LG-FAL : Federated Active Learning Strategy using Local and Global Models

SangMook Kim * sangmook.kim@kaist.ac.kr
SangMin Bae * bsmn0223@kaist.ac.kr
Se-Young Yun † yunseyoung@kaist.ac.kr
Kim Jaechul Graduate School of AI, KAIST

Hwanjun Song † hwanjun.song@navercorp.com
CLOVA AI Research, NAVER

Abstract

Recently, using unlabeled data points of each client in federated learning is attracting attention. Federated active learning is a framework that annotates and utilizes informative unlabeled instances at the client level. In FAL, we observe that there are two types of query selectors, namely 'global' and 'local only' models, and their performance dominance differs across datasets. In this paper, we analyze the advantages of the two query selectors and then propose a new FAL strategy, named LG-FAL, that exploits both benefits. Our experiments confirm that LG-FAL shows excellent performance on various benchmark datasets.

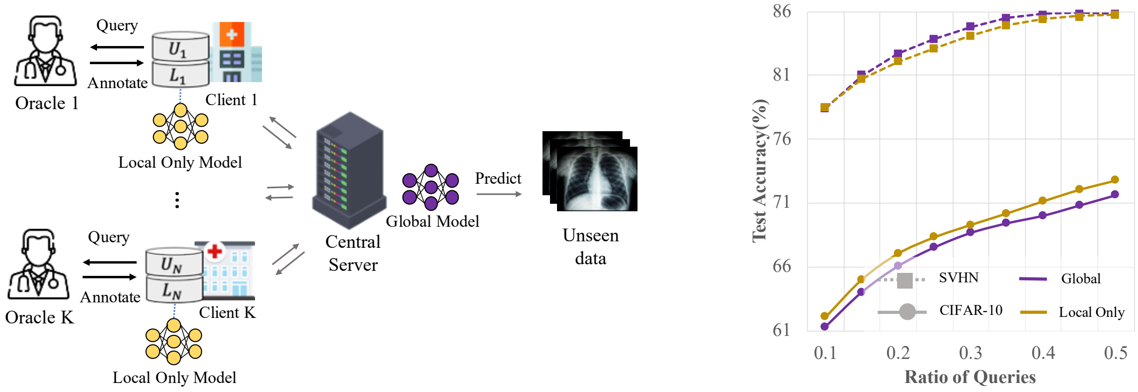
Keywords: Federated Learning, Active Learning

1. Introduction

Federated Learning (FL) is a distributed framework that allows multiple parties to learn machine models cooperatively without direct access to local client data for the privacy-preserving. Existing studies on federated learning assume the known ground-truth labels of the entire data, but in the real-world scenario, each client inevitably contains a large amount of unlabeled data due to the high cost of labeling. Annotating and exploiting the unlabeled set guarantees model performance improvement. However, as the labeling budget is limited in real applications, it is common to choose small but informative training instances to annotate. Therefore, active learning (AL) is becoming a promising learning protocol to reduce the high human-labeling cost, where a small number of maximally-informative instances are selected by a query strategy and labeled by an oracle.

Federated active learning (FAL) (Ahmed et al., 2020; Ahn et al., 2022) merges the philosophy of AL into FL. As demonstrated in Figure 1-(a), each client queries the instances with an AL sampling strategy, selecting potentially the most informative instances iteratively under the FL pipeline. Given a decentralized pipeline in FL, there are *two* types of available models as the query selector: (1) a 'global' model, which is the model globally optimized through model aggregation, and (2) a 'local only' model, which is the model separately trained in the client side only for its local data. So far, the global model has been recog-

[*] Both authors are contributed equally.
[†] Both authors are corresponding authors.



(a) FAL Framework.

(b) Global vs. Local Only Models

Figure 1: (a) shows the FAL framework where the performance of the global model is improved by continuously annotating unlabeled data on the local client-side, and (b) contrasts the performance of two query selectors, the global and the local only models, when BADGE was used as the query criterion.

nized as a better selector for FAL (Ahn et al., 2022), but we found an opposite result that the local only model can outperform the global model with respect to data diversity; the local only model provides more effective queries for FAL in CIFAR-10 data of Figure 1-(b).

Based on this finding, we study a new perspective of query selection in FAL – *the local only model helps increase intra- and inter-class diversity of the selected query instances*. In this paper, we propose a new FAL querying scheme, coined as **LG-FAL** (Local and Global FAL), which leverages both local only and global models throughout two phases, decoupling the role of the two models for query selection: (1) the first phase improves the intra- and inter-class diversity of the selected instances with the local only model (2) while the second phase improves the informativeness of them with the global model. We verified that LG-FAL consistently outperforms other combinations of query selectors, naemly global only, local only, and their ensemble model, on CIFAR-10, SVHN, and medical datasets.

2. Problem Definition

Let U_k^r be the unlabeled set of k-th client at the round r, where $k = \{1, \ldots, K\}$. At the first AL round (*i.e.*, $r = 1$), each client randomly selects B instances $L_k^1 = \{x_1, \ldots, x_B\}$ from U_k^0 and labels them to obtain the initial labeled set $D_k^1 = \{(x_1, y_1), \ldots, (x_B, y_B)\}$. For the next round (*i.e.*, $r \geq 2$), each client selects additional B instances L_k^r from $U_k^{r-1} = U_k^{r-2} \setminus L_k^{r-1}$ based on the given querying criterion, they are labeled by the oracle to expand the previous labeled set to $D_k^r = D_k^{r-1} \cup \{(x_i, y_i) \,|\, x_i \in L_k^r\}$. For the evaluation of the selected instances, a global model θ is trained from the random initialization in a FL fashion. The objective of FL is to obtain the optimal parameter θ^*, minimizing the loss on $D^r = \cup_{k=1}^K D_k^r$ as

$$\theta^* = \arg\min_{\theta} f(\theta), \text{ where } f(\theta) = \frac{1}{|D^r|} \sum_{i=1}^{|D^r|} f_i(\theta), \tag{1}$$

where $f_i(\theta) = \ell(x_i, y_i; \theta)$ and $\ell(\cdot)$ is the loss function determined by the network parameter.

However, due to the data privacy, the global model is optimized based on the reformulated update rule on the partitioned data over clients, as follows:

$$f(\theta^r) = \sum_{k=1}^{K} \frac{|D_k^r|}{|D^r|} F(\theta_k^r), \text{ where } F(\theta_k^r) = \frac{1}{|D_k^r|} \sum_{(x_i,y_i)\in D_k^r} \ell(x_i, y_i; \theta_k^r), \tag{2}$$

where the model θ_k^r is updated locally in the client side for its local data D_k^r and then they are aggregated globally to generate a global model θ^r. The local update and model aggregation procedures are alternated until the global model converges; this is the most popular FL training pipeline proposed by FedAvg (McMahan et al., 2017).

Regarding the query selection per AL round, the converged global model θ^{r*} is typically used as the query selector. Since the global model is distributed to all clients in FL setup, each client queries the unlabeled instances based on the model for labeling. Let $\mathcal{A}(\cdot)$ be the querying function and B be the labeling budget. Then, the query set of the k-th client for the round $r+1$ is constructed by

$$L_k^{r+1} = \mathcal{A}(U_k^r, \theta^{r*}, B) \;\text{ where }\; U_k^r = U_k^{r-1} \setminus L_k^r. \tag{3}$$

The querying function $\mathcal{A}(\cdot)$ in Eq. (3) could vary depending on which active learning algorithm is used. For example, Confidence Sampling (Wang and Shang, 2014) queries the instances with the highest uncertainty as

$$\mathcal{A}(U, \theta, B) = \underset{L\subseteq U,\, |L|=B,\, x\in L}{\arg\max} \; - \, p(y|x;\theta)_{\hat{y}}, \tag{4}$$

where $\hat{y}$ denotes the index of top-1 class probability. After updating the available labeled dataset, *i.e.*, $D_k^{r+1} = D_k^r \cup \{(x_i, y_i) \,|\, x_i \in L_k^{r+1}\}$, we train the global model θ^{r+1} for the next AL round by Eq. (2). This procedure is repeated during the given AL rounds.

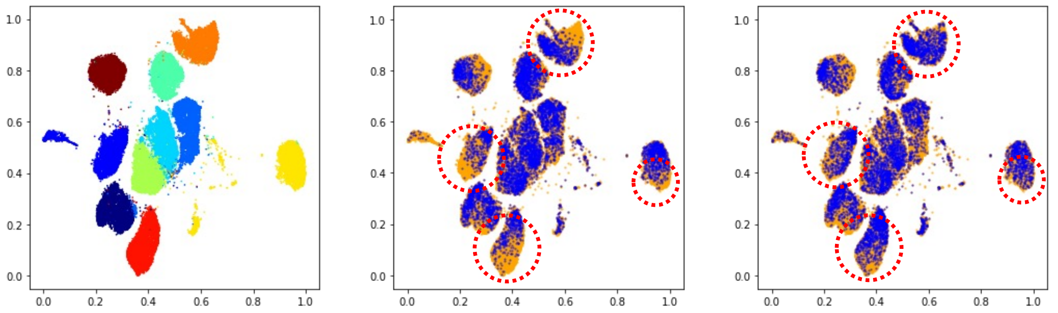
3. Proposed LG-FAL Method

3.1 Local Only vs. Global Models

We consider two different trained models: (1) the local only models optimized in each client for their local data (*i.e.*, the partitioned dataset D_k^r), and (2) the global model optimized for the entire data via the FL pipeline. In this section, we investigate what benefits the two models can provide as the query selector for FAL. We analyze the instances queried by the two different models using the Margin Sampling criterion on CIFAR-10; the other AL uncertainty-based strategies showed consistent trends.

Local Only Model helps increase intra- and inter-class diversity in query selection.

The intra-class diversity indicates how much the queried set can represent the entire unlabeled set within a specific class boundary. To measure this perspective, we visualize the feature embeddings via t-SNE (Van der Maaten and Hinton, 2008) in Figure 2. We can observe that the local only model takes more diverse instances when selecting the query (*i.e.*, the dotted red circles). In contrast, when the global model is used for all clients, the queried instances tend to be skewed towards the specific feature spaces. Meanwhile, for the local only models, the query set respectively labeled by different local models covers a wider region in their coverage because each local model can focus on its own local feature spaces.



(a) Entire Data. (b) Global Model. (c) Local Only Model.

Figure 2: t-SNE visualization to measure intra-class diversity. 'Yellow' and 'blue' colors in (b) and (c) represent the unlabeled set and the query set, respectively. The t-SNE embeddings of CIFAR-10 is extracted using pre-trained ResNet-18.

The inter-class diversity indicates the class balance with respect to the number of instances; that is, the labeled set should maintain the similar number of instances over classes. Table 1 measures the two indicators for inter-class diversity: (1) Imbalance Ratio (IR) (Buda et al., 2018), which is the proportion of the number of the majority class to that of the minority class, and (2) Earth Mover's Distance (EMD) (Zhao et al., 2018), which is the distance between two data distributions, *i.e.*, local client's data and the entire global data (IR and EMD scores are averaged over all clients). As shown in Table 1, since each local only model has high uncertainty on the minor class of its local training data, the class distribution of the locally queried set tends to be balanced, resulting in a better balance of aggregated query set over clients. Therefore, the IR and EMD scores of the local only model are much smaller than those of the global model. In addition, the lower EMD score between the global and local distribution indicates *weight divergence* of hindering model convergence can be alleviated.

Table 1: The quantitative analysis of the inter-class diversity.

Selector	IR	EMD
Global	$27.7_{\pm 18.0}$	$0.33_{\pm 0.06}$
Local Only	$10.5_{\pm 7.9}$	$0.24_{\pm 0.06}$

Based on the results, we confirm that the local only model considers the intra- and inter-class diversity better than the global model.

Global Model helps increase informativeness in query selection. The global model is the subject of learning and performing prediction. Hence, as witnessed by other FAL literature (Ahmed et al., 2020; Ahn et al., 2022), it excels at identifying informative instances in unlabeled data compared with the local only model.

3.2 Two-Phase Strategy Integrating Local Only and Global Models

To construct a query set with high diversity and informativeness, we propose a two-phase way named LG-FAL, integrating the local only and global models for query selection. Figure 3 is the overview of LG-FAL. Our algorithm consists of two phases, leading to (1) high intra- and inter-class diversity using the local only model and (2) high informativeness using the global model. For the ease understanding, let us assume a scenario where the k-th client queries B unlabeled instances in AL round $r+1$:

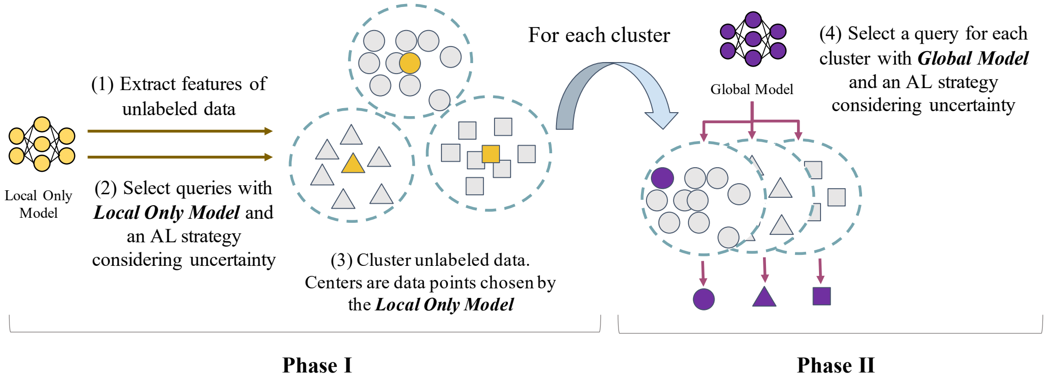

Local Only Model
(1) Extract features of unlabeled data
(2) Select queries with *Local Only Model* and an AL strategy considering uncertainty
(3) Cluster unlabeled data. Centers are data points chosen by the *Local Only Model*
For each cluster
Global Model
(4) Select a query for each cluster with *Global Model* and an AL strategy considering uncertainty
Phase I
Phase II

Figure 3: Overview of LG-FAL strategy. The labeling budget $B = 3$ in this example.

- **Phase I**: First, we extract the feature embeddings of the unlabeled set using the local only model $\theta^r_{k^*}$. Then, we select the query set with a uncertain sampling method $\mathcal{A}(U^r_k, \theta^r_{k^*}, B)$, where the sampling method can be any uncertainty-based or hybrid strategy. Lastly, the features obtained in the first step are clustered into B numbers. Here, the cluster centers correspond to the instances chosen by the local only model. Note that they are not directly used as the query, but the clusters provide representation boundaries with high intra- and inter-class diversity to the next phase.

- **Phase II**: We sample the final query instances based on the informativeness score computed by the global model. However, a stratified sampling is applied using the clusters obtained in Phase I. Let C_i be the i-th cluster. Then, the final B query instances are selected, one for each cluster; the most informative instance by the global model is selected respectively within each cluster boundary,

$$\mathcal{L}^{r+1}_k = \{\mathcal{A}(C_1, \theta^{r_*}, 1), ..., \mathcal{A}(C_B, \theta^{r_*}, 1)\} \tag{5}$$

  By doing this, class diversity can be maintained without losing informativeness. Therefore, our method can leverage the advantages of both global and local only models, achieving high intra- and inter-class diversity and high informativeness.

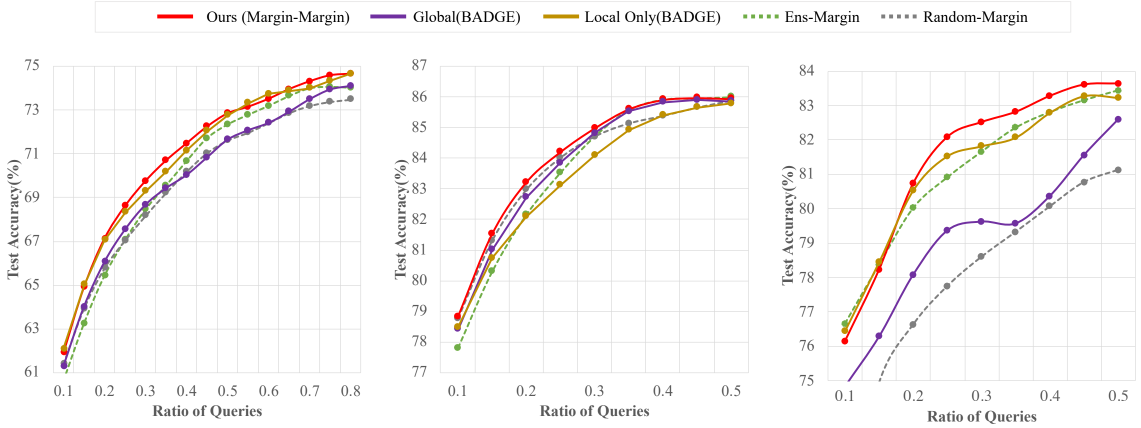
4. Performance Evaluation

The used datasets and evaluation setup is detailed in Appendix B for sake of space.

4.1 Local Only and Global Models vs. LG-FAL

We evaluate that LG-FAL outperforms the baseline methods using global and local only models as their query selector. For LG-FAL, we use Margin sampling in Phase I and II, but any combination of uncertainty-based sampling is possible. For the baselines, we combine them with the most representative AL method, BADGE, to select the query set; BADGE is known to achieve excellent performance by considering uncertainty and diversity even in the FL setting. The comparison with other combinations is provided in Appendix C.

Looking at the Figure 4, the proposed LG-FAL queries more diverse and informative instances, achieving consistently the best test accuracy in every datasets. This is because LG-FAL utilizes the advantages of local only and global models properly through its two



(a) CIFAR-10 (b) SVHN (c) PathMNIST

Figure 4: Comparison of active learning test accuracy on the various benchmark datasets.

phases. In contrast, there is no winner between Global (BADGE) and Local Only (BADGE) since their performance dominance is not consistent across datasets. Therefore, LG-FAL provides data-agnostic performance improvements over other baseliens.

4.2 Simple Ensemble vs. LG-FAL

An ensemble of local and global models can be a simple way to combine them. To compare LG-FAL with it, we analyze the effect of the ensemble by applying *Margin* sampling to the average of prediction vectors of the local only and global model. 'Ens-Margin' of Figure 4 shows the Ensemble-Margin sampling performance. Its performance lies in the middle between the performance of the local only model and the global model. That is, this simple way suffers from a performance trade-off between the two models. However, LG-FAL shows consistently the best performance by decoupling the role of two models through two phases.

4.3 Inter-Class Diversity: Random Sampling vs. LG-FAL

In Phase I of LG-FAL, we split representation space into B clusters that have centroids with high uncertainty; this improves intra- and inter-class diversity. We compare this approach with a simple random sampling method, which is a traditional way to keep intra-class diversity (see 'Random-Margin' in Figure 4). Compared to LG-FAL, there is a significant performance drop when using random sampling in Phase I. This is because random sampling cannot increase 'inter-class' diversity in the presence of the class imbalance problem in local data; that is, it samples more instances from a major class. In contrast, our uncertainty-based clusters give higher sampling weights to the instances in a minor class due to their high uncertainty, therefore achieving high intra- and inter-class diversity simultaneously.

5. Conclusion

We proposed a novel algorithm LG-FAL that selects diverse and informative queries using global and local only models. It was shown from various datasets that LG-FAL queries the instances with high intra- and inter-class diversity as well as high informativeness, compared with other combinations of query selection models and strategies. Our research will provide deep insight and research direction into the field of federated active learning.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST), 10%) and the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration, 90%)

References

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pages 137–153. Springer, 2020.

Lulwa Ahmed, Kashif Ahmad, Naina Said, Basheer Qolomany, Junaid Qadir, and Ala Al-Fuqaha. Active learning based federated learning for waste and natural disaster image classification. *IEEE Access*, 8:208518–208531, 2020.

Jin-Hyun Ahn, Kyungsang Kim, Jeongwan Koh, and Quanzheng Li. Federated active learning (f-al): an efficient annotation strategy for federated learning. *arXiv preprint arXiv:2202.00195*, 2022.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

Yoram Baram, Ran El Yaniv, and Kobi Luz. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5(Mar):255–291, 2004.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9368–9377, 2018.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9583–9592, 2021.

Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

Linton C Freeman. *Elementary applied statistics: for students in behavioral science*. New York: Wiley, 1965.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. *arXiv preprint arXiv:2203.11751*, 2022.

Yonatan Geifman and Ran El-Yaniv. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*, 2017.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Wei-Ning Hsu and Hsuan-Tien Lin. Active learning by learning. In *Twenty-Ninth AAAI conference on artificial intelligence*, 2015.

Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. *arXiv preprint arXiv:2006.12097*, 2020.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Michael Kampffmeyer, Arnt-Borre Salberg, and Robert Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–9, 2016.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Zhuoming Liu, Hao Ding, Huaping Zhong, Weijia Li, Jifeng Dai, and Conghui He. Influence selection for active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9274–9283, 2021.

Nan Lu, Zhao Wang, Xiaoxiao Li, Gang Niu, Qi Dou, and Masashi Sugiyama. Federated learning from only unlabeled data with class-conditional-sharing clients. *arXiv preprint arXiv:2204.03304*, 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Reza Haffari, Anton van den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. *arXiv preprint arXiv:2203.07034*, 2022.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34, 2021.

Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, pages 413–424. Springer, 2006.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Advances in neural information processing systems*, 20, 2007.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. Deep active learning: Unified and principled method for query and training. In *International Conference on Artificial Intelligence and Statistics*, pages 1308–1318. PMLR, 2020.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.

Jamshid Sourati, Murat Akcakaya, Jennifer G Dy, Todd K Leen, and Deniz Erdogmus. Classification active learning based on mutual information. *Entropy*, 18(2):51, 2016.

Jamshid Sourati, Murat Akcakaya, Todd K Leen, Deniz Erdogmus, and Jennifer G Dy. Asymptotic analysis of objectives based on fisher information in active learning. *The Journal of Machine Learning Research*, 18(1):1123–1163, 2017.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pages 112–119. IEEE, 2014.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, pages 1954–1963. PMLR, 2015.

Fengda Zhang, Kun Kuang, Zhaoyang You, Tao Shen, Jun Xiao, Yin Zhang, Chao Wu, Yueting Zhuang, and Xiaolin Li. Federated unsupervised representation learning. *arXiv preprint arXiv:2010.08982*, 2020.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019.

Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4912–4921, 2021.

Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Divergence-aware federated self-supervised learning. *arXiv preprint arXiv:2204.04385*, 2022.

Appendix A. Related Work

A.1 Active Learning

Active Learning (AL) is a promising approach to minimize the labeling effort by querying the most valuable samples among the pool of unlabeled dataset. There are three major types of active learning strategies; uncertainty sampling, representative sampling, and hybrid strategy. Uncertain sampling queries the most uncertain samples that lie on the current decision boundary. The predictive class probability distribution has been widely used to approximate the distance to the decision boundary; Confidence sampling (Wang and Shang, 2014), Margin sampling (Roth and Small, 2006), or Entropy sampling (Wang and Shang, 2014). In addition, there are various criterion for informativeness measure; the sample gradient (EGL (Settles et al., 2007), ISAL (Liu et al., 2021), and GraNd (Paul et al., 2021)), variance of model predictions (QBC (Seung et al., 1992), Variation Ratios (Freeman, 1965), Mean STD (Kampffmeyer et al., 2016), and ENS (Beluch et al., 2018)), mutual information-based objectives (Pess-MI (Sourati et al., 2016) and FIR (Sourati et al., 2017)), Bayesian uncertainty (BALD (Houlsby et al., 2011) and Deep Bayesian AL (Gal et al., 2017)), or the prediction for adversarial samples (DeepFool (Ducoffe and Precioso, 2018) and ALFA-Mix (Parvaneh et al., 2022)).

Representative sampling algorithms select a set of unlabeled instances that represents the entire unlabeled data distribution. Thus, optimization on chosen examples ensures a low error with respect to the full dataset. CoreSet (Sener and Savarese, 2017) and FF-Avtive (Geifman and El-Yaniv, 2017) define the problem of active learning as CoreSet selection and select a subset using the geometry of the datapoints. Inspired by the CoreSet method, Caramalau et al. (2021) adapt the CoreSet method on a sequential Graph Convolution Network (GCN) to measure the relation between labeled and unlabeled samples. Besides, Agarwal et al. (2020) replace the Euclidean distance with the pairwise Contextual Diversity in a CoreSet algorithm.

On the other hand, recent state-of-the-art algorithms support a hybrid of uncertainty and representative sampling. For example, COMB (Baram et al., 2004) and ALBL (Hsu and Lin, 2015) present a meta-active learning algorithm, which combines the various active learning strategies by using the multi-armed bandit. Two-step approaches have also been proposed; first prefilter with uncertainty sampling and then select the representative samples with a submodular data subset selection framework (FASS (Wei et al., 2015)) or K-means clustering (Diverse mini-Batch AL (Zhdanov, 2019)). Meanwhile, there are several existing approaches to simultaneously consider both the uncertainty and diversity via a VAE and discriminator (VAAL (Sinha et al., 2019)), the gradient embedding with a k-MEANS++ initialization scheme (BADGE (Ash et al., 2019)), or an adversarial loss with Wasserstein distance (WAAL (Shui et al., 2020)).

A.2 Federated Learning

McMahan et al. (2017) firstly introduce a Federated Learning (FL) framework, where multiple clients collaboratively train a central model while keeping the training data decentralized. FL can be categorized into cross-device FL and cross-silo FL in terms of the client device type, client statefulness, and distribution scale (Kairouz et al., 2021). Cross-device

FL supposes that the clients are a huge amount of mobile or edge devices, so each client is stateless (i.e., participates only once in a task) and highly unreliable due to wi-fi or slower connections. On the other hand, in a cross-silo FL setting, clients are from different organizations (e.g., medical or financial) and are almost always available during the FL training. However, the more sensitive data there is, the stronger data access restrictions may arise. The main bottleneck for FL is a statistical heterogeneity problem. Beyond the most classic algorithm, FedAvg (McMahan et al., 2017), there have been various algorithms minimizing the *weight divergence* between the local client updates and the aggregated gradient, such as FedProx (Li et al., 2020), SCAFFOLD (Karimireddy et al., 2020), FedDyn (Acar et al., 2021), and FedDC (Gao et al., 2022).

While most previous works focus on a supervised scenario, the assumption for the fully labeled samples is not suitable for FL and real-world scenarios. Therefore, the recent papers propose the methods to utilize the unlabeled data via active learning (Ahmed et al., 2020; **?**; Ahn et al., 2022), semi-supervised learning (Jeong et al., 2020), and self-supervised learning (Zhang et al., 2020; Zhuang et al., 2021, 2022; Lu et al., 2022). In particular, Ahn et al. (2022) argues that using an aggregated global model as a query selector outperforms the case of separately trained models at the client level. However, we found some contradictory results and investigated the reason in terms of the diversity and local class balance. In this paper, we propose a novel federated active learning algorithm that enables to simultaneously utilize the global and local only model.

Appendix B. Implementation Details

Training Settings In a federated active learning framework, we should not violate the fairness issue of labeling costs between clients. Therefore, we assume that 10 clients have the same number of total training samples and query the same number of instances per every AL round. The clients start with an initially labeled set (5% of their training dataset) and obtain the label of 5% samples for one AL round. Moreover, we basically consider a cross-silo FL setting where every client participates in every FL round. In local update steps, we used a stochastic gradient descent (SGD) with a learning rate of 0.01 and a momentum of 0.9 as an optimizer. We decayed the learning rate by 0.1 at half and three-quarters of federated learning rounds to ensure convergence. Besides, we set the number of FL rounds to 100 and local epochs to 5, respectively.

We evaluated our method with two classical public datasets, such as CIFAR-10 (Krizhevsky et al., 2009) and SVHN Netzer et al. (2011), and one medical benchmark, PathMNIST Kather et al. (2019). PathMNIST is a large-scale MNIST-like collection of standardized biomedical images about a colon pathology. It consists of training samples of 89,996 images for nine tissue classes and test samples of 7,180 images. Note that we used a random horizontal flipping for the data augmentation strategy. As for the choice of the architecture, we employed the four layers of convolution neural network for a base architecture. We left the experiments for various architecture and benchmarks as future work.

Heterogeneity Setting For the data heterogeneous FL scenario, we adopt Latent Dirichlet Allocation (LDA) strategy (Wang et al., 2020; Li et al., 2021), where each client k is assigned the partition of classes by sampling $\mathbf{p}_k \sim Dir(\alpha \cdot \mathbb{1})$, where $\mathbb{1} \in \mathbb{R}^C$. α is a concentration parameter that controls the data heterogeneity. The smaller α, the more

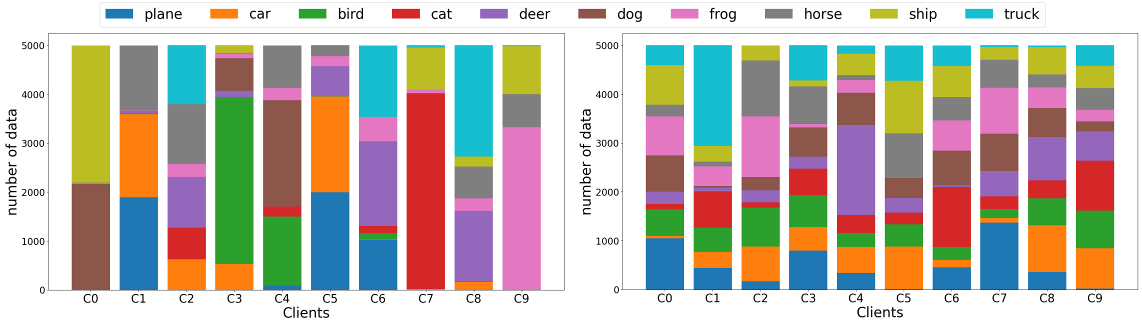


Figure 5: Visualization of the client data distribution on CIFAR-10. The concentration parameter $\alpha = 0.1$ (Left) and $\alpha = 1.0$ (Right). The x-axis and y-axis denote the client index and the number of data, respectively.

heterogeneous scenario. Since we consider a fairness issue in the FAL framework, the total number of samples is equally partitioned for all clients. Therefore, we made a doubly stochastic matrix $P = [\tilde{\mathbf{p}}_1, \ldots, \tilde{\mathbf{p}}_K]^\top$ by scaling $\mathbf{p}_k$ to $\tilde{\mathbf{p}}_k$, when the number of client and image class are same (i.e., P is a square matrix). Note that we set the sum of columns and rows to the proper values for a non-square matrix P. We visualized the examples of CIFAR-10 when $\alpha = 0.1$ and 1.0 in Figure 5.

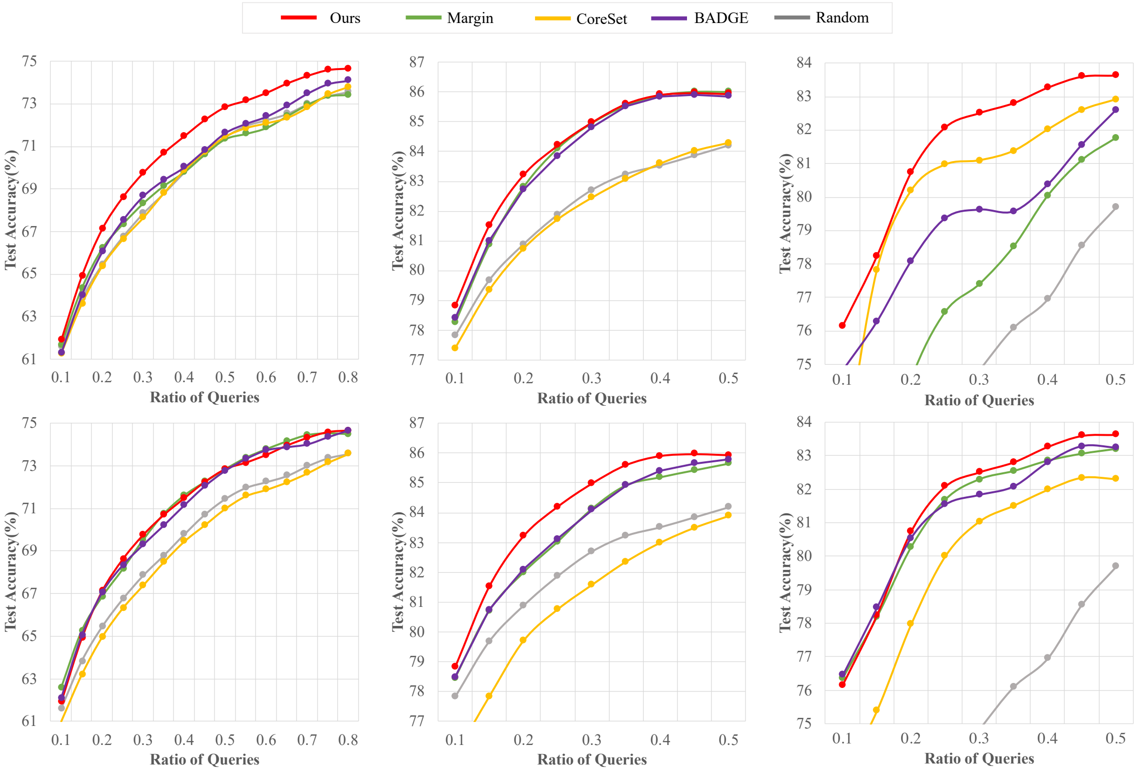
Appendix C. Performance Comparison with AL Baselines

In this Section, we considered the following active learning strategies:

- **Random sampling** randomly selects B samples from the unlabeled pool dataset.
- **Margin sampling** selects the smallest $p(y|x;\theta)_{\hat{y}} - p(y|x;\theta)_{y'}$ where $\hat{y}$ and y' are the indices of the largest and second largest class probability (Roth and Small, 2006).
- **CoreSet selection** chooses the small subset that can represent the whole unlabeled set (Sener and Savarese, 2017).
- **BADGE** selects groups of points that are disparate and high magnitude when represented in a hallucinated gradient space (Ash et al., 2019).

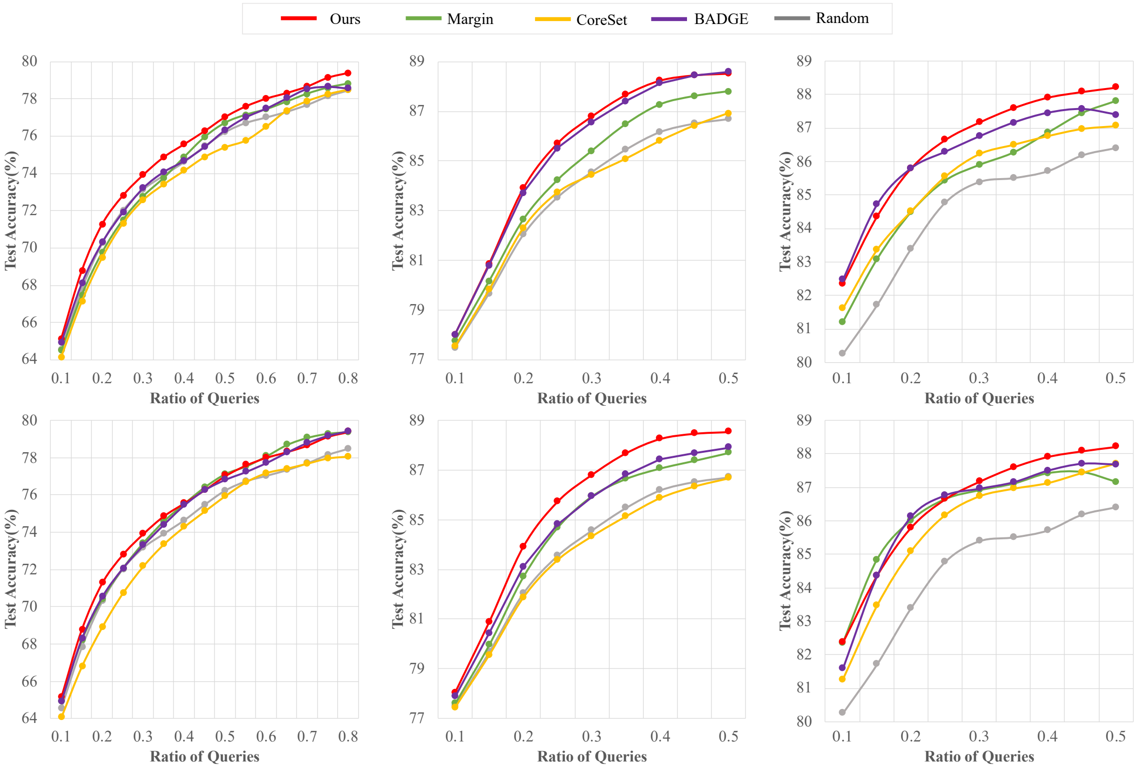
We compared our LG-FAL algorithm to the other active learning strategies in Figure 6 and 7. LG-FAL outperforms all the query selector and AL strategy combinations in various benchmarks and heterogeneity levels.

By the way, CoreSet and Random sampling mostly showed the poor performance against the uncertainty-usage strategies. The reason for the lower accuracy is the lack of the informativeness and inter-class diversity, even though they can consider the intra-class diversity. Figure 7 shows the performance of $\alpha = 1.0$ setting, where the local data heterogeneity between clients is alleviated. Especially, the performance gap between our LG-FAL and BADGE with the global model has been decreased. It is because the intra-class diversiy can guarantee the inter-class diversity as the client has more class balanced dataset.



(a) CIFAR-10 (b) SVHN (c) PathMNIST

Figure 6: The performance comparison with Global (Top) and Local Only models (Bottom) when $\alpha = 0.1$.



(a) CIFAR-10

(b) SVHN

(c) PathMNIST

Figure 7: The performance comparison with Global (Top) and Local Only models (Bottom) when $\alpha = 1.0$.