Supplementary File for: FedSR: A Simple and Effective Domain Generalization Method for Federated Learning

Anonymous Author(s)
Affiliation
Address
email

A Proofs

A.1 Lemma 1

Proof. Recall that the joint distribution of x, y, z for client i is:

$$p_i(x, y, z) = p_i(x, y)p(z|x) = p_i(y)p_i(x|y)p(z|x) \tag{1}$$

We have:

$$\begin{aligned}
I_i(x, z|y) &= \mathbb{E}_{p_i(x,y,z)}\left[\log \frac{p_i(x, z|y)}{p_i(x|y)p_i(z|y)}\right] && (2)\\
&= \mathbb{E}_{p_i(x,y,z)}\left[\log \frac{p_i(x|y)p(z|x)}{p_i(x|y)p_i(z|y)}\right] && (3)\\
&= \mathbb{E}_{p_i(x,y,z)}\left[\log \frac{p(z|x)}{p_i(z|y)}\right] && (4)\\
&= \mathbb{E}_{p_i(x,y,z)}\left[\log \frac{p(z|x)}{p_i(z|y)}\right] && (5)\\
&= \mathbb{E}_{p_i(x,y,z)}\left[\log p(z|x) - \log p_i(z|y)\right] && (6)
\end{aligned}$$

Notice that:

$$\begin{aligned}
&\mathbb{E}_{p_i(y)}\left[\mathrm{KL}[p_i(z|y)|r(z|y)]\right] \geq 0 && (7)\\
\Rightarrow &\mathbb{E}_{p_i(y)}\left[\mathbb{E}_{p_i(z|y)}[\log p_i(z|y) - \log r(z|y)]\right] \geq 0 && (8)\\
\Rightarrow &\mathbb{E}_{p_i(y,z)}\left[\log p_i(z|y) - \log r(z|y)\right] \geq 0 && (9)\\
\Rightarrow &\mathbb{E}_{p_i(x,y,z)}\left[\log p_i(z|y) - \log r(z|y)\right] \geq 0 && (10)\\
\Rightarrow &\mathbb{E}_{p_i(x,y,z)}\left[\log p_i(z|y)\right] \geq \mathbb{E}_{p_i(x,y,z)}\left[\log r(z|y)\right] && (11)\\
& && (12)
\end{aligned}$$

Therefore:

$$\begin{aligned}
I_i(x, z|y) &= \mathbb{E}_{p_i(x,y,z)}\left[\log p(z|x) - \log p_i(z|y)\right] && (13)\\
&\leq \mathbb{E}_{p_i(x,y,z)}\left[\log p(z|x) - \log r(z|y)\right] && (14)\\
&= \mathbb{E}_{p_i(x,y)}\left[\mathbb{E}_{p(z|x)}[\log p(z|x) - \log r(z|y)]\right] && (15)\\
&= \mathbb{E}_{p_i(x,y)}\left[\mathrm{KL}[p(z|x)|r(z|y)]\right] && (16)
\end{aligned}$$

□

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

A.2 Lemma 2

Proof. We need to prove:

$$\mathbb{E}_{p_i(x,y)}\left[\mathrm{KL}[p(z|x)|r(z|y)]\right] \geq \mathbb{E}_{p_i(y)}\left[\mathrm{KL}[p_i(z|y)|r(z|y)]\right] \tag{17}$$

$$\Leftrightarrow \mathbb{E}_{p_i(x,y,z)}[\log p(z|x) - \log r(z|y)] \geq \mathbb{E}_{p_i(y,z)}[\log p_i(z|y) - \log r(z|y)] \tag{18}$$

$$\Leftrightarrow \mathbb{E}_{p_i(x,y,z)}[\log p(z|x)] \geq \mathbb{E}_{p_i(y,z)}[\log p_i(z|y)] \tag{19}$$

$$\text{(since } \mathbb{E}_{p_i(x,y,z)}[\log r(z|y)] = \mathbb{E}_{p_i(y,z)}[\log r(z|y)]) \tag{20}$$

$$\Leftrightarrow \mathbb{E}_{p_i(y)}\left[\mathbb{E}_{p_i(x,z|y)}[\log p(z|x)]\right] \geq \mathbb{E}_{p_i(y)}\left[\mathbb{E}_{p_i(z|y)}[\log p_i(z|y)]\right] \tag{21}$$

$$\tag{22}$$

Therefore, we only need to prove that $\mathbb{E}_{p_i(x,z|y)}[\log p(z|x)] \geq \mathbb{E}_{p_i(z|y)}[\log p_i(z|y)]\ \forall y$.

This is equivalent to:

$$\int_z \int_x p_i(x,z|y) \log p(z|x) dx dz \geq \int_z p_i(z|y) \log p_i(z|y) dz \tag{23}$$

$$\Leftrightarrow \int_z \int_x p_i(x|y)[p(z|x) \log p(z|x)] dx dz \geq \int_z p_i(z|y) \log p_i(z|y) dz \tag{24}$$

$$\Leftrightarrow \int_z \mathbb{E}_{p_i(x|y)}[p(z|x) \log p(z|x)] dz \geq \int_z p_i(z|y) \log p_i(z|y) dz \tag{25}$$

Notice that the function $f(a) = a \log a, a > 0$ is a convex function since $f''(a) = \frac{1}{a} > 0\ \forall a > 0$. Hence, due to Jensen's Inequality we have:

$$\mathbb{E}_{p_i(x|y)}[p(z|x) \log p(z|x)] \geq b \log b \tag{26}$$

where $b = \mathbb{E}_{p_i(x|y)}[p(z|x)] = p_i(z|y)$.

Therefore, $\mathbb{E}_{p_i(x|y)}[p(z|x) \log p(z|x)] \geq p_i(z|y) \log p_i(z|y)$. Then, inequality 25 holds, and we conclude our proof. □

B Details on the ℓ_i^{CMI} objective

As mentioned in the main paper, we use a Gaussian distribution for both $p(z|x)$ and $r(z|y)$, so that the KL term can be computed analytically.

Let w_1 be the parameter of the (probabilistic) representation network. Given an input x_0 (of a datapoint (x_0, y_0)), the network output the mean and standard deviation of the Gaussian distribution $p(z|x_0)$ (both from the final layer of the network), denoted $\mu_{w_1}(x_0)$ and $\sigma_{w_1}(x_0)$. This means that $p(z|x_0) = \mathcal{N}(z; \mu_{w_1}(x_0), \sigma^2_{w_1}(x_0))$.

Also as pointed out in the paper, for a classification problem (where there is a finite number of the labels), we set $r(z|y) = \mathcal{N}(z; \mu_y, \sigma_y^2)$, where μ_y, σ_y ($y = \overline{1..C}$) are the variational parameters to be optimized. Therefore, for the datapoint x_0, y_0, we have $r(z|y_0) = \mathcal{N}(z; \mu_{y_0}, \sigma^2_{y_0})$.

Note that when z is a K-dimensional vector, $\mu_{w_1}(x_0), \sigma_{w_1}(x_0), \mu_{y_0}$ and σ_{y_0} are all K-dimensional.

The KL term in ℓ^{CMI} can be computed analytically as:

$$\log \sigma_{y_0} - \log \sigma_{w_1}(y_0) + \frac{\sigma^2_{w_1}(x_0) + (\mu_{w_1}(x_0) - \mu_{y_0})^2}{2\sigma^2_{y_0}} - \frac{1}{2} \tag{27}$$

If $K > 1$ (i.e., z is high dimensional), the calculations in Eq. 27 are element-wise, and the result are summed across the dimension K at the end.

As this computation is deterministic with respect to w_1 and μ_y, σ_y ($y \in \overline{1..C}$), the gradient w.r.t. these parameters can be computed straight-forwardly.

C Experiments

C.1 Hyper-parameters:

As mentioned in the main text, we use random search to tune the hyper-parameters of our method, namely α^{CMI} and α^{L2R}. Specifically, the tuned value for those parameters are:

RotatedMNIST

- FedL2R: $\alpha^{L2R} = 0.1$
- FedCMI: $\alpha^{CMI} = 0.3$
- FedSR: $\alpha^{L2R} = 0.1, \alpha^{CMI} = 0.3$

PACS

- FedL2R: $\alpha^{L2R} = 0.01$
- FedCMI: $\alpha^{CMI} = 0.01$
- FedSR: $\alpha^{L2R} = 0.01, \alpha^{CMI} = 0.001$

OfficeHome

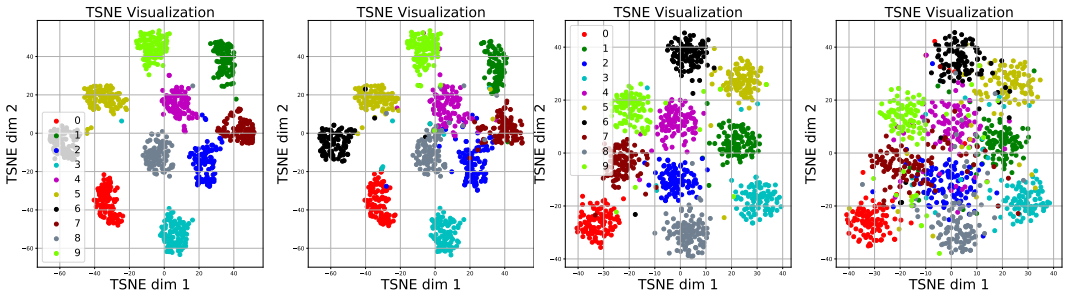
- FedL2R: $\alpha^{L2R} = 0.05$
- FedCMI: $\alpha^{CMI} = 0.0005$
- FedSR: $\alpha^{L2R} = 0.05, \alpha^{CMI} = 0.0005$

DomainNet

- FedL2R: $\alpha^{L2R} = 0.01$
- FedCMI: $\alpha^{CMI} = 0.005$
- FedSR: $\alpha^{L2R} = 0.01, \alpha^{CMI} = 0.0005$

C.2 Visualization of the Representation space

We also visualize the representation space of our method and observe that it aligns the representation much better than the conventional FL baseline FedAVG (which does not actively attempt to align the representation distributions). Specifically, Figure 1 show the t-SNE [1] visualization of the representation of FedSR and FedAVG in the RotatedMNIST experiment, with $\mathcal{M}_0$ (the target domain) and $\mathcal{M}_{15}$ (one of the source domains). We can clearly see that both the marginal and the conditional distributions of the representation are aligned better between the two domains for FedSR.



(a) FedSR (b) FedAVG

Figure 1: Visualization using t-SNE of the representation space of our method FedSR and the baselines FedAVG. For each method, the left subfigure corresponds to one source domain $\mathcal{M}_{15}$ and the right one corresponds to the target domain $\mathcal{M}_0$. Each color represents a digit class.

References

[1] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.