Re-thinking Federated Active Learning based on Inter-class Diversity

SangMook Kim\textsuperscript{1*} Sangmin Bae\textsuperscript{1*} Hwanjun Song\textsuperscript{2†} Se-Young Yun\textsuperscript{1†}

\textsuperscript{1}KAIST AI \textsuperscript{2}NAVER AI LAB

\{sangmook.kim, bsmn0223, yunseyoung\}@kaist.ac.kr ghkwsns91@gmail.com

Abstract

Although federated learning has made awe-inspiring advances, most studies have assumed that the client’s data are fully labeled. However, in a real-world scenario, every client may have a significant amount of unlabeled instances. Among the various approaches to utilizing unlabeled data, a federated active learning framework has emerged as a promising solution. In the decentralized setting, there are two types of available query selector models, namely ‘global’ and ‘local-only’ models, but little literature discusses their performance dominance and its causes. In this work, we first demonstrate that the superiority of two selector models depends on the global and local inter-class diversity. Furthermore, we observe that the global and local-only models are the keys to resolving the imbalance of each side. Based on our findings, we propose LoGo, a FAL sampling strategy robust to varying local heterogeneity levels and global imbalance ratio, that integrates both models by two steps of active selection scheme. LoGo consistently outperforms six active learning strategies in the total number of 38 experimental settings. The code is available at: https://github.com/raymin0223/LoGo.

1. Introduction

Federated learning (FL) is a distributed framework that allows multiple parties to learn a unified deep learning model cooperatively with preserving the privacy of the local client \cite{23, 28, 31}. Typically, FL has been actively studied in a standard supervised learning setting, where all the training instances are labeled, but it is more realistic for each client to contain both labeled and unlabeled data due to the high labeling cost \cite{29, 48}. Here, active learning (AL) can be a promising solution to improve the performance of a cooperated model with the pool of unlabeled data. In practice, federated active learning (FAL) framework has recently attempted to bridge two different philosophies in FL and AL \cite{3, 4}. As illustrated in Figure 1-(a) FAL framework alternates an FL procedure (red line) of training a predictive model collaboratively through local updates and aggregation phases, and an AL procedure (green line) of querying and annotating informative instances separately per client.

Although the overall framework just appears to be a straightforward fusion of two research fields, FL factors introduce two major challenges to the AL procedure. \textbf{First,} the class imbalance of the local dataset originates from heterogeneous distribution across local clients \cite{15, 23, 28}. Hence, in the FAL framework, the active selection algorithm has to ensure \textit{inter-class} diversity from both local and global perspectives. \textbf{Second,} there are two available types of query-selecting models, a \textit{global} model, which is globally optimized through the FL pipeline, and a \textit{local-only} model \cite{9, 33}, which can be separately trained only for each client. In the query selection phase, the global model can leverage the aggregated knowledge of all clients, while the local-only model is able to detect the most valuable instances for the local updates.

Prior FAL literature \cite{3, 4}, which simply adapt conventional AL strategies, had little discussion on these challenges. As our first contribution, we found the significant performance gap between two types of query selector (see Figure 1-(b)), and it is the first study to solve a conundrum of dominance trend by introducing two indicators of inter-class diversity\textsuperscript{1} \textit{local heterogeneity level} ($\alpha$) and \textit{global imbalance ratio} ($\rho$). The first indicator $\alpha$ is the concentration parameter of Dirichlet distribution, commonly seen in the FL literature \cite{1, 15, 23}, resulting in more locally imbalanced class distribution at lower values. Besides, $\rho$ indicator is the ratio of class imbalance for the aggregated global data of all clients \cite{7, 21}. We discovered three meaningful insights on selector dominance: \textbf{(Obs. 1)} Interestingly, the superiority of the two selectors varies depending on the two indicators of inter-class diversity, $\alpha$ and $\rho$. \textbf{(Obs. 2)} When local heterogeneity is severe ($\alpha$ is low), a local-only model is preferred for weighing minority instances for each client, and \textbf{(Obs. 3)} when globally minor classes exist ($\rho$ is high), the knowledge of the entire data distribution, inherent in a global model, is more essential. \textsuperscript{†}See Section 4

\textsuperscript{1}We use the term inter-class diversity interchangeably with class balance throughout the paper.
In a real FAL scenario, the superiority of two query models for a given dataset cannot be known in advance due to privacy preservation. Therefore, as our second contribution, we design a simple yet effective FAL querying approach, LoGo, that simultaneously leverages local-only and global models, to be robust to varying heterogeneity levels and global imbalance ratios. LoGo is a clustering-based sampling strategy, which is composed of macro and micro step exploiting local-only and global models, respectively. The rationale behind our method is that the optimal querying policy needs to evaluate the informativeness of instances with both models, which implicitly learn local and global data distribution, respectively. In a macro step, to improve the local inter-class diversity first, we perform k-means clustering [30] in hallucinated gradient space generated from local-only models. Then, in a micro step, the final query set is determined via one step of the EM algorithm [10], making cluster boundaries using instances from macro step (E-step) and cluster-wise sampling with the global model (M-step). The proposed cluster-wise sampling conservatively guarantees the diversity information of the macro step, i.e., the local inter-class diversity obtained by local-only models, while also considering the global minority classes via the global model. — See Section 5

As our third contribution, we conduct a total number of 38 experiments on five datasets using seven AL strategies including our LoGo algorithm. To verify the superiority of our method in real-world scenarios, we build comprehensive combinations of six categories, including query selector types (local-only vs. global models), local heterogeneity levels ($\alpha \in [0.1, \infty]$), global imbalance ratios ($\rho \in [1, 58]$), model architectures, budget sizes, and model initialization scheme. As a result, the experimental results empirically prove our three observations (Obs. 1–3). Besides, our method outperforms all other AL baselines and naïve implementations for an ensemble of two query selectors in extensive experimental settings. — See Section 6

2. Related Work

Federated Learning is a collaborative learning framework with multiple clients while maintaining the privacy of each client dataset. In this decentralized framework, FedAvg [31] is considered a de facto algorithm in which a server and clients communicate only model parameters efficiently. One of the main bottlenecks for FL is the statistical heterogeneity problem across the clients’ dataset, as the weight divergence from heterogeneous distributions hinders convergence during the aggregation scheme. Hence, several algorithms have tackled the local heterogeneity through the alignment between the local-updated gradient and the aggregated gradient, as a form of correction term [15, 23], regularization loss [1, 28], or distillation-based loss [25, 26]. Furthermore, there are existing literature to alleviate the class imbalance on the local side and they eventually achieve the global balance via rebalancing datasets [11] or a weighted loss function with a monitoring scheme [43].

Active Learning minimizes the labeling effort by querying the most informative instances from the unlabeled data pool. There are three major types of active learning strategies, namely uncertainty-based sampling, diversity-based sampling, and hybrid strategy. Uncertainty-based sampling queries the most uncertain instances that lie on the current decision boundary [12,14,35,41], while diversity-based sampling selects a set of unlabeled instances that represents the entire unlabeled data distribution [16,37]. Recently, hybrid strategies simultaneously consider both uncertainty and diversity. BADGE [6] utilized the gradient embedding as the uncertainty measure and selected the diverse query set by a k-means++ initialization scheme [5]. Several hybrid
AL methods are based on a submodular data subset selection [44], pairwise contextual diversity [2], or feature mixing [34]. In addition to the three categories, model-based strategies have also been recently proposed to train additional networks for query selection, such as a VAE and discriminator [38] or a sequential graph neural network [8].

**Federated Active Learning** has been recently studied to address a more realistic scenario, where clients have lots of the unlabeled instances [20, 47, 48]. However, previous works [3, 4] have not deeply discussed the challenges of FAL framework and naively applied existing AL strategies. Ahn et al. [4] have even considered the local-only query selector, but concluded that the global model is superior to the local-only model with limited experimental settings, three benchmarks and one heterogeneity level. In this work, we observe the counterparts in varying benchmarks and heterogeneity levels, which motivates us to design a novel sampling strategy by exploiting both global and local-only models. Our extensive analysis and experimental results encourage future research on the FAL problem.

### 3. Preliminary

**AL Procedure.** For the ease of understanding, we summarize notations in Table 1. At the first AL round (i.e., \( r = 1 \)), each client \( k \) randomly selects \( B \) instances, \( D_k^1 = \{x_1, \ldots, x_B\} \), from \( U_k^1 \), and oracles annotate them to obtain the initial labeled set \( D_k^1 = \{(x_1, y_1), \ldots, (x_B, y_B)\} \). For the next round \( (r \geq 2) \), based on the given querying strategy \( \mathcal{A}(\cdot) \) and the model parameters \( \Theta \), the query set of the \( k \)-th client at round \( r \) is sampled by

\[
L_k^r = \mathcal{A}(U_k^r, \Theta, B), \quad \text{where} \quad U_k^r = U_k^{r-1} \setminus L_k^{r-1}. \tag{1}
\]

The querying function \( \mathcal{A}(\cdot) \) in Eq. (1) depends on which AL algorithm is used. For example, Entropy sampling [41] queries the instances with the highest uncertainty like:

\[
\mathcal{A}(U, \Theta, B) = \arg \max_{x_i \in L, |L|=B} H(p(y|x_i; \Theta)) \tag{2}
\]

where \( H(p) = -\sum_{c=1}^C p_c \ln p_c \), and \( p \) is the predictive probability. The query set is annotated by the oracle and expanded to assemble the available labeled set, i.e., \( D_k^r = D_k^{r-1} \cup \{ (x_i, y_i) \mid x_i \in L_k^r \} \).

**FL Procedure.** After each AL round, we perform the FL procedure of which the objective is to obtain the optimal parameter \( \Theta^{r*} \) such that it minimizes the target loss on the given labeled set for all clients, \( D^r = \cup_{k=1}^K D_k^r \),

\[
\Theta^{r*} = \arg \min_{\Theta} \ell(x_i, y_i ; \Theta) \quad \text{and} \quad \ell(\cdot) \quad \text{is the loss function determined by the network parameter } \Theta.
\]

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

\[
f(\Theta^r) = \sum_{k=1}^K \frac{|D_k^r|}{|D^r|} F(\Theta_k^r),
\]

where

\[
F(\Theta_k^r) = \frac{1}{|D_k^r|} \sum_{(x_i, y_i) \in D_k^r} \ell(x_i, y_i ; \Theta_k^r). \tag{4}
\]

The model \( \Theta_k^r \) is updated locally on the client side for its local data \( D_k^r \), and then it is aggregated globally to generate a global model \( \Theta^r \). The local update and model aggregation steps are alternated until the global model converges; this corresponds to the most popular FL training pipeline, FedAvg proposed by [31].

The previous studies [3, 4] have typically used the converged global model \( \Theta^{r*} \) in the next AL round as the query selector of Eq. (2). However, considering the hierarchy structure in the FAL framework, it is also possible to use a separately optimized model on local partitioned data; replacing \( D^r \) with \( D_k^r \) in Eq. (3). It is often referred to as the local-only model [9, 33], and we denote it as \( \Theta_k^{r*} \). In the following section, we investigate what these models are specialized in and when they are beneficial to use.

### 4. Observation and Analysis

In this section, we analyze the performance trend between global and local-only models as the query selector, with respect to the degree of class imbalance in local and global data distribution. We synthetically adjust two indicators of inter-class diversity, \( \alpha \in \{0, 1, 0.5, \infty\} \) and \( \rho \in \{1, 5, 10, 20\} \), on CIFAR-10 benchmark. As \( \alpha \) and \( \rho \) get lower and higher, the levels of local heterogeneity and global imbalance increase, respectively (refer to Appendix A for the detailed data distribution). For both query selectors, we use Entropy sampling [41] as an active learning algorithm, and the training set is progressively labeled with the query ratio of 10% in each AL round.
The value of winning rate becomes 1 if the strategy
Here, the strategy
Definition 1. [36] Let \( \alpha_i \) and \( \alpha_j \) be the set of accuracies for two different FAL strategies \( i \) and \( j \). Then, t-score at AL round \( r \) is formulated as:

\[
t^r_{ij} = \frac{\sqrt{4\mu^r_{ij}}}{\sigma^r_{ij}}, \quad \text{where} \quad \mu^r_{ij} = \frac{1}{4} \sum_{l=1}^{4} (a^i_{r,l} - a^j_{r,l})
\]

\[
\sigma^r_{ij} = \sqrt{\frac{1}{3} \sum_{l=1}^{4} \left( (a^i_{r,l} - a^j_{r,l}) - \mu^r_{ij} \right)^2}.
\]  

Here, the strategy \( i \) is considered to beat the strategy \( j \) if \( t^r_{ij} > 2.776 \). Therefore, the winning rate for all AL rounds is formulated as follows:

\[
\text{win}^{ij} = \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}_{t^r_{ij} > 2.776}.
\]  

The value of winning rate becomes 1 if the strategy \( i \) beats the strategy \( j \) over all AL rounds.

Observation 1. The superiority of local-only and global query-selecting models varies according to the degree of local heterogeneity and global imbalance ratio.

In Figure 2, we summarize the performance gap between two query models depending on the local heterogeneity level (indicated by different shapes) and the global imbalance ratio (increased along x-axis). The y-axis represents the gap of the winning rate in Eq. (6) between global and local-only models; thus, the value becomes positive up to +1 if the global model beats the local-only model, otherwise negative up to -1. At a glance, there is a clear and consistent superiority of two query models according to \( \alpha \) and \( \rho \), where the dominance has intensified toward both extremes (e.g., upper right and lower left). This observation contradicts the previous findings that the global model has always outperformed the local-only model as the query selector in a FAL framework [4]. We provide more in-depth analysis in following Obs. 2 and Obs. 3.

Observation 2. As local heterogeneity increases (\( \alpha \downarrow \)), a local-only query selector is preferred due to the increased significance of local inter-class diversity.

As the collapse of the local inter-class balance (i.e., lower \( \alpha \)) incurs severe performance degradation due to a weight divergence [11, 31], addressing local imbalance can improve the learning stability and performance. Since the local-only models are separately trained on each client, in general, the local-only model has shown higher confidence for its own data distribution than the global model [9, 33]. In Figure 3, we visualized the number of class instances and the class-wise test accuracy in the first AL round (see the caption for details). Specifically, we confirmed a high correlation between counts and accuracy in Figure 3-(a), such that the local-only models have higher accuracy than the global model for major classes of their data distribution. Thus, by the nature of favoring low-confident instances in AL, the local-only model tends to select the instances with local minority classes as the query.

More precisely, we verify that the local-only model indeed queries the locally balanced set using earth mover’s distance (EMD) [46]. In Table 2, local EMD measures the mean of distance between class distribution of local query sets and a uniform distribution. The lower the value, the more balanced the locally queried instances. As shown in Table 2-(a) with high local heterogeneity, the local EMD of the local-only model (L) is lower than that of the global model (G). That is, the local-only model queries more diverse instances than the global model with respect to the local inter-class diversity.

Meanwhile, in the case of (b), the global model, trained with more samples, has higher accuracy over the classes due to little distribution discrepancy. Although the more accurate model is likely to have the higher prediction confidence, it does not mean that it is better at identifying the required instances based on the current local dataset, which the global model had not directly learned. In practice, the local-only model still chose the more locally balanced query set (Table 2-(b)), and we supposed this contradiction makes no sizeable winning gap of the case (b) in Figure 2.
Low global imbalance (\(\rho = 1\)) and high heterogeneity (\(\alpha = 0.1\)).

Observation 3. As the degree of global class imbalance increases (\(\rho \uparrow\)), it is more advantageous to exploit a global model that alleviates the global class imbalance.

Based on Obs. 2, the local-only model should outperform the global model in the case (c), but there is no clear superiority between them with respect to the winning rate in the case (c) of Figure 2. The only answer for this conundrum is the presence of global minority classes due to the high global imbalance ratio. The local heterogeneity is obviously a crucial factor by Obs. 2, but global class imbalance is also another factor that significantly degrades the classification performance in the FAL framework. Here, the major challenge is that neither the central server nor local clients cannot access any information of aggregated data due to privacy preservation. The only way to address this problem, we should utilize the global model that implicitly learns the knowledge of the entire data distribution through the aggregation phase in Eq. (4).

We introduce an additional global EMD, the indicator of measuring the inter-class diversity of the aggregated queried set over all clients. As can be seen in Table 2-(c), where the global imbalance ratio is high, we confirm that the global query selector (G) favors to query global minority classes. The global EMD of the global model is lower than that of the local-only model, i.e. the more globally balanced query set, but the local EMD is the opposite.

Meanwhile, in the case of (d), the minority classes are always the same from the global and local perspectives. It is different from the case (b), where the minority classes with respect to the accuracy differ in each client depending on the informativeness of instances despite the same instance number. Therefore, in this scenario, the global model has high confidence even in local datasets, leading to significantly overwhelming the local-only model in the case of (d) of Figure 2. In conclusion, the global inter-class diversity is an essential factor when global existence matters.
5. Method

From three observations in Section 4, we confirmed that FAL framework requires delicate consideration of local and global inter-class diversity. However, since clients are reluctant to share their data information, we should simultaneously utilize local-only and global models to ensure both sides of inter-class diversity. To this end, we propose a novel query sample strategy, named LoGo, that composed of macro and micro steps. Pseudo algorithm for LoGo is provided in Appendix C. Before describing details of each step, let us assume a scenario where k-th client queries B unlabeled instances at the round of r.

Macro Step: Clustering with Local-Only Model.

The ultimate goal of macro step is to satisfy local inter-class diversity by primarily sampling the informative instances by the local-only model. In detail, we introduce k-means clustering [30] on the hypothetical gradient embedding. Let \( z \) be the embedding vector before forwarding to the last layer \( W \) for \( x \in U_\beta^r \). Here, we utilize the gradient of negative cross-entropy loss, induced by a pseudo label, with respect to the last layer of encoder as follows:

\[
g_b^e = -\frac{\partial}{\partial W_{ce}} \ell_{CE}(x, \hat{y}; \Theta_{k+}^e) = z \cdot (1_{[\hat{y}=c]} - p_c), \tag{7}
\]

where \( \hat{y} = \arg\max_{c \in [C]} p_c \) and \( W_{ce} \) is the weights connected to \( c \)-th neuron of logits. Gradient embedding is widely used in conventional AL algorithm [6, 40], and we consider only the gradients corresponding to the pseudo label (i.e., \( g_b^e \)) for computation efficiency.

Then, we calculate \( B \) number of centroids on the hallucinated gradient space via EM algorithm [10] of \( k \)-means clustering by minimizing

\[
J = \sum_{i=1}^{N} \sum_{b=1}^{B} w_{ib} \| \tilde{g}_b^{x_i} - \mu_b \|^2, \tag{8}
\]

where \( w_{ib} \) is an indicator function whether \( g_b^{x_i} \) is assigned to \( \mu_b \) for E-step. Eq. (7) shows that the gradient embedding is a just scaling of feature embedding \( z \), especially with the scale of uncertainty. In other words, if an instance is uncertain to predict (low value of \( p_b \)), its gradient will be much highly scaled. In this sense, Eq. (8) can be regarded as a weighted \( k \)-means clustering [13, 39] on the feature embedding space by using the function of softmax response [17].

Hence, the macro step of LoGo enable the query set to contain both diversity in the embedding space and uncertainty from the perspective of local-only model.

Micro Step: Cluster-wise Sampling with Global Model.

In micro step, the global model selects the final instances that results in the higher global inter-class diversity. Given \( B \) selected instances from the macro step, the most uncertain instance is selected for each cluster:

\[
L_k^r = \{ A(C_1, \Theta^{+}, 1), ..., A(C_B, \Theta^{+}, 1) \} \tag{9}
\]

\( C_b \) denotes \( b \)-th cluster, generated by the query set from the macro step. We simply used Entropy sampling for \( A \), but we should note that \( A \) can be any AL sampling strategy.

The cluster-wise sampling in the micro step is a simple yet effective strategy to leverage the advantages of both query selector models, ensuring local and global inter-class diversity. Here, we discuss how LoGo can be a promising sampling strategy for the FAL framework.

Remark 1. The micro step is same with one step of EM algorithm. We first update cluster assignments with \( B \) number of centroids from the macro step (E-step). Then, within each cluster, we select the most informative instance based on the uncertainty score. This can be regarded as one M-step of the weighted k-means clustering, using infinitely scaled weight function of uncertainty measure.

Then, let \( c_b \) be a centroid of \( C_b \) and we define \( M \) as follows:

\[
M = \sum_{b=1}^{B} \| c_b - \tilde{x}_b \|^2, \quad \text{where } \tilde{x}_b = \arg \min_x \| c_b - x \|^2 \tag{10}
\]

where \( x \in L_k^r \) (Eq. (9)) and \( \tilde{x} \) is one-to-one mapping with the minimal transport cost. The lower value of \( M \) means that the final query set much consider the diversity in the macro step.

Remark 2. EM-based sampling in the micro step guarantees the diversity information of the macro step. Through sampling one instance per cluster, each \( \tilde{x}_b \) is disjointly assigned to one cluster. Therefore, LoGo conservatively ensures the local inter-class diversity from the macro step (i.e., lower \( M \) value), comparing to any other strategy that query at least two instances for one cluster.

6. Evaluation

6.1. Experimental Configuration

Training Settings. In a FAL framework, a central server should carefully consider the fairness of labeling and training costs between clients. Therefore, we assume that ten clients have the same size of unlabeled data pool and query the same number of instances per every AL round. Moreover, we focus on a cross-silo FL setting [22] where every client participates in every FL round.

We compared LoGo with six active learning strategies on five benchmark dataset (CIFAR-10 [24], SVHN [32], PathMNIST, DermaMNIST, and OrganAMNIST [45]). We considered 38 comprehensive experimental settings combined into six categories (see Figure 4). Except for learning ablations of the architectures, labeling budgets, and initialization schemes, in default, we implemented four layers of CNN and trained the encoder from scratch with the labeling budget of 5% for every AL round. We repeat all experiments four times and report their averaged values. Refer to Appendix D for detailed experimental settings.
Figure 4. Winning percentage across six categories. We also added defeat percentage, the black hatched bar that represents the percentage at which LoGo has been defeated by each baseline. Among total experiments, only statistically reliable values ($t$-score > 2.776) are considered. Thus, the lower value of the colored bar and the higher value of the black bar indicate a more comparable baseline.

Figure 5. Pairwise penalty matrix over 38 experimental settings. The value $P_{ij}$ indicates the number of times that the $i$-th strategy outperforms the $j$-th strategy (i.e., sum of win$^{ij}$ in Eq. (6) over 38 settings). The last row is the average number of times the $j$-th strategy is defeated by the rest strategies; the lower, the better.

**Baselines.** We considered six standard AL strategies. Random randomly samples the instances from the unlabeled pool. Entropy selects instances with the largest entropy [41]. CoreSet chooses the small subset that can represent the whole unlabeled set [37]. BADGE selects the diverse points with high magnitude in a hallucinated gradient space [6]. GCNAL adapts CoreSet on a sequential graph convolution network to measure the relation between labeled and unlabeled instances [8]. ALFA-Mix identifies the valuable instances by seeking inconsistencies in the prediction of mixing features [34]. We adopt these sampling strategies with either the global model or local-only model in FedAvg pipeline. Refer to Appendix E for the comparison of query selection cost. Besides, we summarized the results with various FL methods in Appendix F.

### 6.2. Overall Comparison

#### 6.2.1 Pairwise Penalty Matrix

We summarized the overall comparison results as a pairwise penalty matrix $P$ in Figure 5, following the recent work [6, 34]. Each cell of the pairwise penalty matrix represents the summation of the winning rate of Eq. (6) calculated above for each of the 38 experimental sets. As the rows of the matrix $P$, indicate how many times $i$-th algorithm outperforms the other algorithms, the brighter color means the better algorithm (conversely, the darker columns are the better). Note that we only considered statistically reliable results throughout the two-sided t-test.

LoGo defeats all baselines in general (see 7-th row values) while under-performed only 0.9 out of 38 times on average (see 7-th column values in the last row). In particular, LoGo outperforms BADGE and Entropy sampling, which are top-2 baselines, with 13.7 and 12.0 out of 38 times, while LoGo only loses 2.4 and 1.9 out of 38 times, respectively. This results demonstrate the robust effectiveness of our LoGo algorithm in various experimental settings.

#### 6.2.2 Winning Rate Bar Plots

Figure 4 summarizes the comparison results according to the six categories, which is the breakdown of Figure 5 on a more systematic categorization. For example, the colored bar of the overall category is calculated by dividing the total setting number of 38 in the 7-th row vector of Figure 5 (the black bar is from the 7-th column vector). The colored bar represents the average percentage at which LoGo defeats each baseline algorithm, and the higher bar means that LoGo has won more. Appendix G.1 provides the detailed comparison penalty values according to each category.

Overall, LoGo consistently shows an overwhelming winning percentage toward any baseline against the defeat percentage. Interestingly, ALFA-Mix, the latest AL strategy, shows performance considerably varying depending on which query selector is selected (see the query selector in Figure 4). The reason would be that a local-only model, which is separately trained on the highly heterogeneous dataset, is not appropriate for the feature mixing or the sensitivity to lots of hyperparameters. Moreover, CoreSet, which does not consider uncertainty, cannot resolve the local and global imbalance, showing similar performance as Random sampling. Thus, LoGo is a superior algorithm for a FAL framework in that it is robust to the type of query selector or most combinations of experimental settings.
6.2.3 Under Query Selector Category

Table 3 shows the test accuracy according to the increased labeling budgets over rounds on four datasets. Even with the same active learning strategy, a gap in test accuracy occurs depending on which query selector is used because the global imbalance varies by 1.0–58.7 across datasets. For example, in general, global models outperform local-only models for SVHN and DermaMNIST, while the opposite trend is observed for CIFAR-10 and PathMNIST. However, irrespective of the benchmarks and query model types, our LoGo shows the best performance in most cases. Two-step selection strategy enables LoGo to be robust by utilizing both the benefits of global and local-only models. Because we cannot know the degree of local and global imbalance in advance, our proposed method has a strong advantage in providing data-agnostic performance improvements over all baselines. Appendix G.2 provides a detailed performance comparison between LoGo and baselines under various experimental settings.

6.3. LoGo vs. Simple Ensemble

To prove that LoGo is an effective way to leverage the advantages of both models, we compared LoGo with three naive implementation of ensemble methods for global and local-only models: (1) the average of logits (for Entropy sampling) or gradient embedding (for BADGE) from two models, (2) weighing the instances based on the selection ranks (i.e., more weights if it is selected from both models), and (3) fine-tuning the global model on local datasets.

In Table 4, LoGo consistently shows better classification accuracy over increasing labeling budgets than three counterparts. Compared with the results in both Tables 3 and 4, all three ensemble methods show lower performance than using a single superior query selector. That is, the naive ensemble suffers from a performance trade-off between two query selector models and, therefore, their results fall somewhere in the middle of using global and local models.

7. Conclusion

We discovered the superiority of the two query selectors according to the local heterogeneity level and global imbalance ratio. Based on our findings, the local-only and global models are both crucial since global and local inter-class diversity affects their performance dominance. To this end, we propose LoGo algorithm that incorporates local-only and global models into macro and micro steps of cluster-based active learning. LoGo preferentially selects samples that simultaneously alleviate local and global imbalance as a query, making it robust to local and global imbalance. Our experiments verified that LoGo consistently outperforms six baselines under comprehensive settings of 38 combinations using six categories.

Acknowledgement. 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) and No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration).
References


[27] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In CVPR, pages 10713–10722, 2021. 11, 16


[31] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agueray Arcas. Communication- efficient learning of deep networks from decentralized data. In AISTATS, pages 1273–1282. PMLR, 2017. 1, 2, 3, 4
[32] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bis-
sacco, Bo Wu, and Andrew Y Ng. Reading digits in natural
images with unsupervised feature learning. 2011. 6

[33] Jaehoon Oh, Sangmook Kim, and Se-Young Yun. Fedbabu:
Towards enhanced representation for federated image classi-

[34] 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. 3, 4, 7, 8

[35] Dan Roth and Kevin Small. Margin-based active learning for
structured output spaces. In ECML, pages 413–424, 2006. 2


[37] Ozan Sener and Silvio Savarese. Active learning for convolu-
arXiv:1708.00489, 2017. 2, 7, 8

[38] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Vari-

[39] Helmuth Spath. Cluster analysis algorithms for data reduc-
tion and classification of objects. Ellis Horwood Chichester,
1980. 6

[40] Bindya Venkatesh and Jayaraman J Thiagarajan. Ask-n-
learn: Active learning via reliable gradient representations
2020. 6

deep learning. In 2014 International Joint Conference on
Neural Networks), pages 112–119, 2014. 2, 3, 7, 8

[42] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Pa-
pailiopoulos, and Yasaman Khazaeni. Federated learning
2020. 11

[43] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Address-
ing class imbalance in federated learning. In AAAI, vol-
ume 35, pages 10165–10173, 2021. 2

[44] Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in
data subset selection and active learning. In ICML, pages

[45] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin
Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni.
Medmnist v2: A large-scale lightweight benchmark for 2d
and 3d biomedical image classification. arXiv preprint
arXiv:2110.14795, 2021. 6

[46] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon
Civin, and Vikas Chandra. Federated learning with non-iid

[47] Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang,
and Shuai Yi. Collaborative unsupervised visual representa-
tion learning from decentralized data. In ICCV, pages 4912–
4921, 2021. 3

Divergence-aware federated self-supervised learning. arXiv