Feature Distribution Matching for Federated Domain Generalization

Yuwei Sun[1,2] **and Ng Chong**[3] **and Hideya Ochiai**[1]

[1]**The University of Tokyo,** [2]**RIKEN AIP,** [3]**United Nations University**

Asian Conference on Machine Learning

Abstract

We propose a new federated domain generalization method called Federated Knowledge Alignment (FedKA). FedKA leverages feature distribution matching in a global workspace such that the global model can learn domain-invariant client features under the constraint of unknown client data. The results show that FedKA can significantly reduce negative transfer, improving the performance gain via model aggregation.

Introduction

- One of the most challenging problems in Federated Learning (FL) is to improve the model generality in tackling client data that show particular sample features from different domains.
- The learned knowledge from a client might not facilitate the learning of others. Simply aggregating these models will not guarantee a better global model.
- The difficulty in federated domain generalization is that client data are not available for domain transfer, hindering effective knowledge sharing in FL.

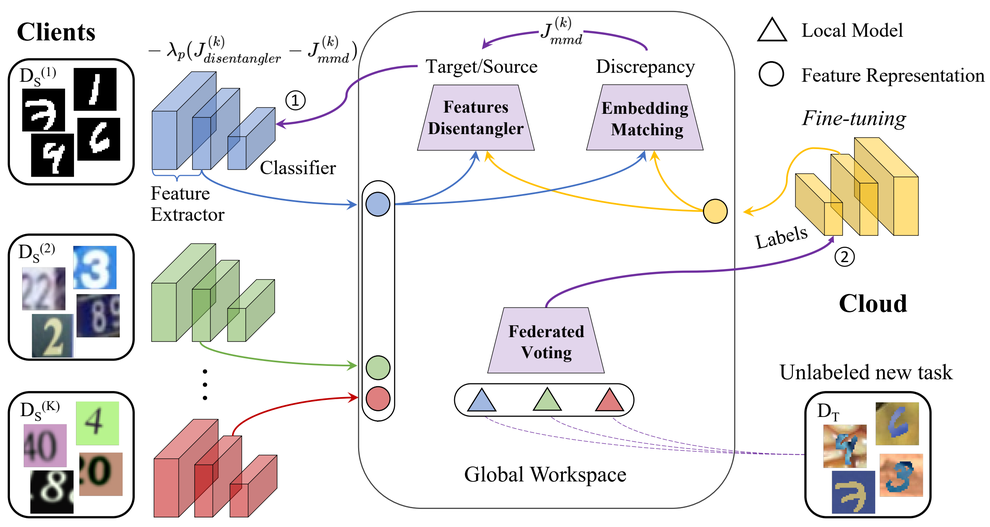


Figure 1: Federated Knowledge Alignment (FedKA) leverages distributed data domains based on feature distribution matching in a global workspace. The negative transfer is alleviated by 1) local model representation learning on client domains and 2) global model fine-tuning on the unlabeled cloud domain.

Methodology

We proposed Federated Knowledge Alignment (FedKA) that reduces feature discrepancy between clients improving the global model's generality to unseen tasks using three building blocks, i.e., Global Feature Disentangler, Embedding Matching, and Federated Voting.

Global Feature Disentangler

- To learn an encoder $f_e^{(k)}$ that disentangles the domain-invariant features H from $X_S^{(k)}$, we devise the global features disentangler by introducing a domain classifier f_d in the server that takes the representations H as the input and outputs a binary variable q for each input sample h, which indicates whether h comes from the client k ($h \in H^{(k)} = f_e^{(k)}(X_S^{(k)})$ if $q = 0$) or from the target domain in the PS ($h \in H^G = f_e^G(X_T)$ if $q = 1$).
- When the features disentangler cannot distinguish whether an input representation is from the client domain or the cloud domain, $f_e^{(k)}$ outputs feature vectors that are close to the ones from the target domain. $\hat{f_e^{(k)}} = \arg\max_{f_e^{(k)}} J^{(k)}_{\text{disentangler}}(\hat{f}_d, f_e^{(k)}, \hat{f_e^G})$.

Embedding Matching

We further enhance the disentanglement of features by measuring the high-dimensional distribution difference between feature representations from a client and the target domain.

- We employ the MK-MMD loss to perform embedding matching between the learned representations of a client $H_S^{(k)} = f_e^{(k)}(X_S^{(k)})$ and the target domain $H_T = f_e^G(X_T)$.
- The local model of client k can be updated based on the embedding matching loss $J^{(k)}_{\text{mmd}}$ as follows

$$J^{(k)}_{\text{mmd}} = \frac{1}{5}\sum_{r=1}^{5} \text{MMD}^2_{e_r}(f_e^{(k)}(X_S^{(k)}), f_e^G(X_T)),\ \hat{f_e^{(k)}} = \arg\min_{f_e^{(k)}} J^{(k)}_{\text{mmd}}(f_e^{(k)}, \hat{f_e^G}).$$

Global Model Fine-Tuning Based on Federated Voting

- Federated voting fine-tunes the global model based on the pseudo-labels generated by the consensus from learned client local models.
- Given an unlabeled input sample x_i from the target domain D_T, the federated voting method aims to attain the optimized classification label y_i^* based on the plurality voting of local models. $y_i^* = \arg\max_{c\in\{1,2,\ldots,C\}} \sum_{k=1}^{K} \mathbb{1}\{y_i^{(k)} = c\}$.
- Then, we fine-tune the global model based on samples from the target domain and the generated labels Y^*.

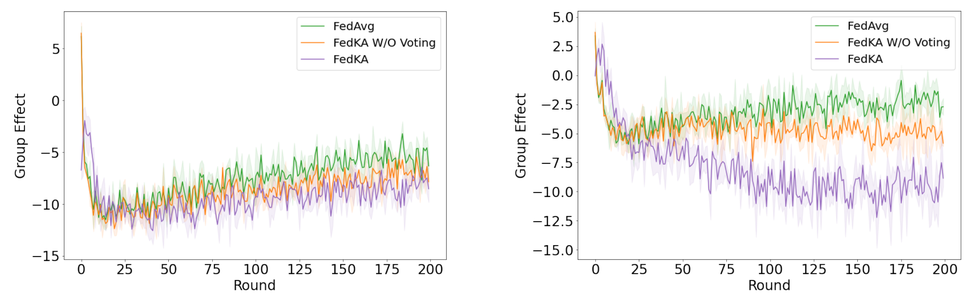
Results

Table 1: Test accuracy on the Digit-Five dataset for different transfer learning tasks.

Models/Tasks	→mt	→mm	→up	→sv	→sy	Avg
FedAvg	93.5±0.15	62.5±0.72	90.2±0.37	12.6±0.31	40.9±0.50	59.9
f-DANN	89.7±0.23	70.4±0.69	88.0±0.23	11.9±0.50	43.8±1.04	60.8
f-DAN	93.5±0.26	62.1±0.45	90.2±0.13	12.1±0.56	41.5±0.76	59.9
Vote-S	**93.7**±0.18	63.4±0.28	**92.6**±0.25	14.2±0.99	45.3±0.34	61.8
Vote-L	93.5±0.18	64.8±1.01	92.3±0.21	14.3±0.42	45.6±0.57	62.1
Disentangler + Vote-S	91.8±0.20	71.2±0.40	91.0±0.58	14.4±1.09	48.7±1.19	63.4
Disentangler + Vote-L	92.1±0.16	71.8±0.48	90.9±0.36	15.1±0.91	49.1±1.03	63.8
Disentangler + MMD	90.0±0.49	70.4±0.86	87.5±0.25	12.2±0.70	44.3±1.18	60.9
FedKA-S	91.8±0.19	72.5±0.91	90.6±0.14	**15.2**±0.46	48.9±0.48	63.8
FedKA-L	92.0±0.26	**72.6**±1.03	91.1±0.24	14.8±0.41	**49.2**±0.78	**63.9**

Group Effect (GE) throws light on negative transfer in the model aggregation of FL, which has not yet been studied to our best knowledge. We formulate GE by $\text{GE}_t = \frac{1}{K}\sum_{k\in\{1,2,\ldots,K\}} \text{TTA}_f(G_t + \Delta_t^{(k)}) - \text{TTA}_f(G_{t+1})$.

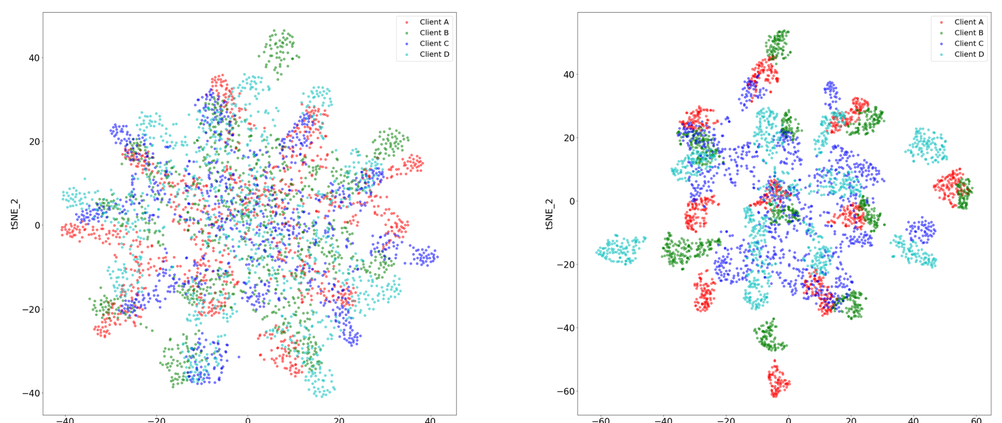
- A higher GE value reflects more information loss from the aggregation and a negative value represents a performance gain via the aggregation.
- The learning progress had high GE values at early stages, implicating that the model aggregation results in information loss. As learning progresses, the GE values keep decreasing showing the gradual convergence of client models toward the target domain distribution.



(a) mt,sv,sy,up→mm (b) mt,mm,sv,up→sy

Figure 2: Group Effect GE_t during the 200 rounds of FL. Lower is better.

T-SNE was employed to visualize the feature distributions of different client domains. The global model based on FedKA learns better representations.



(a) FedAvg (Resnet18) (b) FedKA (Resnet18)

Figure 3: T-SNE visualization of different client domain feature distributions.

Conclusion

The data discrepancy between clients hinders the effectiveness of Federated Learning (FL). Traditional domain adaptation methods cannot benefit FL under the constraint of data confidentiality. We proposed FedKA to allow domain feature matching in the shared global workspace, improving the transferability of local knowledge. The experiments showed that FedKA improved the global model's generality to unseen image and text classification tasks.

arxiv.org/abs/2203.11635