Disentangled Federated Learning for Tackling Attributes Skew via Invariant Aggregation and Diversity Transferring

Zhengquan Luo [1 2] **Yunlong Wang** [2] **Zilei Wang** [1] **Zhenan Sun** [2] **Tieniu Tan** [2]

Abstract

Attributes skew hinders the current federated learning (FL) frameworks from consistent optimization directions among the clients, which inevitably leads to performance reduction and unstable convergence. The core problems lie in that: 1) Domain-specific attributes, which are non-causal and only locally valid, are indeliberately mixed into global aggregation. 2) The one-stage optimizations of entangled attributes cannot simultaneously satisfy two conflicting objectives, i.e., generalization and personalization. To cope with these, we proposed disentangled federated learning (DFL) to disentangle the domain-specific and cross-invariant attributes into two complementary branches, which are trained by the proposed alternating local-global optimization independently. Importantly, convergence analysis proves that the FL system can be stably converged even if incomplete client models participate in the global aggregation, which greatly expands the application scope of FL. Extensive experiments verify that DFL facilitates FL with higher performance, better interpretability, and faster convergence rate, compared with SOTA FL methods on both manually synthesized and realistic attributes skew datasets.

1. Introduction

Federated learning (McMahan et al., 2017), as a decentralized and privacy-preserving machine learning framework, aims to build a client-server system that can adapt to different distributions without access to local raw data. A key challenge lies in the Non-i.i.d factors between clients' data domains. The huge distribution shift comes from the diverse contextual information across devices/locations, which causes suboptimal or detrimental performance. In addition, the aggregated global model could be unstable and even nonconvergent due to the different optimization directions of client models. To mitigate this problem, a plethora of techniques are introduced to adjust the system, such as local model adaptation, global reweighting aggregation, and optimization direction correction. These methods make FL rapidly adapt to the local distributions and corresponding tasks. Although effective, the basic problem still exists. Especially, the attributes skew, as one challenging factor of Non-i.i.d, indicates the scenarios in which the representation distribution across attributes on each client is different from each other. The domain-specific attributes are inevitably extracted by the single-branch client model and mixed into the global aggregation. These attributes are decision-correlated but not causal, which is only locally valid. It leads to essential differences in the optimization directions of client models. Taking bird classification as an example in the left part of Figure 1. The attributes of the bird itself and the blue sky are extracted by the client

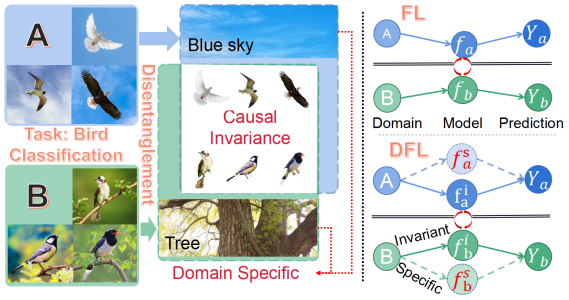

Figure 1. Taking bird classification as an example to illustrate the principle of DFL. The domain-specific attributes, such as the blue sky or trees, are stripped out and contributed locally. The invariant aggregation employs the partial client model, which only focuses on the invariant and causal attributes, like the bird itself. The two-branch local model is employed in DFL to replace the single-branch in FL, which shows in the right part.

[1]University of Science and Technology of China (USTC) [2]Institute of Automation, Chinese Academy of Sciences (CASIA). Correspondence to: Zhengquan Luo <zhengquan.luo@cripac.ia.ac.cn>, Yunlong Wang <yunlong.wang@cripac.ia.ac.cn>, Zilei Wang <zlwang@ustc.edu.cn>, Zhenan Sun <znsun@nlpr.ia.ac.cn>, Tieniu Tan <tnt@nlpr.ia.ac.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

network simultaneously by the local extractor, which both contribute to the task decision in the flying birds' domain. Unfortunately, this blue sky as domain-specific attributes will be infused into the global server by model aggregation. It causes performance degradation to the domain where birds all sit on tree branches. Because there is no blue sky in this data domain. This phenomenon of attributes skew distribution is widespread, which causes concerns about the robustness and the trustworthiness of the FL system.

To mitigate this negative transfer[1] caused by attributes skew, we proposed disentangled federated learning (DFL). The motivation is to spin off the domain-specific attributes from model aggregation. However, the single-branch-based client model of the traditional FL framework cannot support DFL. The reason is that the specific and invariant attributes are entangled and extracted by a single extractor. Although these two attributes both contribute to the task decision locally, the huge differences are that: 1) invariant attributes are intrinsic and causal, which is cross-domain generic; 2) specific attributes are only locally valid, which may bring performance degradation to other domains. Thus, DFL applied two complementary branches client model, which are presented in the right part of Figure 1. These two attributes are disentangled by mutual information (MI) constraints and focused by two branches respectively. 1) Domain-specific branch is only trained locally. 2) Domain-invariant branch is applied for global aggregation. Except for redesigning the local model, the other two innovations are proposed as invariant aggregation and diversity transferring.

Invariant aggregation is proposed to employ the local invariant branch for global model aggregation. The MI maximization between the local invariant branch and the global invariant model restricts clients' optimizations in the same direction. The combination of the invariant aggregation and MI constraint drives the local invariant branch to focus on the intrinsic and causal attributes (e.g. bird itself). In the theory of (Schölkopf et al., 2012; Peters et al., 2016), these cross-domain invariant attributes can provide transferable and reliable knowledge, which leads to proper domain adaptation and lighter catastrophic forgetting. In addition, MI minimization is employed to disentangle the invariant and specific locally. Although these domain-specific attributes are dropped out from model aggregation, they can still contribute to the final decision (e.g., things that fly in the blue sky are most likely birds). It is not wise to throw these task-correlated domain-specific attributes away directly. **Diversity transferring** is proposed to make full use of these attributes, enhancing the diversity of representation. In theoretical analysis, this diversity augmentation has proved to mitigate overfitting and correct inaccurate distributions (Shorten & Khoshgoftaar, 2019). Specifically, the diverse representations are extracted by combining the local invariant extractor and transferred specific extractors. Thus, the local client model is forced to pay attention to these tailed attributes, which the local extractor may ignore but has a decision contribution (e.g. trees may appear in the blue-sky domain).

We emphasize that two-stage training in an alternating manner is an essential innovation because it changes the optimization purpose of FL. Compared with one-stage global optimization, alternating optimization easily finds the optimal solution of the global invariant model based on multiple local optimal points. Specifically, part of the optimization process is separated into local clients and is personalized trained. It provides both generalized adaption and personalized performance. Besides, the theoretical analysis proves that DFL is convergent even if incomplete client models participate in the global aggregation, based on the bounded gradient of the remaining model part. As far as we know, this work is the first to provide the convergence guarantee of partial aggregation. With sufficient theoretical guarantees, we design the following experiments: 1) **Clarification experiments** demonstrate the unstable convergence and performance degradation with the introduction of manually synthesized attributes skew in Colored-MNIST (Arjovsky et al., 2019); 2) **Verification experiments** verify the superiority of DFL on convergence rate and classification accuracy, compared with other SOTA personalized FL methods. In addition, the ablation study proves the complementary effectiveness of invariant aggregation and diversity transferring; 3) **Application experiments** on DomainNet (Peng et al., 2019a) and Office-Caltech (Gong et al., 2012) point out that DFL can adapt to the realistic attributes skew. The accuracies are all greatly improved in different backbones. The visualization on DomainNet verifies that the specific and invariant attributes can be successfully disentangled, proving the interpretability improvement by DFL. In summary, the main contributions of the paper are as follows:

- Disentangled federated learning (DFL) is proposed to overcome the attributes skew essentially, which spins off the domain-specific attributes from model aggregation.
- Theory deduction proves the convergence analysis of invariant aggregation based on the bounded local-specific gradient.
- Invariant aggregation and diversity transferring are proposed to correct the optimization directions and augment representation diversity.
- Alternating local-global optimization is proposed to simultaneously meet generalized adaption and personalized performance.

[1]Negative transfer, i.e., leveraging other clients' knowledge undesirably reduces the learning performance of the local client (Wang et al., 2019b).

2. Related Work

Federated learning research with clients' Non-i.i.d distributions aims to enhance the stability and convergence of FL system which suffers from distribution shift caused by Non-i.i.d factors. Several kinds methods have been proposed from different perspectives: 1) **Local model adaptation** is proposed to adjust the local model for heterogeneous distributions, such as fine-tuning (Wang et al., 2019a), Meta-learning-based different initialization (Fallah et al., 2020; Chen et al., 2018), and personalized prediction layers (Arivazhagan et al., 2019; Liang et al., 2020), etc. 2) **Global reweighting aggregation** is introduced to employ different global aggregated weights for each client. The weights can be based on data distribution similarity (Huang et al., 2021), local model contribution differences (Zhang et al., 2020), and same consensus focus (Feng et al., 2020), etc. 3) **Optimization direction correction** is proposed to mitigate the gap between local and global models. The additional constraints are introduced to the loss function or optimization, such as regularization terms (Hanzely & Richtárik, 2020), proximal terms (Li et al., 2018), gradient correction (Acar et al., 2021), Moreau Envelopes (Dinh et al., 2020), attentive message (Huang et al., 2021), control variate (Karimireddy et al., 2020), etc. Some adversarial-based methods are proposed, such as domain adaptation (**?**) and debiasing (Hong et al., 2021).

MI-based disentanglement aims at interpreting underlying interaction factors. MI can measure the degree of interdependence between any two variables. MI can be applied to 1) quantify the separation of distributions in learning binary hash codes by (Cakir et al., 2017; 2019); 2) perform unsupervised learning (Hjelm et al., 2018); 3) disentangle the specific features in advertising training by (Peng et al., 2019b). In addition, the MI maximization techniques are expended by (Belghazi et al., 2018) via a neural network to estimate MI between two random variables. In this work, MI is employed for two intentions: 1) disentanglement of local specific and invariant attributes; 2) similarity enhancement of local and global invariant model.

3. Disentangled Federated Learning

Rethinking the limitation of single-branch model sharing, we proposed alternating training, which changes the optimization purpose in Section 3.1. Two innovations, invariant aggregation and diversity transferring, are proposed to mitigate the attributes skew, and the MI-based disentanglement technique is introduced in Section 3. DFL supported by these methods disentangled the specific and invariant attributes. First, we provide the convergence analysis of DFL in Section 3.3, which proves that the FL system is convergent even if the shared model is incomplete.

3.1. Definition of Alternating Optimization

The task of FL can be defined as follows:

$$\min_{\omega} \left\{ f(\omega) := \frac{1}{N} \sum_{k=1}^{N} h_k(\omega) \right\} \tag{1}$$

The optimization of a local client is to minimize the loss of each client $k \in |K|$. The key problem of this purpose is the Non-i.i.d dilemma in FL. Specifically, the personalized loss minimization pushes the client parameter converging to the local optimal point following the client data distribution D_k^*. However, the model aggregation of FL drives server model gradient descent toward the global direction, which has a huge direction shift from the local client's optimization $D_i^* \neq D_j^* \neq D^*, 1 \leqslant i \neq j \leqslant K$. The reason is that the domain-specific and invariant attributes are entangled and indiscriminately extracted by the single-branch local extractor. It leads to generalization problems. The MI-based disentanglement is introduced to disentangle these two attributes into two local branches, which are optimized independently. The client network is divided as representation extractors, i.e., the invariant branch E_c^k and the specific branch E_s^k. The prediction module P^k takes the concatenated representations from these two branches as input for final decisions. The entire model framework is shown in Figure 2.

It should be emphasized that attribute disentanglement is leveraged to break through the limitation of one-stage global optimization of traditional FL methods. One-stage global optimization strives to find an optimal solution that simultaneously meets two conflicting objectives: generalized adaption and personalized performance. However, such efforts are usually in vain. To overcome this problem, we proposed two-stage alternating optimization. Specifically, only the invariant extraction branch of each client participates in the global model aggregation. The specific branches are optimized locally. The optimization purpose of DFL is changed to:

$$\min_{\omega_c} \begin{Bmatrix} f(\omega_c) := \frac{1}{N} \sum_{k=1}^{N} \min_{\omega_{k,s}} h_k(\omega_i) \\ \omega_i = M(\omega_c, \omega_{k,s}) = P_c \omega_c + P_s \omega_{k,s} \end{Bmatrix} \tag{2}$$

where M represents the model combination of two branches, ω_c is the parameter of the aggregated invariant model. h_k is local loss function. P_c and P_s are the weighting vectors of an invariant and specific model in local combinations. The minimized local specific branch model is defined as $\omega_{k,s}^*$, which satisfies the condition as:

$$\omega_{k,s}^* = \arg\min_{\omega_{k,s}} h_k(M(\omega_c, \omega_{k,s})) \tag{3}$$

Compared with the one-stage methods, the optimization of local specific branches in DFL is locally trained at first.

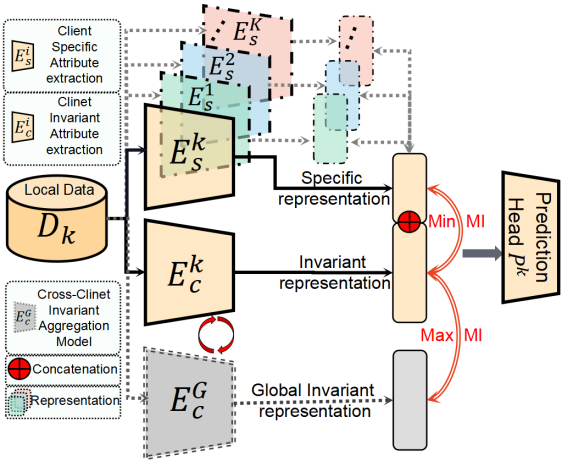

Figure 2. The framework of DFL.

Then, the entire local model parameter is $\omega_k^* = M(\omega_c, \omega_{k,s}^*)$ based on multiple optimal points $\omega_{k,s}^*$ of local specific branches. The global purpose turned to:

$$\min_{\omega_c} \left\{ f(\omega_c) := \frac{1}{N} \sum_{k=1}^{N} h_k(\omega_k^*) \right\} \tag{4}$$

which aims to find the global optimal solution of the invariant branch ω_c. In summary, one-stage optimization is divided into two alternating local and global parts. The advantages are: 1) Finding the optimal global invariant aggregated model with better generalization is easier. 2) The specific branch is optimized locally, which provides personalized adaption. 3) The stability of convergence is enhanced by alternating local-global optimization.

3.2. Framework

Representation disentanglement: For effective attributes disentanglement, the MI-based disentanglement technique (Belghazi et al., 2018) is introduced to disentangle the local extractors. MI has been applied in two aspects: 1) MI maximization between local invariant and global invariant branches enhance the cross-domain similarity. 2) MI minimization between local invariant and local specific branches disentangles the mixed attributes. Thus, the adversarial objective function of the client k model is defined as:

$$L_{MI}^k := I_s(E_s^k(x^k), E_c^k(x^k)) - I_c(E_c^k(x^k), E_c^G(x^k)) \tag{5}$$

Deep InfoMax proposed by (Hjelm et al., 2018) is employed as MI estimation, which is based on Jensen-Shannon estimator, i.e.$I^{JSD}(X,Z) = D_{JS}(\mathbb{P}_{XZ}||\mathbb{P}_X\mathbb{P}_Z)$.

Invariant aggregation is proposed to apply the weighted averaging to the invariant extractors from selected clients as:

$$\mathbb{E}_c^G = \omega_k \mathbb{E}_c^k = \sum_{k=1}^{K} \frac{n_k}{N} \mathbb{E}_c^k \tag{6}$$

At the beginning of optimization, the aggregated model may contain inadvertently mixed local domain-specific attributes. When enough clients participate, one domain-specific attribute is reduced to $\frac{n_k}{N}$ in the aggregated model by model averaging. Thus, for better specific/invariant disentanglement, parameters of the aggregated model are frozen. MI maximization drives the local invariant branch close to global aggregation, which abandons the domain-specific attributes and pays more attention to the cross-domain invariant attributes. After that, the optimal local invariant branch contributes to the global invariant aggregation.

Diversity transferring With the successful disentanglement, the domain-specific attributes are stripped out from the model aggregation. Most pioneering FL methods directly ignored these attributes, which causes substantial waste. The domain-specific attributes contain not only the personalized requirements of local clients but also the diversities of data distributions. Thus, diversity transferring is proposed to augment the local representation sets as:

$$\left\{ R_A^{k,j} \right\} := \left\{ E_s^j(x^k) \bigoplus E_c^k(x^k) | j \in |K| \right\} \tag{7}$$

These representations are extracted by combining the local invariant extractor and cross-domain specific extractors. It forces the local client to pay attention to other domains' specific attributes. These attributes may also exist in the local domain as tail attributes, which are mistakenly ignored by the local extractor. The representation augmentation improves the diversity of local distribution in latent space, mitigating the overfitting and enhancing generalization. Besides, the final loss function of local client k is replaced with:

$$\begin{aligned} h_k &= F_k(\omega) + I_s(\omega) - I_c(\omega) \\ F_k(\omega) &= l_{CE}^k + \frac{\lambda}{K-1} \sum_{j=1, j\neq k}^{K} l_{CE}^k(R_A^{k,j}) \end{aligned} \tag{8}$$

where the l_{CE} is the cross-entropy loss; $F_k(\omega)$ is the loss combination of local and augmented representations; I_s, and I_c are MI estimations. λ are hyperparameters, which is set to 1.0 in the experiments.

Optimization process The entire training consists of local and global two-stage alternating optimization. During each process, the parameters of some model parts are frozen for more targeted training. In the local process, the local invariant parameter is frozen for spinning off the specific attributes into the local specific branch, which is only optimized locally. Then, the optimal local specific and aggregated model parameters are frozen in the global process

Algorithm 1 Optimization process of DFL.

Input: The model initialization $\omega^0 = M(\omega_c^0, \omega_s^0)$; The distributed Non-i.i.d datasets$\{D_k\}_{k=1}^N$; Total optimization round T; Number of participating clients K.
for $i = 0$ **to** $T-1$ **do**
The client subset of t-th round S_t is selected from the N clients.
Send the global aggregation invariant parameter ω_c^t to replace the local invariant branch.
for $k \in S_t$ **do**
The optimization of the local model is divided into two steps.
Step 1: Disentanglement of local specific attributes with freezing the local invariant branch. Client k finds a $\widehat{\omega_k^{t+1}} = M(\omega_c^t, \omega_{k,s}^{t+1,*})$ which is a $\gamma-$inexact minimizer of:
$\omega_{k,s}^{t+1,*} \approx \arg\min_{\omega_{k,s}} h_k(\omega', \omega_c^t)$ where $\omega' = M(\omega_{k,c} = \omega_c^t, \omega_{k,s})$
Step 2: Disentanglement of cross-domain invariant attributes with freezing the local specific branch. Client k finds a $\widetilde{\omega_k^{t+1}} = M(\widetilde{\omega_{k,c}^{t+1}}, \omega_{k,s}^{t+1,*})$ which is a $\gamma-$inexact minimizer of:
$\widetilde{\omega_{k,c}^{t+1}} \approx \arg\min_{\omega_{k,c}} h_k(\omega'', \omega_c^t)$ where $\omega'' = M(\omega_{k,c}, \omega_{k,s} = \omega_{k,s}^{t+1,*})$
end for
Step 3: Invariant aggregation. $\widetilde{\omega_c^{t+1}} = \mathbb{E}_{s_t}[\widetilde{\omega_{k,c}^{t+1}}]$
end for

for better disentanglement of domain-invariant attributes. MI maximization drives the local invariant branch to concentrate on the decision-casual and cross-domain invariant attributes. Afterward, the optimal local invariant parameters are sent to the server for the next round of invariant aggregation. The diversity transfer is employed for representation augmentation in both two training stages. The detailed optimization process is in Algorithm 1, and the parameter updating is shown in Algorithm 2.

3.3. Convergence Analysis

As the optimization process is alternately performed, the training process is changed from a one-stage global optimization to a two-stage partial optimization. In global optimization, only part of the extractor from each client participates in the model aggregation. This work proved the convergence of the system with the following assumptions:

1. Non-convex and L-Lipschitz smoothness of f:

$$\|\bigtriangledown f(\omega) - \bigtriangledown f(\omega')\| \leqslant L\|\omega - \omega'\|, \forall \omega, \omega' \tag{9}$$

Algorithm 2 Parameter updating of DFL.

Step 1: $\omega_{k,s}$ do not participate in model aggregation, which only update locally. $\omega_{k,s}^{t+1} = \omega_{k,s}^t - \eta_s(P_s^T \bigtriangledown_\omega F_k(\omega) + \bigtriangledown_{\omega_s} I_s(\omega_{k,s}^t, \omega_c^t))$
Step 2: The updating of $\omega_{k,c}^{t+1}$ based on global aggregation parameter ω_c^t. $\omega_{k,c}^{t+1} = \omega_c^t - \eta_c(P_c^T \bigtriangledown_\omega F_k(\omega) - \bigtriangledown_{\omega_c} I_c(\omega_{k,c}^t, \omega_c^t))$
Step 3: invariant aggregation $\omega_c^{t+1} = \frac{1}{K}\sum_{k \in s_t} \omega_{k,c}^{t+1}$

2. Polyak-Łojasiewicz of I_c, I_s:

$$\begin{aligned}\left\|\bigtriangledown I_c\left(\omega, \omega_c^t\right) - \bigtriangledown I_c\left(\omega', \omega_c^t\right)\right\| &\geqslant \mu_{I_c}\|\omega - \omega'\|, \forall \omega, \omega' \\ \left\|\bigtriangledown I_s\left(\omega, \omega_c^t\right) - \bigtriangledown I_s\left(\omega', \omega_c^t\right)\right\| &\geqslant \mu_{I_s}\|\omega - \omega'\|, \forall \omega, \omega'\end{aligned} \tag{10}$$

3. $\overline{\mu}-$strongly convex of h_k and Polyak-Łojasiewicz:

$$\begin{aligned}&\left\|\bigtriangledown h_k(M(\omega_c, \omega_{k,s}^{t+1,*}), \omega_c^t) - \bigtriangledown h_k(M(\omega_c', \omega_{k,s}^{t+1,*}), \omega_c^t)\right\| \\ &\geq \overline{\mu}\|\omega_c - \omega_c'\|\end{aligned} \tag{11}$$

4. Bounded second moments of I_c, I_s gradient:

$$\begin{aligned}\mathbb{E}_k\left[\left\|\bigtriangledown I_c\left(\omega, \omega_c^t\right)\right\|^2\right] &\leqslant \epsilon_c^2, \exists \epsilon_c \\ \mathbb{E}_k\left[\left\|\bigtriangledown I_s\left(\omega, \omega_c^t\right)\right\|^2\right] &\leqslant \epsilon_s^2, \exists \epsilon_s\end{aligned} \tag{12}$$

The definitions of the $\gamma-$inexact solution and B-local dissimilarity are the same as FedProx (Li et al., 2018) in the Appendix. First, the expected aggregation model parameter is defined as: $\overline{\omega}_c^{t+1} = \mathbb{E}_k\left[\omega_{k,c}^{t+1}\right]$. Then, we establish the updating relationship between the expected and empirical parameters of the aggregation model, using local-global two-stage optimization. The optimization purpose of the local and global training processes is defined in Algorithm 1.

Definition 3.1. In client k, local specific optimal parameter is $\widehat{\omega_k^{t+1}} = M(\omega_c^t, \omega_{k,s}^{t+1,*})$, and the invariant optimal parameter is $\widetilde{\omega_k^{t+1}} = M(\widetilde{\omega_{k,c}^{t+1}}, \omega_{k,s}^{t+1,*})$. Besides, the empirical parameters applied for the updating of client k is $\omega_k^{t+1} = M(\omega_{k,c}^{t+1}, \omega_{k,s}^{t+1,*})$.

With the bounded gradient of local specific optimization and the empirical updating in the Algorithm 2, The following inequality can be derived:

$$\begin{aligned}\left\|\widetilde{\omega_{k,c}^{t+1}} - \omega_{k,c}^{t+1}\right\| &\leqslant \frac{\gamma}{\overline{\mu} P_c}\left\|\bigtriangledown h_k(\widehat{\omega_k^{t+1}})\right\| \\ \left\|\omega_{k,c}^{t+1} - \omega_c^t\right\| &\leqslant \frac{1+\gamma}{\overline{\mu} P_c}\left\|\bigtriangledown h_k(\widehat{\omega_k^{t+1}})\right\|\end{aligned} \tag{13}$$

Equation (32) verifies that 1) the distance between expected and empirical of the local invariant extractor's parameter; and 2) the local invariant updating are both bounded by the

gradient of the local specific extractor. Thus, the updating relationship between the expected and empirical parameters is derived as:

$$\begin{aligned}&\left\|\overline{\omega}_c^{t+1}-\omega_c^t\right\|^2 \leqslant \mathbb{E}_k\left[\left\|\omega_{k,c}^{t+1}-\omega_c^t\right\|^2\right]\\&\leqslant \frac{(1+\gamma)^2}{\overline{\mu}^2 P_c^2}\left(2B^2\left\|\bigtriangledown f(\omega_c^t)\right\|^2+\left(\frac{8L^2P_s^2}{\mu_{I_s}^2}+2\right)\epsilon_s^2\right)\end{aligned} \tag{14}$$

After that, the local Lipschitz continuity of the function f is applied to approximate ω_c^{t+1}.

$$\begin{aligned}&f\left(\omega_c^{t+1}\right)\leqslant f\left(\overline{\omega}_c^{t+1}\right)+L_0\left\|\omega_c^{t+1}-\overline{\omega}_c^{t+1}\right\|\\&L_0\leqslant\left\|\bigtriangledown f\left(\omega_c^t\right)\right\|+L\left(\left\|\overline{\omega}_c^{t+1}-\omega_c^t\right\|+\left\|\omega_c^{t+1}-\omega_c^t\right\|\right)\end{aligned} \tag{15}$$

After finishing the derivation process, the convergence of DFL is obtained as:

Theorem 3.2. Convergence of disentangled federated learning. *Let Assumptions 1-4 hold. Suppose that ω_c^t is not a stationary solution and the local functions F_K is B-locally dissimilar, i.e. $B(\omega_c^t)\leqslant B$. If the hyperparameters in α,β are chosen such that:*

$$\begin{aligned}\alpha=&\frac{\eta_c}{2}+2\eta_c P_c^2B^2-\frac{LB^2(1+\gamma)^2}{\overline{\mu}^2P_c^2}\\&-\frac{2(1+\gamma)B}{\overline{\mu}P_c\sqrt{K}}-\frac{2(2\sqrt{2K}+2)L(1+\gamma)^2B^2}{K\overline{\mu}^2P_c^2}>0\\\beta=&\frac{\sqrt{K}\overline{\mu}P_c(1+\gamma)+BL(K+4\sqrt{2K}+4)(1+\gamma)^2}{KB\overline{\mu}^2P_c^2}\\&\left(\frac{4L^2P_s^2}{\mu_{I_s}^2}+1\right)-\frac{8\eta_cL^2P_c^2P_s^2}{\mu_{I_s}^2}\end{aligned} \tag{16}$$

then at the t-th round of optimization, the expected decrease in the global objective is:

$$\mathbb{E}_{s_t}[f(\omega_c^{t+1})]\leqslant f(\omega_c^t)-\alpha\left\|\bigtriangledown f(\omega_c^t)\right\|+\beta\epsilon_s^2-\eta_c\epsilon_c^2 \tag{17}$$

which means the convergence is that:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\bigtriangledown f(\omega_c^t)\right\|\leqslant\frac{1}{\alpha T}\left(f(\omega_c^0)-f^*\right)+\beta\epsilon_s^2-\eta_c\epsilon_c^2 \tag{18}$$

where f^* is the minimum value of the problem and the ω_c^0 is the initialization of the model.

The convergence proof is detailed in Appendix. As far as we know, Theorem 3.2 first provides the convergence guarantee of partial aggregation. It means that the DFL system is convergent even if only part of the extractor participates in the aggregation, based on the bounded gradient of the local specific branch. Compared with FedAvg, the convergence rate of DFL can be sped up with better optimization of local specific branches. This process is only trained locally, which reduces the entire time cost with fewer communication rounds. The other benefits of this theory are that: 1) The constraint of the client's model consistency is relaxed, amplifying the FL application. 2) The incomplete model makes the raw data reconstruction more difficult, enhancing privacy security.

4. Experiment

This section consists of three experimental parts: clarification, verification, and application. The first part shows the complex problems caused by attributes skew in FL, which is the independent component of Non-i.i.d and exists ubiquitously. Clarification experiments clarify that attributes skew can cause unstable convergence and performance degradation. Verification experiments focus on manually synthesized attributes skew, which aims to verify the effectiveness, loss convergence, and performance improvement of DFL compared with other critical related works. Application experiments pay attention to the performance of DFL in datasets with realistic attributes skew, which tries to prove that DFL can adapt to the practical environment.

Benchmark datasets: The clarification experiments are performed on the MNIST and **colored-MNIST (Arjovsky et al., 2019)**. The former is attributes balanced, and the latter has skewed attributes as different foreground/background colors in different clients. These digit classification datasets are both divided into 20 clients. The verification experiments are performed on: 1) **colored-MNIST** (Arjovsky et al., 2019) with attributes skew as the background color (BG color in Table 3). 2) **3dshapes** (Burgess & Kim, 2018) is employed the background colors (BG color in Table 3) or scales(Scale in Table 3) as attributes skew. The shape is employed for classification. 3) **dSprites** (Matthey et al., 2017) performs object scale classification, with the attributes skew as orientation (Orientation in Table 3). Besides, the training sampling ratio in verification experiments is reduced to increase the classification difficulty. However, one thing is sure the comparisons are fair for every FL method in each experiment. The application experiments employ two datasets: 1) **Office-Caltech10** (Gong et al., 2012), which contains ten overlapping classes from four domains acquired in different cameras or environments. 2) **DomainNet** (Peng et al., 2019a) has 345 overlapping classes from six domains with different image styles. For both application datasets, the representation distribution across attributes on each domain is different from each other. In addition, both whole training datasets are employed for distributed training.

Backbones: For clarification and verification, the network architecture follows as (McMahan et al., 2017). The representation extraction networks contain three groups of Conv-BN-Relu layers. The classifier is composed of two fully connected layers. For application, the AlexNet (Krizhevsky

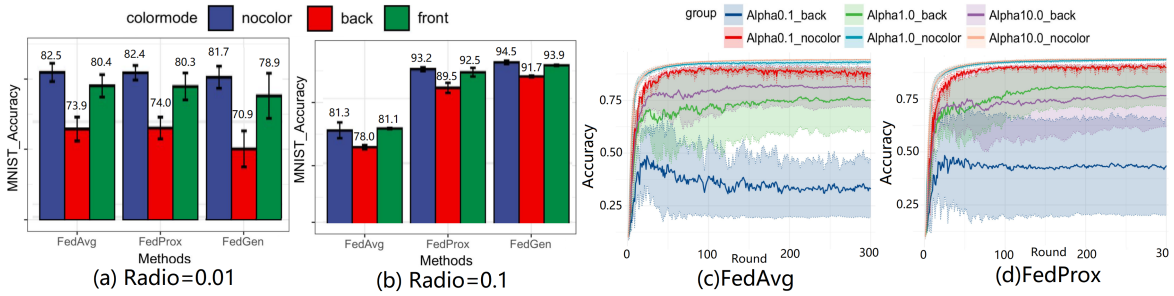

Figure 3. (a) and (b) show the huge accuracy degradation of the MNIST classification with the introduction of attributes skew, in different FL methods and sampling ratios. Similarly, (c) and (d) represents the unstable convergence of the Colored-MNIST.

Table 1. Top-1 test accuracy of verifications on Colored-MNIST, 3Dshapes, dSprites. BG color means that each client has different floor color and wall color.

Dataset	Attributes	clients	FedAvg	FedProx	FedGen	DFL
Colored-MNIST	BG color	10/20	88.88±0.28	89.93±0.87	93.47±0.26	**95.91±0.13**
3Dshapes	BG color	20/50	98.57±0.46	98.16±0.79	98.38±0.47	**99.37±0.09**
3Dshapes	Scale	10/10	89.34±1.25	89.93±1.43	76.57±9.18	**90.38±0.56**
dSprites	Orientation	20/40	73.55±4.78	71.64±5.23	82.69±1.82	**86.74±2.09**

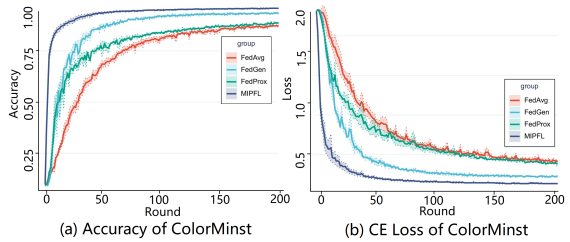

Figure 4. Accuracy and cross-entropy loss curves with global communication rounds increase. DFL improves classification performance and convergence stability compared with other baseline FL methods in Colored-MNIST. DFL reaches better performance with fewer communication rounds.

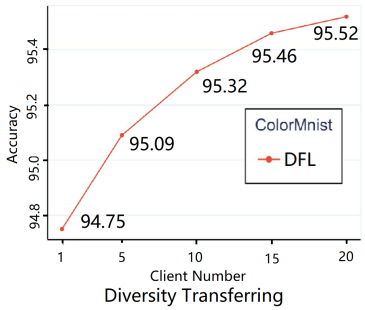

Figure 5. Accuracy curve as client number increases.

Table 2. Ablation study of DFL in Colored-MNIST.

	Invariant Aggregation	Diversity Transferring	DFL
10/20 Ratio=0.5 BG-color	√		95.11±0.13
		√	95.29±0.33
	√	√	**96.02±0.30**

Table 3. Top-1 test accuracy of application on Office-Caltech10. A, C, D , and W are abbreviations for Amazon, Caltech, DSLR and WebCam.

Dataset / Methods	Office-Caltech10, Backbone=AlexNet				
	A	C	D	W	Avg
Fedavg	60.64	54.22	87.50	96.61	74.69
FedProx	59.38	53.33	84.38	94.92	73.00
FedBN	69.27	55.00	96.88	97.31	79.61
DFL	**73.96**	**55.56**	**100.00**	**98.31**	**81.96**

et al., 2012) without pre-training and ResNet101 (He et al., 2016) with pre-training are selected as representation extraction modules. The classifier consists of 3 fully connected layers and two batch normalization layers.

Comparison Five pioneering FL methods are fairly compared. FedAvg (McMahan et al., 2017) is the classic FL using simple aggregation. FedProx (Li et al., 2018) employed regularization terms for optimization correction that achieve good personalized performance. FedGen (Zhu et al., 2021) proposed a data-free knowledge distillation to mitigate the distribution shift. FedEnsemble (Shi et al., 2021) introduced model ensembling to FL using random permutations for updating. FedBN (Li et al., 2021) applied local batch normalization to alleviate the feature shift.

Configurations: Unless otherwise mentioned, the global communication round set 200 for clarification/verification

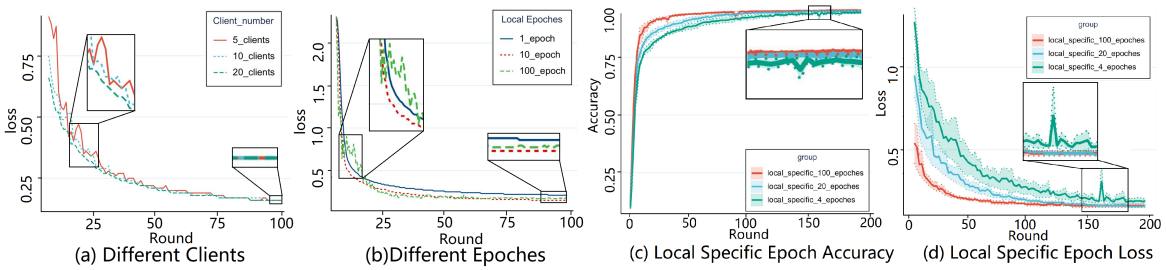

Figure 6. (a), (b) are the loss curves with different client participation and different local updating epochs.(c), (d) are the accuracy and cross-entropy loss curves.

Table 4. Top-1 test accuracy of application on DomainNet.

		Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg
FedAvg	DomainNet	77.70	37.29	62.84	73.00	70.67	72.56	65.68
FedProx	Backbone	77.71	38.96	62.20	72.50	71.08	71.12	65.60
FedBN	=AlexNet	76.43	35.31	65.11	83.60	74.45	74.55	68.24
DFL	Top-10 Classes	**77.76**	**41.55**	**66.88**	**84.10**	**76.42**	**74.65**	**70.23**
FedAvg	DomainNet	96.32	60.12	94.83	82.10	95.81	93.68	87.14
FedProx	Backbone	96.58	60.27	94.67	82.90	95.15	94.04	87.27
FedBN	=ResNet101	**97.15**	61.34	94.80	87.00	96.63	94.95	88.65
DFL	Top-10 Classes	96.20	**61.64**	**95.01**	**89.60**	**96.73**	**95.67**	**89.14**
SingleSet	ResNet101	69.3	34.5	66.3	66.8	80.1	60.7	62.95
DFL	All 345 Classes	**78.4**	**38.2**	**71.2**	**70.4**	**82.7**	**68.6**	**68.25**

and 100 for application. The testing sample ratio of verification experiments is 0.5, which means half of the images are used for testing. The training sample ratios are set to: 1) 0.01 or 0.1 for MNIST and Colored-MNIST in the clarification experiments; 2) 0.5 for Colored-MNIST, 0.01 for 3Dshapes, and 0.2 for dSprites in verification experiments; 3) 1.0 for Office-Caltech10 and DomainNet in the application experiments, which means using all training data for FL training. The data distribution of the training and testing set is the same in each experiment in one client, but different between clients. The local updating step is E=20, and the mini-batch size is B=32. The learning rate lr=0.0.1. The total client numbers are 20 for Colored-MNIST, 50 for 3Dshapes with BG color skew, 10 for 3Dshapes with scale skew, 40 for dSprites, 4 for Office-Caltech10, and 6 for DomainNet.

4.1. Clarification

(a) and (b) of Figure 3 shows huge accuracy degradation with the introduction of attributes skew. With the decrease in sampling ratio, the performance damage rapid growth, which increased from less than 4% at 0.1 ratio to over 10% at 0.01 ratio. In addition, the negative transfer caused by the background color attributes skew is worse than the foreground color skew. To further explore the influence of attributes skew on the FL system, the attributes skew and label skew are mixed, which is closer to reality. The accuracy curves as (c) and (d) in Figure 3 show both the performance and convergence are severely damaged in any FL methods, which even tends to be un-convergent. In summary, the stability, convergence, and performance of the FL system suffer from the introduction of attributes skew. Worse yet, this phenomenon widely exists in reality, which seriously hinders the application of FL. Thus, It is crucial to delve into and mitigate the bad influence of attributes skew.

4.2. Verification

Performance: The verification results are shown in Table 3. DFL outperforms FedAvg, FedProx, and FedGen on all three datasets with different attributes skew and different classification tasks. Especially, FedGen applied global representation generation to complement local latent space. The higher performance of DFL means that the diversity transferring augmented the local representations with a more accurate distribution. The accuracy curve as (a) in Figure 4 indicates that DFL provides a faster and more stable performance improvement.

Convergence: The cross-entropy loss curve as (b) in Figure 4 shows that DFL achieves a higher rate and more stable convergence compared with SOTA FL methods. The small volatility of the DFL loss curve verifies the similar optimization directions provided by invariant aggregation. The accuracy and loss curve of (c) and (d) in Figure 6 repre-

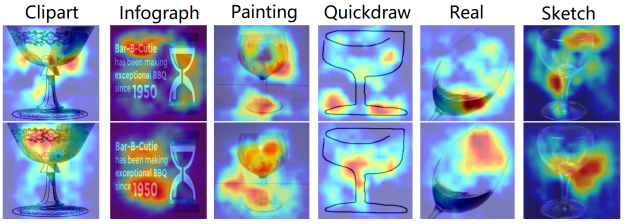

Figure 7. Visualization of DomainNet. The first line shows the Grad-CAM (Selvaraju et al., 2017) generated by invariant branches, and the second line shows specific branches.

sent that the convergence is sped up, and the accuracy is improved with the increase of local specific training epochs. It also means that the DFL can achieve better performance with fewer communication rounds. In addition, the influence of participating clients number increasing is analyzed. The loss curve as (a) in Figure 6 shows that the convergence is more stable with more clients participating, although the final losses are the same. Similarly, ten local updating epochs setting achieves the best and fastest convergence compared with 1 or 100, shown in (b) of Figure 6.

Communication cost: At first, the (c) and (d) of Figure 6 present that the convergence rate is sped up due to better local specific optimization, which is trained independently with more local epochs. It means DFL needs fewer communication rounds. Second, diversity transferring is an optional component. Analyzing the results in Table 1 and the ablation study in Table 2, DFL only with invariant aggregation has fewer communication costs and outperforms other SOTA FL methods. Third, Figure 5 quantitatively presents the improvement of absolute accuracy as client number increases, wherein more clients participating in the diversity transferring means more communication cost. The performance gain tends to stabilize with more participating clients. It means that only a few clients selected are enough for diversity transferring. The communication overhead will not be very large.

Ablation: Table 2 is the ablation study of DFL, which verifies the complementarity of invariant aggregation and diversity transferring. The one-stage optimization is also tested, but the training diverged, and the results are not shown in the body text.

4.3. Application

Application experiments verify that DFL can adapt to realistic environments with realistic attribute shifts. Table 3 and Table 4 show that DFL outperforms other baselines with a considerable margin. In Office-Caltech10, DFL significantly improves all categories and mean accuracy by almost 3%. For DomainNet, regardless of whether the backbone is AlexNet or ResNet101, the mean accuracy of DFL is improved. In addition, the performance of DFL is better than SingleSet training, which is only trained by the source data. It proves that DFL transfers knowledge from other domains successfully. The visualization in Figure 7 is an example of mug classification from DomainNet. The heatmaps of invariant branches are highlighted around the shape of the objects, which is causal to the final decision. The specific branches from different clients pay attention to different attributes such as liquid, description, and color, which are decision-correlated but domain-specific. It demonstrates the effectiveness of disentanglement in DFL. In future work, we would implement DFL on MindSpore[2], which is a new deep learning computing framework suitable for FL applications.

5. Conclusion

In this paper, we propose a novel FL paradigm that applies MI-based disentanglement to mitigate the negative transfer caused by distributed attributes skew. The invariant aggregation is proposed to spin off the mixed domain-specific attributes, which essentially corrects the optimization direction. The proposed diversity transferring augmented the representation in local latent space. The convergence of global aggregation using an incomplete client model expands the application scope of FL. Extensive experiments, guided by our solid convergence analysis, verify that DFL benefits FL with higher performance, better interpretability, and fewer communication rounds.

Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant No.62006225, 61906199, 62071468), the Strategic Priority Research Program of Chinese Academy of Sciences (CAS) (Grant No. XDA27040700), and sponsored by CAAI-Huawei Mindspore Open Fund.

References

Acar, D. A. E., Zhao, Y., Zhu, R., Matas, R., Mattina, M., Whatmough, P., and Saligrama, V. Debiasing model updates for improving personalized federated training. In *International Conference on Machine Learning*, pp. 21–31. PMLR, 2021.

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information

[2]https://www.mindspore.cn/

neural estimation. In *International Conference on Machine Learning*, pp. 531–540. PMLR, 2018.

Burgess, C. and Kim, H. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Cakir, F., He, K., Adel Bargal, S., and Sclaroff, S. Mihash: Online hashing with mutual information. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 437–445, 2017.

Cakir, F., He, K., Bargal, S. A., and Sclaroff, S. Hashing with mutual information. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2424–2437, 2019.

Chen, F., Luo, M., Dong, Z., Li, Z., and He, X. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*, 2018.

Dinh, C. T., Tran, N. H., and Nguyen, T. D. Personalized federated learning with moreau envelopes. *arXiv preprint arXiv:2006.08848*, 2020.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33, 2020.

Feng, H.-Z., You, Z., Chen, M., Zhang, T., Zhu, M., Wu, F., Wu, C., and Chen, W. Kd3a: Unsupervised multi-source decentralized domain adaptation via knowledge distillation. *arXiv preprint arXiv:2011.09757*, 2020.

Gong, B., Shi, Y., Sha, F., and Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2066–2073. IEEE, 2012.

Hanzely, F. and Richtárik, P. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Hong, J., Zhu, Z., Yu, S., Wang, Z., Dodge, H. H., and Zhou, J. Federated adversarial debiasing for fair and transferable representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 617–627, 2021.

Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., and Zhang, Y. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7865–7873, 2021.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Li, X., Jiang, M., Zhang, X., Kamp, M., and Dou, Q. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

Liang, P. P., Liu, T., Ziyin, L., Allen, N. B., Auerbach, R. P., Brent, D., Salakhutdinov, R., and Morency, L.-P. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019a.

Peng, X., Huang, Z., Sun, X., and Saenko, K. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning*, pp. 5102–5112. PMLR, 2019b.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shi, N., Lai, F., Kontar, R. A., and Chowdhury, M. Fed-ensemble: Improving generalization through model ensembling in federated learning. *arXiv preprint arXiv:2107.10663*, 2021.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Wang, K., Mathews, R., Kiddon, C., Eichner, H., Beaufays, F., and Ramage, D. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019a.

Wang, Z., Dai, Z., Póczos, B., and Carbonell, J. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11293–11302, 2019b.

Zhang, M., Sapra, K., Fidler, S., Yeung, S., and Alvarez, J. M. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020.

Zhu, Z., Hong, J., and Zhou, J. Data-free knowledge distillation for heterogeneous federated learning. *arXiv preprint arXiv:2105.10056*, 2021.

A. Appendix A

A.1. Proof of Convergence

Theorem 3.2. Convergence of disentangled federated learning. Let Assumptions 1-4 hold. Suppose that ω_c^t is not a stationary solution and the local functions F_K is B-locally dissimilar, i.e. $B(\omega_c^t) \leqslant B$. If the hyperparameters in α, β are chosen such that:

$$\begin{aligned}\alpha &= \frac{\eta_c}{2} + 2\eta_c P_c^2 B^2 - \frac{LB^2(1+\gamma)^2}{\overline{\mu}^2 P_c^2} - \frac{2(1+\gamma)B}{\overline{\mu} P_c \sqrt{K}} - \frac{2(2\sqrt{2K}+2)L(1+\gamma)^2 B^2}{K\overline{\mu}^2 P_c^2} > 0 \\ \beta &= \frac{\sqrt{K}\overline{\mu} P_c(1+\gamma) + BL(K+4\sqrt{2K}+4)(1+\gamma)^2}{KB\overline{\mu}^2 P_c^2}\left(\frac{4L^2 P_s^2}{\mu_{I_s}^2}+1\right) - \frac{8\eta_c L^2 P_c^2 P_s^2}{\mu_{I_s}^2}\end{aligned} \tag{19}$$

then at the t-th round of optimization, the expected decrease in the global objective is:

$$\mathbb{E}_{s_t}[f(\omega_c^{t+1})] \leqslant f(\omega_c^t) - \alpha \left\| \bigtriangledown f(\omega_c^t) \right\| + \beta \epsilon_s^2 - \eta_c \epsilon_c^2 \tag{20}$$

which means the convergence is that:

$$\frac{1}{T}\sum_{t=0}^{T-1} \left\| \bigtriangledown f(\omega_c^t) \right\| \leqslant \frac{1}{\alpha T}\left(f(\omega_c^0) - f^*\right) + \beta \epsilon_s^2 - \eta_c \epsilon_c^2 \tag{21}$$

where f^* is the minimum value of the problem, and the ω_c^0 is the initialization of the model.

Proof.

Definition A.1. $\gamma-$**inexact solution:** For local function: $h_k(M(\omega_c, \omega_s), \omega_c^t) = F_k(M(\omega_c, \omega_s)) + I_s(\omega_s, \omega_c^t) - I_c(\omega_c, \omega_c^t)$, and $\gamma \in [0,1]$, $\omega_{k,c}^*$ is defined as $\gamma-$inexact solution of $\underset{\omega_{k,c}}{min}\left\{h_k(M(\omega_{k,c}, \omega_{k,s}^{t+1,*}), \omega_c^t)\right\}$ if:

$$\left\| \bigtriangledown h_k(M(\omega_{k,c}^*, \omega_{k,s}^{t+1,*}), \omega_c^t) \right\| \leqslant \gamma \left\| \bigtriangledown h_k(M(\omega_c^t, \omega_{k,s}^{t+1,*}), \omega_c^t) \right\| \tag{22}$$

Definition A.2. B-local dissimilarity: The local functions F_k is B-locally dissimilar at ω_c^t if:

$$\mathbb{E}_k[\left\| \bigtriangledown F_k(M(\omega_c^t, \omega_{k,s}^*)) \right\|^2] \leqslant \left\| \bigtriangledown f(\omega_c^t) \right\|^2 B^2 \tag{23}$$

and the B is defined as:

$$B(\omega_c^t) = \sqrt{\frac{\mathbb{E}_k[\left\| \bigtriangledown F_k(M(\omega_c^t, \omega_{k,s}^*)) \right\|^2]}{\left\| \bigtriangledown f(\omega_c^t) \right\|^2}} \tag{24}$$

Definition A.3. In global, the expected aggregation model parameters is $\overline{\omega}_c^{t+1} = \mathbb{E}_k\left[\omega_{k,c}^{t+1}\right]$. In client k, local specific optimal parameter is $\widehat{\omega_k^{t+1}} = M(\omega_c^t, \omega_{k,s}^{t+1,*})$, and the invariant optimal parameter is $\widetilde{\omega_k^{t+1}} = M(\widetilde{\omega_{k,c}^{t+1}}, \omega_{k,s}^{t+1,*})$. Besides, the empirical parameters applied for the updating of client k is $\omega_k^{t+1} = M(\omega_{k,c}^{t+1}, \omega_{k,s}^{t+1,*})$.

Let Assumptions 1-4 hold, we have:

$$\left\| \overline{\omega}_c^{t+1} - \omega_c^t \right\| \leqslant \mathbb{E}_k\left[\left\| \omega_{k,c}^{t+1} - \omega_c^t \right\|\right] \tag{25}$$

$$\left\| \widehat{\omega_k^{t+1}} - \widetilde{\omega_k^{t+1}} \right\| = P_c \left\| \widetilde{\omega_{k,c}^{t+1}} - \omega_c^t \right\| \tag{26}$$

$$\left\| \widehat{\omega_k^{t+1}} - \widetilde{\omega_k^{t+1}} \right\| \leqslant \frac{1}{\overline{\mu}} \left\| \bigtriangledown h_k(\widehat{\omega_k^{t+1}}) - \bigtriangledown h_k(\widetilde{\omega_k^{t+1}}) \right\| = \frac{1}{\overline{\mu}} \left\| \bigtriangledown h_k(\widehat{\omega_k^{t+1}}) \right\| \tag{27}$$

Thus,

$$\left\|\widetilde{\omega_{k,c}^{t+1}}-\omega_c^t\right\| \leqslant \frac{1}{\overline{\mu}P_c}\left\|\bigtriangledown h_k(\widehat{\omega_k^{t+1}})\right\| \tag{28}$$

Similar,

$$\left\|\widetilde{\omega_k^{t+1}}-\omega_k^{t+1}\right\| = P_c\left\|\widetilde{\omega_{k,c}^{t+1}}-\omega_{k,c}^{t+1}\right\| \leqslant \frac{1}{\overline{\mu}}\left\|\bigtriangledown h_k(\omega_k^{t+1})\right\| \leqslant \frac{\gamma}{\overline{\mu}}\left\|\bigtriangledown h_k(\widehat{\omega_k^{t+1}})\right\| \tag{29}$$

$$\left\|\widetilde{\omega_{k,c}^{t+1}}-\omega_{k,c}^{t+1}\right\| \leqslant \frac{\gamma}{\overline{\mu}P_c}\left\|\bigtriangledown h_k(\widehat{\omega_k^{t+1}})\right\| \tag{30}$$

Trigonometric inequality is introduced:

$$\left\|\omega_{k,c}^{t+1}-\omega_c^t\right\| \leqslant \left\|\widetilde{\omega_{k,c}^{t+1}}-\omega_c^t\right\| + \left\|\widetilde{\omega_{k,c}^{t+1}}-\omega_{k,c}^{t+1}\right\| \tag{31}$$

Thus,

$$\left\|\omega_{k,c}^{t+1}-\omega_c^t\right\| \leqslant \frac{1+\gamma}{\overline{\mu}P_c}\left\|\bigtriangledown h_k(\widehat{\omega_k^{t+1}})\right\| \tag{32}$$

$$\left\|\overline{\omega}_c^{t+1}-\omega_c^t\right\|^2 \leqslant \mathbb{E}_k\left[\left\|\omega_{k,c}^{t+1}-\omega_c^t\right\|^2\right] \leqslant \frac{(1+\gamma)^2}{\overline{\mu}^2P_c^2}\mathbb{E}_k\left[\left\|\bigtriangledown h_k(\widehat{\omega_k^{t+1}})\right\|^2\right] \tag{33}$$

Next, the expectation calculation is introduced:

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|\bigtriangledown h_k(\widehat{\omega_k^{t+1}})\right\|^2\right] = \mathbb{E}_k\left[\left\|\bigtriangledown F_k\left(M(\omega_c^t,\omega_{k,s}^{t+1,*})\right)+\bigtriangledown I_s\left(\omega_{k,s}^{t+1,*},\omega_c^t\right)\right\|^2\right] \\
&\leqslant 2\mathbb{E}_k\left[\left\|\bigtriangledown F_k\left(M(\omega_c^t,\omega_{k,s}^{t+1,*})\right)\right\|^2\right] + 2\mathbb{E}_k\left[\left\|\bigtriangledown I_s\left(\omega_{k,s}^{t+1,*},\omega_c^t\right)\right\|^2\right] \\
&\leqslant 2\mathbb{E}_k\left[\left\|\bigtriangledown F_k\left(M(\omega_c^t,\omega_{k,s}^{t+1,*})\right)\right\|^2\right] + 2\epsilon_s^2
\end{aligned} \tag{34}$$

$$\begin{aligned}
\left\|\bigtriangledown F_k\left(M(\omega_c^t,\omega_{k,s}^{t+1,*})\right)\right\|^2 &\leqslant 2\left\|\bigtriangledown F_k\left(M(\omega_c^t,\omega_{k,s}^{*})\right)\right\|^2 + 2\left\|\bigtriangledown F_k\left(M(\omega_c^t,\omega_{k,s}^{t+1,*})\right)-\bigtriangledown F_k\left(M(\omega_c^t,\omega_{k,s}^{*})\right)\right\|^2 \\
&\leqslant 2\left\|\bigtriangledown F_k\left(M(\omega_c^t,\omega_{k,s}^{*})\right)\right\|^2 + 2L^2P_s^2\left\|\omega_{k,s}^{t+1,*}-\omega_{k,s}^{*}\right\|^2
\end{aligned} \tag{35}$$

$$\left\|\omega_{k,s}^{t+1,*}-\omega_{k,s}^{*}\right\|^2 \leqslant \frac{1}{\mu_{I_s}^2}\left\|\bigtriangledown I_s\left(\omega_{k,s}^{t+1,*},\omega_c^t\right)-\bigtriangledown I_s\left(\omega_{k,s}^{*},\omega_c^t\right)\right\|^2 \leqslant \frac{4}{\mu_{I_s}^2}\left\|\bigtriangledown I_s\left(\omega_{k,s}^{*},\omega_c^t\right)\right\|^2 \leqslant \frac{4}{\mu_{I_s}^2}\epsilon_s^2 \tag{36}$$

$$\mathbb{E}_k\left[\left\|\bigtriangledown F_k\left(M(\omega_c^t,\omega_{k,s}^{*})\right)\right\|^2\right] \leqslant B^2\left\|\bigtriangledown f(\omega_c^t)\right\|^2 \tag{37}$$

$$\mathbb{E}_k\left[\left\|\bigtriangledown h_k(\widehat{\omega_k^{t+1}})\right\|^2\right] \leqslant 2B^2\left\|\bigtriangledown f(\omega_c^t)\right\|^2 + \left(\frac{8L^2P_s^2}{\mu_{I_s}^2}+2\right)\epsilon_s^2 \tag{38}$$

$$\left\|\overline{\omega}_c^{t+1}-\omega_c^t\right\|^2 \leqslant \mathbb{E}_k\left[\left\|\omega_{k,c}^{t+1}-\omega_c^t\right\|^2\right] \leqslant \frac{(1+\gamma)^2}{\overline{\mu}^2P_c^2}\left(2B^2\left\|\bigtriangledown f(\omega_c^t)\right\|^2 + \left(\frac{8L^2P_s^2}{\mu_{I_s}^2}+2\right)\epsilon_s^2\right) \tag{39}$$

Based on the L-Lipschitz smoothness of f and Taylor expansion, it is:

$$f\left(\overline{\omega}_c^{t+1}\right) \leqslant f\left(\omega_c^t\right) + \left\langle\bigtriangledown f\left(\omega_c^t\right),\overline{\omega}_c^{t+1}-\omega_c^t\right\rangle + \frac{L}{2}\left\|\overline{\omega}_c^{t+1}-\omega_c^t\right\|^2 \tag{40}$$

$$\mathbb{E}_{s_t}\left[\left\langle \bigtriangledown f\left(\omega_c^t\right), \overline{\omega}_c^{t+1}-\omega_c^t\right\rangle\right] = \mathbb{E}_{s_t}\left[\left\langle \bigtriangledown f\left(\omega_c^t\right), \mathbb{E}_k[\omega_{k,c}^{t+1}-\omega_c^t]\right\rangle\right] = -\eta_c \mathbb{E}_{s_t}\left[\left\langle \bigtriangledown f\left(\omega_c^t\right), \mathbb{E}_k[\bigtriangledown_{\omega_c} h_k(\widehat{\omega_k^{t+1}})]\right\rangle\right] \quad (41)$$

Since the following inequality established as:

$$\langle a, b\rangle = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a-b\|^2\right) \quad (42)$$

similarly, we have:

$$\begin{aligned}\mathbb{E}_{s_t}\left[\left\langle \bigtriangledown f\left(\omega_c^t\right), \overline{\omega}_c^{t+1}-\omega_c^t\right\rangle\right] &\leqslant -\frac{\eta_c}{2}\mathbb{E}_{s_t}(\left\|\bigtriangledown f\left(\omega_c^t\right)\right\|^2 + \left\|\bigtriangledown_{\omega_c} h_k(\widehat{\omega_k^{t+1}})\right\|^2 - \left\|\bigtriangledown f\left(\omega_c^t\right) - \mathbb{E}_k[\bigtriangledown_{\omega_c} h_k(\widehat{\omega_k^{t+1}})]\right\|^2) \\ &\leqslant -\frac{\eta_c}{2}(\left\|\bigtriangledown f\left(\omega_c^t\right)\right\|^2 + \mathbb{E}_{s_t}\left[\left\|\bigtriangledown_{\omega_c} h_k(\widehat{\omega_k^{t+1}})\right\|^2\right])\end{aligned} \quad (43)$$

$$\begin{aligned}\mathbb{E}_{s_t}\left[\left\|\bigtriangledown_{\omega_c} h_k(\widehat{\omega_k^{t+1}})\right\|^2\right] &= \mathbb{E}_{s_t}\left[\left\|P_c \bigtriangledown F_k\left(\widehat{\omega_k^{t+1}}\right) - \bigtriangledown I_c\left(\omega_{k,c}^t, \omega_c^t,\right)\right\|^2\right] \leqslant 2P_c^2 \mathbb{E}_{s_t}\left[\left\|\bigtriangledown F_k\left(\widehat{\omega_k^{t+1}}\right)\right\|^2\right] + 2\epsilon_c^2 \\ &\leqslant 4P_c^2 B^2\left\|\bigtriangledown f(\omega_c^t)\right\|^2 + \frac{16L^2P_c^2P_s^2}{\mu_{I_s}^2}\epsilon_s^2 + 2\epsilon_c^2\end{aligned} \quad (44)$$

The local Lipschitz continuity of the function f is applied to approximate ω_c^{t+1}.

$$\begin{aligned}f\left(\omega_c^{t+1}\right) &\leqslant f\left(\overline{\omega}_c^{t+1}\right) + L_0\left\|\omega_c^{t+1} - \overline{\omega}_c^{t+1}\right\| \\ L_0 &\leqslant \left\|\bigtriangledown f\left(\omega_c^t\right)\right\| + L\left(\left\|\overline{\omega}_c^{t+1} - \omega_c^t\right\| + \left\|\omega_c^{t+1} - \omega_c^t\right\|\right)\end{aligned} \quad (45)$$

Following the derivation process, we have:

$$\begin{aligned}\mathbb{E}_{s_t}\left[f(\omega_c^{t+1})\right] &\leqslant f(\overline{\omega}_c^{t+1}) + \sqrt{\frac{2}{K}\mathbb{E}_k\left[\left\|\omega_{k,c}^{t+1} - \omega_c^t\right\|^2\right]}\left\|\bigtriangledown f(\omega_c^t)\right\| + \frac{(2\sqrt{2K}+2)L}{K}\mathbb{E}_k\left[\left\|\omega_{k,c}^{t+1} - \omega_c^t\right\|^2\right] \\ &\leqslant f(\overline{\omega}_c^{t+1}) + \frac{2(1+\gamma)B}{\overline{\mu}P_c\sqrt{K}}\sqrt{\left\|\bigtriangledown f(\omega_c^t)\right\|^4 + \left(\frac{4L^2P_s^2}{\mu_{I_s}^2}+1\right)\frac{\epsilon_s^2}{B^2}\left\|\bigtriangledown f(\omega_c^t)\right\|^2} \\ &+ \frac{2(2\sqrt{2K}+2)L(1+\gamma)^2}{K\overline{\mu}^2P_c^2}\left(B^2\left\|\bigtriangledown f(\omega_c^t)\right\|^2 + \left(\frac{4L^2P_s^2}{\mu_{I_s}^2}+1\right)\epsilon_s^2\right) \\ &\leqslant f(\overline{\omega}_c^{t+1}) + \left(\frac{2(1+\gamma)B}{\overline{\mu}P_c\sqrt{K}} + \frac{2(2\sqrt{2K}+2)L(1+\gamma)^2B^2}{K\overline{\mu}^2P_c^2}\right)\left\|\bigtriangledown f(\omega_c^t)\right\|^2 \\ &+ \frac{\sqrt{K}\overline{\mu}P_c(1+\gamma) + 2BL(2\sqrt{2K}+2)(1+\gamma)^2}{KB\overline{\mu}^2P_c^2}\left(\frac{4L^2P_s^2}{\mu_{I_s}^2}+1\right)\epsilon_s^2\end{aligned} \quad (46)$$

$$\begin{aligned}f(\overline{\omega}_c^{t+1}) &\leqslant f(\omega_c^t) - \left(\frac{\eta_c}{2} + 2\eta_c P_c^2 B^2 - \frac{LB^2(1+\gamma)^2}{\overline{\mu}^2P_c^2}\right)\left\|\bigtriangledown f(\omega_c^t)\right\|^2 \\ &+ \left(\frac{L(1+\gamma)^2}{\overline{\mu}^2P_c^2}\left(\frac{4L^2P_s^2}{\mu_{I_s}^2}+1\right) - \frac{8\eta_c L^2P_c^2P_s^2}{\mu_{I_s}^2}\right)\epsilon_s^2 - \eta_c\epsilon_c^2\end{aligned} \quad (47)$$

At last, we get the convergence as:

$$\begin{aligned}&\mathbb{E}_{s_t}\left[f(\omega_c^{t+1})\right] \leqslant f(\omega_c^t) - (\frac{\eta_c}{2} + 2\eta_c P_c^2 B^2 - \frac{LB^2(1+\gamma)^2}{\overline{\mu}^2P_c^2} - \frac{2(1+\gamma)B}{\overline{\mu}P_c\sqrt{K}} - \frac{2(2\sqrt{2K}+2)L(1+\gamma)^2B^2}{K\overline{\mu}^2P_c^2})\left\|\bigtriangledown f(\omega_c^t)\right\|^2 \\ &+ (\frac{\sqrt{K}\overline{\mu}P_c(1+\gamma) + BL(K+4\sqrt{2K}+4)(1+\gamma)^2}{KB\overline{\mu}^2P_c^2}\left(\frac{4L^2P_s^2}{\mu_{I_s}^2}+1\right) - \frac{8\eta_c L^2P_c^2P_s^2}{\mu_{I_s}^2})\epsilon_s^2 - \eta_c\epsilon_c^2\end{aligned} \quad (48)$$

□

A.2. Cross-entropy Loss Functions

With the optimization purpose changing, and the introduction of invariant aggregation and diversity transferring, the loss function of local clients also has huge changes. The optimization purpose of local client is to minimize the loss of each client $k \in |K|$.

$$\arg\min_{\theta_k} \mathbb{E}_{(x_k,y_k)\in D_k^*} \left[\mathfrak{h}_k\left(x_k, y_k, E\right)\right] \tag{49}$$

where D_k^* is the data distribution of client k, $D_i^* \neq D_j^*, 1 \leqslant i \neq j \leqslant K$ shows the Non-i.i.d factors between domains. In reality, the empirical loss function is employed for instead in actual calculations as:

$$\hat{h}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \mathfrak{h}_k\left(x_k, y_k, \omega_k\right) \tag{50}$$

The training samples of k client are defined as $D_k = \left\{\left(x_k^j, y_k^j\right), x_k^j \in \mathbb{R}^d, y_k^j \in \mathbb{R}\right\}_{j=1}^{n_k}$, n_k is the training sample number of client k. With the application of the two-branch local model replacing single-branch, the local empirical loss function tends to:

$$\widehat{h}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \mathfrak{h}_k \left(P^k \left(E_c^k(x_k^j) \bigoplus E_s^k(x_k^j)\right), y_k^j\right) \tag{51}$$