Importance Sampling Based Unsupervised Federated Representation Learning

Nazreen Shah*[1], Prachi Goyal*[1], Ranjitha Prasad[2]

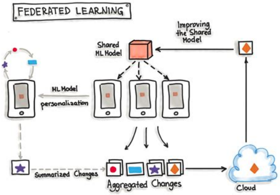
IEEE ICASSP, Seoul, South Korea, 2024

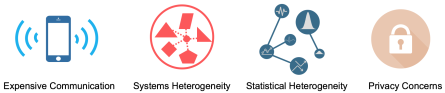
April 18, 2024

[1]Student Author
[2]Academic Advisor

Overview

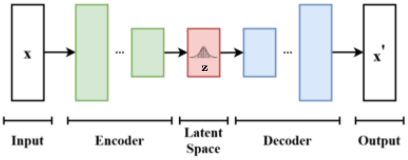
- Federated Learning(FL)[1] - Revolutionized the area of machine learning and data privacy.
- A significant challenge is unannotated/unlabeled data.
- Solution - unsupervised FL.





https://ml-ops.org/content/three-levels-of-ml-software.html
https://blog.ml.cmu.edu/2019/11/12/federated-learning-challenges-methods-and-future-directions/

- Representation learning - Popular way of unsupervised learning.
- Rich representations - important for downstream tasks.
- Variational Autoencoder(VAE)s are suitable for extracting meaningful representations.
- VAE enables uncertainty quantification and reduces overfitting.



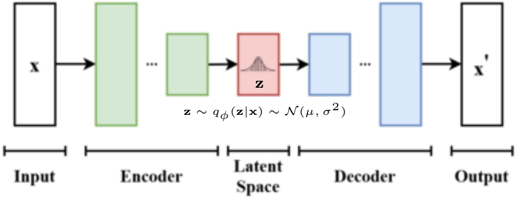
Can we formulate a distributed VAE model to achieve federated representation learning?

Contributions

- An importance-sampling based federated variational autoencoder framework: IS-FedVAE
  - Novel framework based on distributed evidence lower bound for VAE.
  - Enables generating globally relevant samples at the clients.
- Robustness to FL attributes like statistical heterogeneity, local epochs, and client participation.
- Demonstrated the effectiveness of samples for classification as a downstream task.

VAE decoder-encoder parameters are inferred through ELBO over N samples, where for a given sample $\mathbf{x}_i$ we have

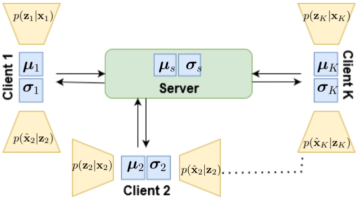
$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)}[\log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)||p(\mathbf{z})) = \mathcal{L}(\phi, \boldsymbol{\theta}; \mathbf{x}_i).$$



Architecture

- Assumption: Global latent distribution satisfies mean field decomposition.
- The optimal global variational distribution $q_\phi(\mathbf{z}|\mathbf{x}_i) \sim \mathcal{N}(\mu_{s,i}, \sigma^2_{s,i})$,

$$\mu_{s,i} = \frac{1}{K}\sum_{k=1}^{K}\mu_{k,i}, \quad \sigma^2_{s,i} = \frac{1}{K}\sum_{k=1}^{K}[\sigma^2_{k,i} + \mu^2_{k,i} - \mu^2_{s,i}].$$



Intuition

<u>GENERAL ELBO</u> :

$$\log p_{\boldsymbol\theta}(\mathbf{x}_i) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_i)}[\log p_{\boldsymbol\theta}(\mathbf{x}_i|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{x}_i|\mathbf{z})||p(\mathbf{z})) = \mathcal{L}(\boldsymbol\phi, \boldsymbol\theta; \mathbf{x}_i).$$

We rewrite the reconstruction error using the importance sampling approach at each client as,

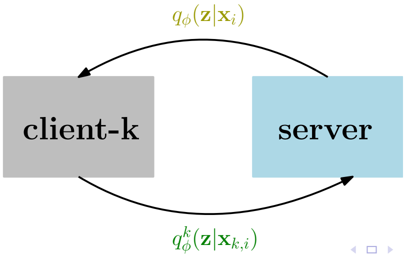
$$\mathbb{E}_{q_{\boldsymbol\phi}(\mathbf{z}|\mathbf{x}_{k,i})}[\log p^k_{\boldsymbol\theta}(\mathbf{x}_{k,i}|\mathbf{z})]$$

$$= \mathbb{E}_{q^k_{\boldsymbol\phi}(\mathbf{z}|\mathbf{x}_{k,i})}\left[\frac{q_{\boldsymbol\phi}(\mathbf{z}|\mathbf{x}_i)}{q^k_{\boldsymbol\phi}(\mathbf{z}|\mathbf{x}_{k,i})} \log p^k_{\boldsymbol\theta}(\mathbf{x}_{k,i}|\mathbf{z})\right].$$

KL divergence is computed as follows:

$$\mathrm{KL}(q_{\boldsymbol\phi}(\mathbf{z}|\mathbf{x}_i)||p(\mathbf{z})) =$$

$$\frac{1}{2}\sum_{i=1}^{d}\left[-\log\left(\sigma^2_{s,i}\right) + \sigma^2_{s,i} + \mu^2_{s,i} - 1\right].$$



Algorithm: IS-FedVAE: Importance Sampling based Federated Variational Autoencoder

Input: Dataset $\mathcal{D}_k$ at the k-th client, Number of communication rounds C, Number of local epochs E, Learning rate η, Initialize $p(\mathbf{z})$ at all clients;

for *C communication rounds* **do**

  At Server:

  Combine all the $q_{\phi}^{k}(\mathbf{z}|\mathbf{x}_{k,i})$;

  Communicate $q_{\phi}(\mathbf{z}|\mathbf{x}_k)$: to compute KLD and importance weights;

  At Client:

  for *E epochs* **do**

    $\forall k \in [K]$, sample mini-batch $B_k \subset \mathcal{D}_k$

    Optimize at each client to obtain $\{\boldsymbol{\theta}_k, \boldsymbol{\phi}_k\}$ by computing the loss based on importance-sampling based reconstruction loss along with KLD ;

    Communicate $q_{\phi}^{k}(\mathbf{z}|\mathbf{x}_{k,i}), \forall i$ to the server;

  end

end

Output: Per-client VAE: $\{\boldsymbol{\theta}_k, \boldsymbol{\phi}_k\}$ after C rounds.

Experimental Settings

- Qualitative analysis of the representations.
- Main attributes of FL: Local epochs, Statistical heterogeneity, and client participation.
- Comparison with baselines on the classification task.
  - Evaluation - Linear probe.
  - Metric - Accuracy.
- Datasets : CIFAR10, CIFAR100
- Data partitioning scheme: Dirichlet partitioning.
- Notations : E - Local epochs, α - Dirichlet parameter, K - No.of Clients, C - Communication rounds.

t-SNE and Scalability

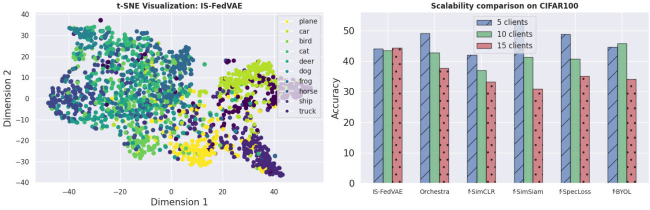


Figure 2: Left: t-SNE visualization of the latent space, Right: Scalability comparison on CIFAR100 dataset.

- Representations are well *separated*.
- *Scalable* and achieves similar performance across settings.

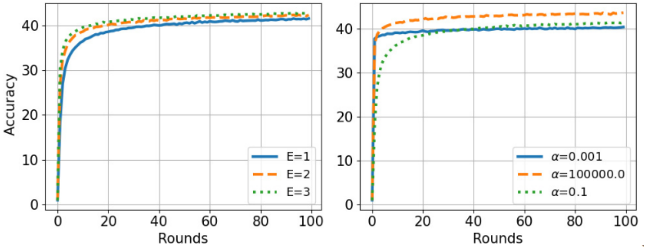


Figure 3: Left: Epoch-wise test accuracies on CIFAR100, Right: Test accuracies for varying levels of statistical heterogeneity

- Beneficial in scenarios where there are *high communication costs or computational constraints.*
- IS-FedVAE is *robust to varying levels of statistical heterogeneity*.

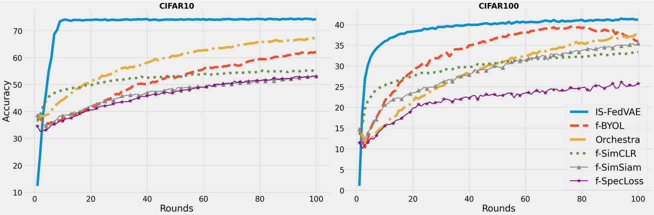


Figure 4: Convergence of IS-FedVAE as compared to baseline methods for $C = 100$ rounds.

- *Faster convergence* - addressing the challenge of communication bottle-neck

Accuracy: Comparison with baselines

Methods	CIFAR10	CIFAR100
IS-FedVAE	**77.19**	**43.05**
Orchestra [2]	70.64	35.64
f-BYOL [3]	66.18	38.88
f-SpecLoss [4]	64.53	35.99
f-SimSiam [5]	61.95	36.92
f-SimCLR [6]	58.15	33.49

Top-1 accuracy (%) comparison under the Linear evaluation protocol for statistically heterogeneous ($\alpha = 0.1$) setting for CIFAR10 and CIFAR100 datasets.

Conclusions

- We addressed a crucial problem of unsupervised federated representation learning.
- Proposed the novel IS-FedVAE using a distributed ELBO formulation.
- Demonstrated the robustness to FL attributes.
- Outperformed the state-of-the-art baselines.
- Future works
  - Other methods to utilize latent distributions in a federated setting.
  - improving representations to carry out specific downstream tasks.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.

[2] E. S. Lubana, C. I. Tang, F. Kawsar, R. P. Dick, and A. Mathur, "Orchestra: Unsupervised federated learning via globally consistent clustering," *arXiv preprint arXiv:2205.11506*, 2022.

[3] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

[4] J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma, "Provable guarantees for self-supervised deep learning with spectral contrastive loss," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5000–5011, 2021.

[5] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

Thank you!