding paper.

B. Performance Comparison

We compare FedAKD with eight SOTA methods, including FedAvg [30], MOON [19], Fedprox [25], Fedproc [26], FedNTD [31], FPL [12], DaFKD [13] and FedIR [32]. These methods typically use FedAvg as the base algorithm, so we integrate key modules into FedAvg to form FedAKD$_{FedAvg}$. Additionally, we incorporate the proposed AM module into the global branch of MOON, forming FedAKD$_{MOON}$. The following results can be derived from Table I.

- Both FedAKD$_{FedAvg}$ and FedAKD$_{MOON}$ demonstrate significant improvements in classification accuracy over their respective baseline models. This highlights the model-agnostic nature of the FedAKD approach.
- FedAKD consistently outperforms other methods in terms of classification accuracy. This is understandable since FedAKD can capture meaningful objects in images while mitigating the negative effects of irrelevant elements.
- The causal modeling approach focuses on extracting key information in latent spaces and eliminating the interference of irrelevant elements, which provides a meaningful direction for mitigating the out-of-distribution issues. Notably, a refined causal feature learning could further enhance performance.
- Incorporating prototypical contrastive learning to guide different clients in learning consistent class-level representations proves to be more effective in CIFAR10 and CIFAR100 than in other datasets (Fedproc and FPL). This is due to the increased difficulty of learning consistent features under significant attribute variance.

C. Ablation Study

This section further studies the effectiveness of different modules of FedAKD. The results are summarized in Table III, including
- Incorporating the Attentive Modeling (AM) and Attention-Guided Representation Learning (AGRL) modules significantly improve the performance, indicating their role in enhancing causal discovery.
- Leveraging different loss functions (KL, L2 and JS) for attention-guided representation learning yielded similar
TABLE III
ABLATION STUDY ON THE EFFECTIVENESS OF DIFFERENT COMPONENTS OF FEDAKD ON THE NICO-ANIMAL AND COLORMNIST.

<table>
<thead>
<tr>
<th>Task Channel</th>
<th>NICO-Animal</th>
<th>COLORMNIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>29.84±0.6</td>
<td>57.48±0.4</td>
</tr>
<tr>
<td>+ AM_input + AGRL_d</td>
<td>33.12±0.5</td>
<td>57.98±0.7</td>
</tr>
<tr>
<td>+ AM_layer + AGRL_d2</td>
<td>32.61±0.7</td>
<td>58.04±0.9</td>
</tr>
<tr>
<td>+ AM_layer + AGRL_d4</td>
<td>32.92±0.5</td>
<td>58.13±0.9</td>
</tr>
<tr>
<td>+ AM_input + AGRL_d2</td>
<td>34.97±0.6</td>
<td>58.31±0.3</td>
</tr>
<tr>
<td>+ AM_input + AGRL_d4</td>
<td>34.72±0.4</td>
<td>59.15±0.5</td>
</tr>
<tr>
<td>+ AM_input + AGRL_d4</td>
<td>35.56±0.4</td>
<td>59.03±0.2</td>
</tr>
</tbody>
</table>

4(a) illustrates that both FedAvg and FedAKD made correct predictions, with FedAKD achieving a more precise focus on the object. However, as shown in Figure 4(b), FedAvg struggles with complex contexts, while FedAKD accurately focuses on the causal region. This can be attributed to the guidance provided by attentive modeling. In the Figure 4(c), FedAKD faces challenges with undue attention to context, making it difficult to distinguish between the object and context. FedAvg also fails to focus on the core object despite a correct prediction. As illustrated in Figure 4(d), both FedAvg and FedAKD encounter challenges in focusing on the object within intricate contexts. However, FedAKD demonstrates reduce the attention to the irrelevant region, which decrease the prediction disparity between 'dog' and the top-1 category. This highlights the advantage of FedAKD in federated classification.

V. CONCLUSION

This paper presents a novel attentive modeling and distillation mechanism in federated learning, termed FedAKD, to handle the out-of-distribution issue. It performs attention-guided representation learning to instruct the local models to focus on the meaningful objects within the images. Experimental results show that FedAKD can effectively improve the performance by focusing on the important regions. This enhances the generalization ability of local models and the collaborative effect among them.

There are some directions for further exploration in this study. First, stronger attentive modeling techniques [39], [40] that more accurately identify causal regions can provide the meaningful information. Second, better feature learning methods can further improve the performance [41]–[45]. Third, it is anticipated that applying FedAKD to some challenging tasks would be promising [46]–[55].

REFERENCES
