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

SangMook Kim[1*]   Sangmin Bae[1*]   Hwanjun Song[2†]   Se-Young Yun[1†]

[1]KAIST AI   [2]NAVER AI LAB

{sangmook.kim, bsmn0223, yunseyoung}@kaist.ac.kr   ghkswns91@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 [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 [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 [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. **First**, the class imbalance of the local dataset originates from heterogeneous distribution across local clients [15, 23, 28]. Hence, in the FAL framework, the active selection algorithm has to ensure *inter-class* diversity from both local and global perspectives. **Second**, there are two available types of query-selecting models, a *global* model, which is globally optimized through the FL pipeline, and a *local-only* model [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 [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[1] – *local heterogeneity level* (α) and *global imbalance ratio* (ρ). The first indicator α is the concentration parameter of Dirichlet distribution, commonly seen in the FL literature [1, 15, 23], resulting in more locally imbalanced class distribution at lower values. Besides, ρ indicator is the ratio of class imbalance for the aggregated global data of all clients [7, 21]. We discovered three meaningful insights on selector dominance: **(Obs. 1)** Interestingly, the superiority of the two selectors varies depending on the two indicators of inter-class diversity, α and ρ. **(Obs. 2)** When local heterogeneity is severe (α is low), a local-only model is preferred for weighing minority instances for each client, and **(Obs. 3)** when globally minor classes exist (ρ is high), the knowledge of the entire data distribution, inherent in a global model, is more essential. ▷ See Section 4
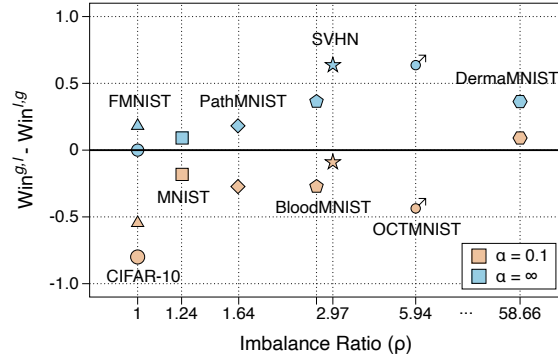
---

*equal contribution
†corresponding authors

[1]We use the term inter-class diversity interchangeably with class balance throughout the paper.
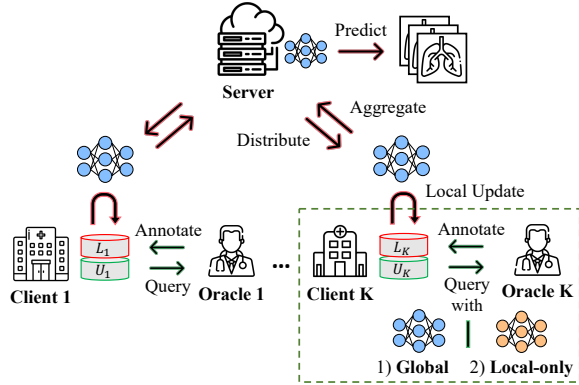
(a) Federated Active Learning framework.

(b) Superiority change of query selector.

Figure 1. Motivation: (a) The red and green lines correspond to the conventional FL and AL framework. We focus on a sampling strategy for FAL (green box) with the considerations of hierarchy structure and two available query-selecting models. (b) For a fixed querying strategy (Entropy sampling), the performance gap occurs only by changing the query selector. The y-axis is the gap in the winning rate in total active rounds (refer to Section 4). Closer to 1 indicates that global models outperforms than local-only models, and -1 is the opposite.

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 clutering-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

Indices:	
c	Index for a class ($c \in \{1, \ldots, C\}$)
r	Index for AL round ($r \in \{1, \ldots, R\} = [R]$)
k	Index for a client ($k \in \{1, \ldots, K\} = [K]$)
Parameters:	
B	Labeling budget for each AL round r
α	Local heterogeneity level
ρ	Global imbalance ratio
Data:	
U_k^r	Pool of unlabeled instances for a client k at round r
L_k^r	A queried instance set from U_k^r at round r
D_k^r	An available labeled set at round r
Weights:	
Θ^{r*}	Aggregated weights via FL phases on D^r (*global* model)
Θ_{k*}^r	Separately optimized weights on D_k^r (*local-only* model)

Table 1. Summary of notations throughout the paper.

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 naïvely 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, $L_k^1 = \{x_1, \ldots, x_B\}$, 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 Θ, the query set of the k-th client at round r is sampled by

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

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:

$$\mathcal{A}(U, \Theta, B) = \underset{x_i \in L, |L|=B, L \subseteq U}{\arg\max} H(p(y|x_i; \Theta)) \quad (2)$$

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 assembled to expand the available labeled set, *i.e.*, $D_k^r = D_k^{r-1} \cup \{(x_i, y_i) \,|\, x_i \in L_k^r\}$.

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

$$\Theta^{r*} = \underset{\Theta}{\arg\min} \, f(\Theta^r) = \underset{\Theta}{\arg\min} \frac{1}{|D^r|} \sum_{i=1}^{|D^r|} f_i(\Theta^r) \quad (3)$$

where $f_i(\Theta) = \ell(x_i, y_i; \Theta)$ and $\ell(\cdot)$ is the loss function determined by the network parameter Θ. However, due to data privacy, the global model is optimized based on the reformulated update rule on the partitioned data over clients:

$$f(\Theta^r) = \sum_{k=1}^K \frac{|D_k^r|}{|D^r|} F(\Theta_k^r),$$
$$\text{where} \ F(\Theta_k^r) = \frac{1}{|D_k^r|} \sum_{(x_i, y_i) \in D_k^r} \ell(x_i, y_i; \Theta_k^r). \quad (4)$$

The model Θ_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 Θ^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 Θ^{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 Θ_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, $\alpha \in \{0.1, 1.0, \infty\}$ and $\rho \in \{1, 5, 10, 20\}$, on CIFAR-10 benchmark. As α and ρ 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.
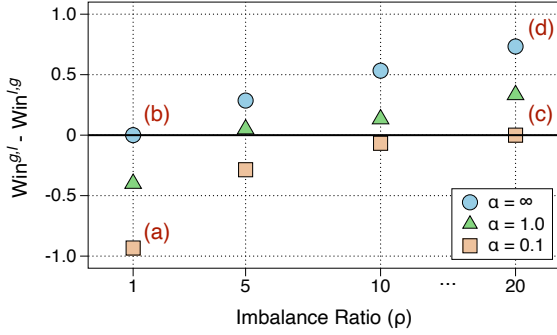
Figure 2. Gap between the winning rate of global and local-only models by varying global imbalance ratio (ρ) and local heterogeneity (α) on CIFAR-10 benchmark. The experimental setups for (a)-(d) are also compatible with Figure 3 and Table 2.

Comparison Metric. We evaluate the superiority over AL rounds through pairwise comparison [6, 34], widely used in conventional AL literature. We repeat each experimental setup, a pair of α and ρ, with four different seeds and obtain a set of four accuracy results $a_r = \{a_{r,1}, ..., a_{r,4}\}$ at each round r. Then, we conduct a two-sided t-test, where t-score is defined by Definition 1 for two given strategies i and j. Note that the strategy denotes the combination of the sampling strategy and the type of query selector.

Definition 1. [36] *Let a_r^i and a_r^j be the set of accuracies for two different FAL strategies i and j. Then, t-score at AL round r is formulated as:*

$$t_r^{ij} = \frac{\sqrt{4}\mu_r^{ij}}{\sigma_r^{ij}}, \quad \text{where } \mu_r^{ij} = \frac{1}{4}\sum_{l=1}^{4}\left(a_{r,l}^i - a_{r,l}^j\right)$$

$$\text{and } \sigma_r^{ij} = \sqrt{\frac{1}{3}\sum_{l=1}^{4}\left(\left(a_{r,l}^i - a_{r,l}^j\right) - \mu_r^{ij}\right)}. \tag{5}$$

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:

$$\text{win}^{ij} = \sum_{r=1}^{R}\frac{1}{R}\mathbb{1}_{t_r^{ij}>2.776}. \tag{6}$$

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 α and ρ, 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 α) 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. □

# of Instances

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
k1	0	0	0	0	0	220	1	2	277	0
k2	190	163	2	0	6	0	1	138	0	0
k3	0	75	0	68	98	0	26	122	0	111
k4	0	55	344	0	9	71	9	0	12	0
k5	8	0	124	22	0	225	29	92	0	0
k6	173	214	1	0	66	0	20	26	0	0
k7	108	0	11	22	173	0	58	0	0	128
k8	0	5	0	404	0	0	7	0	80	4
k9	0	21	0	1	163	0	25	49	18	223
k10	0	0	0	0	0	0	340	64	92	4
G.	479	534	484	520	519	521	522	500	487	479

Test Accuracy

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
k1	0	0	0	0	0	95.2	0	0	95.1	0
k2	83.7	86.7	0.5	0	2	0	0	82.2	0	0
k3	0	58.3	0	36.6	31.7	0	20.4	68.3	0	69.7
k4	0	70.4	89	0	0.4	37.7	11	0	24.8	0
k5	6.2	0	50.4	3.4	0	69.4	28.2	55.6	0	0
k6	81	80.8	0	0	65.9	0	27.4	37.8	0	0
k7	66.5	0	2.4	18.9	68.5	0	42	0	0	79.7
k8	0	0.9	0	97.2	0	0	3.5	0	78.1	2.2
k9	0	9.1	0	0.1	73.3	0	30.4	41	27.5	84.9
k10	0	0	0	0	0	0	95	62.6	88.4	5.4
G.	65.2	78.3	42.7	38.3	57	50.4	77.9	71.4	76.6	68.9

(a) Low global imbalance ($\rho = 1$) and high heterogeneity ($\alpha = 0.1$).

# of Instances

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
k1	50	51	50	49	48	50	50	50	52	50
k2	52	49	50	50	51	50	50	48	50	50
k3	50	50	50	50	49	51	50	52	50	48
k4	48	50	49	50	51	50	51	51	50	50
k5	50	50	50	51	50	50	50	49	48	52
k6	50	50	50	50	50	50	50	50	50	50
k7	50	50	50	50	50	49	50	50	51	50
k8	49	50	51	51	49	51	49	50	50	50
k9	51	50	49	50	50	50	50	50	50	50
k10	50	50	50	50	52	49	50	50	49	50
G.	500	500	500	500	500	500	500	500	500	500

Test Accuracy

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
k1	44.9	62.4	31.6	35.9	20.6	24.9	57.9	51.1	64.1	39.7
k2	58.9	53.6	25.3	29.6	20.4	42.4	37	37.7	27.9	29.2
k3	51.9	39.1	26.9	30.5	31.2	33.8	36.7	38.6	37.8	34.5
k4	38.2	56	28.7	23.6	38.9	29	41.9	29.5	47.3	34.6
k5	34.4	42.4	29.7	25	24.4	21.3	39.1	42	41.6	53.3
k6	45.7	41.7	27	22.8	22.8	36.1	37	49.2	39.2	36.5
k7	26.8	66	23.4	23	10.6	36.9	33.5	43.3	47.7	32.5
k8	39.9	43.9	19.2	11.9	38.6	26.9	46.4	42.2	47.9	41.7
k9	47.1	47.2	22.4	36.2	35.2	25.1	44.5	36.1	44.1	43.8
k10	38.3	37.3	27.4	11.2	35.6	26	24.8	43.2	50.3	33.4
G.	70.7	78.6	53	49	59.1	57.9	76.1	69.3	75.9	78.7

(b) Low global imbalance ($\rho = 1$) and low heterogeneity ($\alpha = \infty$).

# of Instances

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
k1	0	0	0	143	1	0	0	21	0	5
k2	7	163	0	0	0	0	0	0	0	0
k3	45	0	57	7	61	0	0	0	0	0
k4	7	17	98	35	0	0	7	0	0	6
k5	16	101	0	0	36	0	0	0	15	2
k6	44	1	0	11	23	81	0	0	0	10
k7	58	23	89	0	0	0	0	0	0	0
k8	48	62	0	0	8	5	47	0	0	0
k9	116	0	3	0	0	0	4	15	30	2
k10	170	0	0	0	0	0	0	0	0	0
G.	489	370	250	197	121	100	67	59	26	21

Test Accuracy

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
k1	0	0	0	98	0	0	0	37.5	0	15.6
k2	0	100	0	0	0	0	0	0	0	0
k3	79.9	0	37.1	3	73.7	0	0	0	0	0
k4	8.4	60.5	63.8	63.1	0	0	0.2	0	0	8.8
k5	40.7	90	0	0	68.3	0	0	0	27.2	0.4
k6	66.4	0.5	0	4.2	43.6	71.8	0	0	0	21.4
k7	75.5	27.8	82.7	0	0	0	0	0	0	0
k8	52.1	91.3	0	0	2.9	0.8	86.6	0	0	0
k9	84.9	0	0.8	0	0	0	21.1	31.9	32.1	1.3
k10	100	0	0	0	0	0	0	0	0	0
G.	81	84.6	56.4	35	41.2	29.2	42.5	12	12.8	8.3

(c) High global imbalance ($\rho = 20$) and high heterogeneity ($\alpha = 0.1$).

# of Instances

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
k1	48	35	29	16	18	6	7	4	5	2
k2	47	32	25	23	13	10	5	5	5	5
k3	51	40	22	16	12	13	7	5	3	1
k4	47	37	24	17	13	14	4	4	6	4
k5	47	41	28	15	13	11	5	5	2	3
k6	44	36	28	24	19	8	7	4	0	0
k7	56	32	22	18	12	12	8	3	4	3
k8	49	37	32	12	16	9	4	6	3	2
k9	45	33	32	22	14	9	7	2	3	3
k10	51	36	29	15	12	10	6	4	3	4
G.	485	359	271	178	142	102	60	42	34	27

Test Accuracy

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
k1	68.3	78.2	38.5	34.6	40	10.6	15.5	2.6	0.3	1.2
k2	70.8	31.6	51.8	31.4	13.1	2.7	7.4	3.3	0.6	1
k3	78.2	79.4	19.2	33.9	5.5	18.9	3.3	5.3	0.6	0
k4	43	66.7	24.9	28.8	23.9	35.9	0.1	0.6	29	9.8
k5	33.6	73.1	27.9	45.4	6.9	1.2	28.1	2.2	8.5	7.4
k6	52.3	67.7	40.8	67.9	2.5	1	5.1	4.8	0	0
k7	55.2	79.1	26.2	43.8	4.2	21.4	0.4	0	9.4	0.7
k8	42.7	83.3	42.3	16.9	35.4	15.6	0	8.8	1.3	2.3
k9	72.9	25.2	25.5	51.4	33.5	11.2	3.7	0.9	9.8	2.9
k10	63.6	47.1	14.8	68.1	15.8	19.2	2.5	12.5	0.9	3.6
G.	84.1	87.2	56.7	51.5	47.4	31.5	42.5	27.4	17.2	8.7

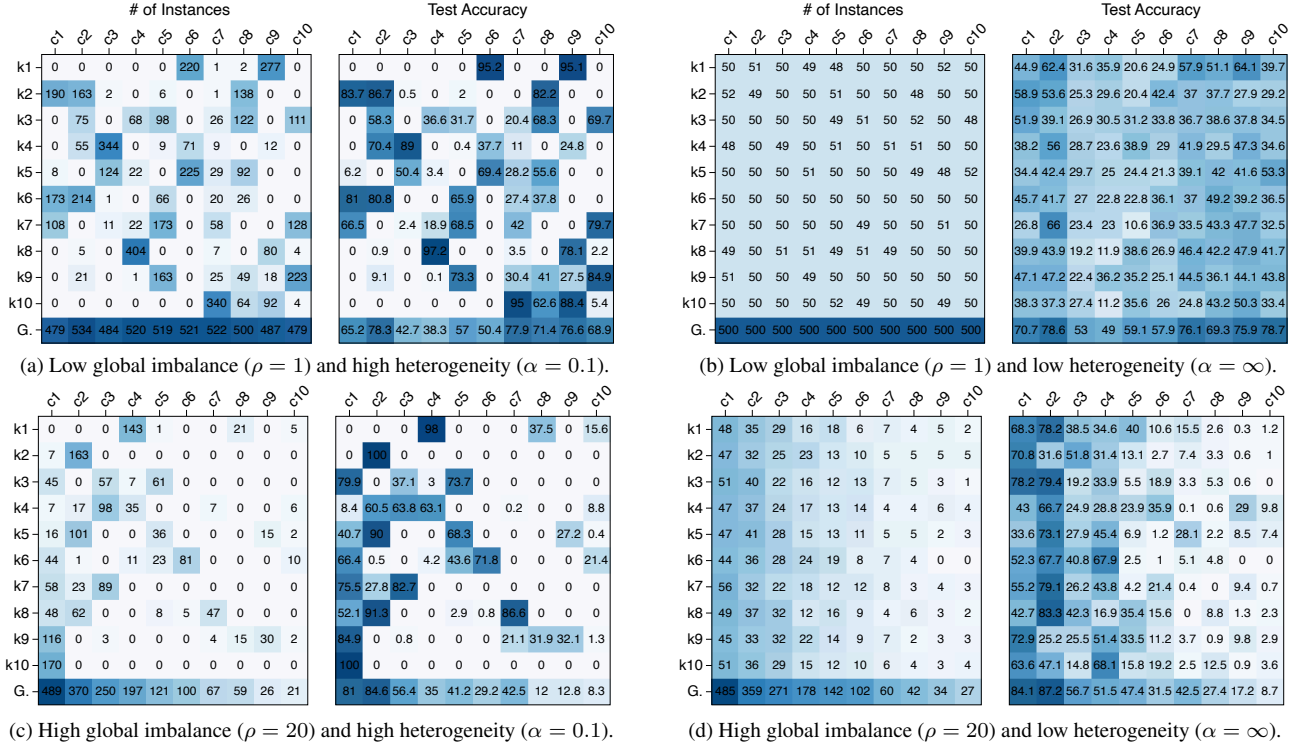(d) High global imbalance ($\rho = 20$) and low heterogeneity ($\alpha = \infty$).

Figure 3. Matrices of data count (left) and class-wise accuracy (right) for CIFAR-10, with four combinations of $\rho = \{1, 20\}$ and $\alpha = \{0.1, \infty\}$. k1–10 denote ten clients, c1–10 are ten classes of CIFAR-10. There are two types of data counts, # of per-client local instances (1-10th rows) and aggregated instances over all clients (the last row). Similarly, the test accuracy is measured with local-only models (1-10th rows) and a global model (the last row). (a)–(d) setups correspond to those of Figure 2. See Appendix B.1 for the more cases.

Case	Model	Obs. 2: Local EMD ($\downarrow$)				Obs. 3: Global EMD ($\downarrow$)			
		10%	20%	30%	40%	10%	20%	30%	40%
(a)	G	0.632	0.638	0.641	0.643	0.019	0.064	0.086	0.095
	L	0.632	0.597	0.592	0.595	0.019	0.050	0.050	0.046
(b)	G	0.049	0.077	0.070	0.084	0.014	0.070	0.066	0.063
	L	0.049	0.042	0.054	0.059	0.014	0.025	0.044	0.053
(c)	G	0.692	0.680	0.676	0.674	0.377	0.300	0.294	0.294
	L	0.692	0.641	0.633	0.636	0.377	0.334	0.326	0.321
(d)	G	0.371	0.298	0.284	0.274	0.368	0.294	0.282	0.272
	L	0.371	0.313	0.293	0.290	0.368	0.309	0.287	0.288

Table 2. Local EMD and global EMD on CIFAR-10. We summarize the results of four AL rounds with the labeling budget of 10% per round. (a)–(d) setups correspond to those of Figure 2. See Appendix B.2 for EMDs of more cases.

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 (d) of Figure 2. In conclusion, the global inter-class diversity is an essential factor when global imbalance exists. □

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 $x \in U_k^r$. 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:

$$g_c^x = -\frac{\partial}{\partial W_c} \ell_{CE}(x, \hat{y}; \Theta_{k*}^r) = z \cdot (\mathbb{1}_{[\hat{y}=c]} - p_c), \quad (7)$$

where $\hat{y} = \arg\max_{c \in [C]} p_c$ and W_c 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.*, $g_{\hat{y}}^x$) 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

$$J = \sum_{i=1}^{N} \sum_{b=1}^{B} w_{ib} \| g_{\hat{y}}^{x_i} - \mu_b \|^2, \quad (8)$$

where w_{ib} is an indicator function whether $g_{\hat{y}}^{x_i}$ is assigned to μ_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 $p_{\hat{y}}$), 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:

$$L_k^r = \{\mathcal{A}(\mathcal{C}_1, \Theta^{r*}, 1), ..., \mathcal{A}(\mathcal{C}_B, \Theta^{r*}, 1)\} \quad (9)$$

C_b denotes b-th cluster, generated by the query set from the macro step. We simply used Entropy sampling for $\mathcal{A}$, but we should note that $\mathcal{A}$ 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:

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

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.
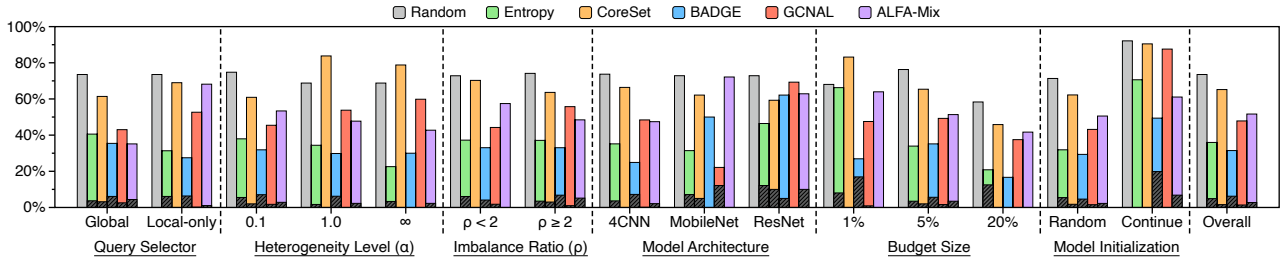
Figure 4. Winning percentage across six categories. We also added defeat percentage, the black hatched bar that represents the percentage at which LoGo has been defeated by each baseline. Among total experiments, only statistically reliable values (t-score > 2.776) are considered. Thus, the lower value of the colored bar and the higher value of the black bar indicate a more comparable baseline.
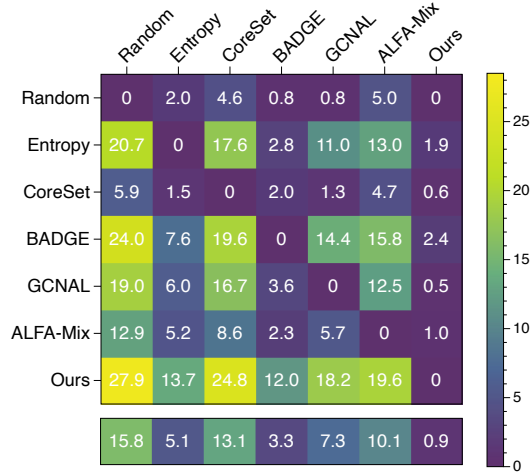

Figure 5. Pairwise penalty matrix over 38 experimental settings. The value P_{ij} indicates the number of times that the i-th strategy outperforms the j-th strategy (*i.e.*, sum of win^{ij} in Eq. (6) over 38 settings). The last row is the average number of times the j-th strategy is defeated by the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

Method	Model	CIFAR-10				SVHN				PathMNIST				DermaMNIST			
		20%	40%	60%	80%	20%	30%	40%	50%	20%	30%	40%	50%	20%	30%	40%	50%
Random	-	64.19	69.07	71.63	72.81	80.90	83.07	84.22	84.77	68.41	72.70	73.76	75.49	71.70	72.57	72.66	72.86
Entropy [41]	G	64.02	69.12	71.87	73.33	82.08	84.61	85.88	86.31	71.54	74.39	75.91	76.65	72.49	72.63	73.02	73.20
	L	66.29	71.45	73.51	74.02	82.09	84.58	85.69	86.18	76.52	78.29	78.71	79.10	71.38	72.04	72.22	72.65
Coreset [37]	G	64.66	69.43	71.75	73.1	80.94	82.74	83.81	84.46	74.84	76.24	76.85	76.80	72.02	72.16	72.34	72.74
	L	64.06	68.79	71.49	73.28	80.94	82.92	83.78	84.48	72.53	76.06	76.28	76.86	71.13	71.48	72.15	72.38
BADGE [6]	G	65.12	69.57	72.11	73.53	82.81	84.82	85.89	86.2	72.21	74.38	75.53	76.97	72.59	73.09	73.23	73.45
	L	66.32	71.28	73.41	74.28	82.69	84.67	85.61	86.1	76.48	78.51	78.42	78.68	71.35	72.13	72.25	72.99
GCNAL [8]	G	65.40	70.05	72.41	73.42	82.05	84.07	85.09	85.61	75.51	77.79	78.13	78.81	72.01	72.60	73.07	73.17
	L	65.62	70.18	72.36	73.42	81.92	83.58	84.55	85.10	74.85	76.46	77.18	77.45	71.95	72.91	72.91	73.29
ALFA-Mix [34]	G	65.45	69.87	72.24	73.29	83.02	84.99	86.05	86.33	73.34	74.83	76.31	77.43	72.39	73.14	73.27	73.10
	L	64.14	68.79	71.03	72.6	81.08	82.55	83.62	84.33	71.10	75.01	75.81	76.70	71.51	72.18	72.94	73.28
LoGo (ours)	G, L	**66.50**	**71.70**	**73.80**	**74.49**	**83.46**	**85.31**	86.02	**86.38**	76.32	**78.72**	**79.51**	**79.58**	**72.61**	**73.18**	**73.33**	**73.77**

Table 3. Comparison of test accuracy on four benchmarks with $\alpha = 0.1$. We reported the results with four random seeds. The baselines, except for Random sampling, are combined with two query selector models, G and L that stands for a global or local-only model, respectively. **Bold** and underline mean Top-1 and Top-2, respectively.

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 querying 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 naïve 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 naïve ensemble suffers from a performance trade-off between two

Method	Strategy	CIFAR-10				SVHN		
		20%	40%	60%	80%	20%	30%	40%
Ens. Logit	+Entropy	64.53	70.36	73.02	74.28	81.81	84.64	85.87
	+BADGE	65.55	70.31	72.83	73.97	82.77	84.76	85.90
Ens. Rank	+Entropy	65.90	70.92	73.34	74.20	82.15	84.38	85.64
	+BADGE	66.21	70.98	73.15	74.01	83.02	85.05	85.86
Fine-tuning	+Entropy	65.10	70.75	73.21	74.23	82.53	85.05	86.01
	+BADGE	65.82	70.95	72.94	74.12	82.59	84.89	85.82
LoGo (ours)	-	**66.50**	**71.70**	**73.80**	**74.49**	**83.46**	**85.31**	**86.02**

Table 4. Comparison of test accuracy on two benchmarks (α=0.1) with baselines using both global and local information.

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.

References

[1] 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. 1, 2

[2] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *ECCV*, pages 137–153, 2020. 3

[3] 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. 1, 3

[4] 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. 1, 3, 4

[5] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 2

[6] 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. 2, 4, 6, 7, 8

[7] 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. 1

[8] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *CVPR*, pages 9583–9592, 2021. 3, 7, 8

[9] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *ICML*, pages 2089–2099. PMLR, 2021. 1, 3, 4

[10] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 2, 6

[11] Moming Duan, Duo Liu, Xianzhang Chen, Yujuan Tan, Jinting Ren, Lei Qiao, and Liang Liang. Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications. In *ICCD*, pages 246–254, 2019. 2, 4

[12] Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018. 2

[13] Richard O Duda, Peter E Hart, et al. *Pattern classification*. John Wiley & Sons, 2006. 6

[14] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*, pages 1183–1192, 2017. 2

[15] 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. 1, 2

[16] Yonatan Geifman and Ran El-Yaniv. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*, 2017. 2

[17] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017. 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 16

[19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 16

[20] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with interclient consistency & disjoint learning. *arXiv preprint arXiv:2006.12097*, 2020. 3

[21] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019. 1

[22] 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. 6

[23] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, pages 5132–5143, 2020. 1, 2, 18

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[25] Gihun Lee, Yongjin Shin, Minchan Jeong, and Se-Young Yun. Preservation of the global knowledge by not-true self knowledge distillation in federated learning. *arXiv preprint arXiv:2106.03097*, 2021. 2

[26] 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. 2

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

[28] 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. 1, 2, 18

[29] Ekdeep Singh Lubana, Chi Ian Tang, Fahim Kawsar, Robert P Dick, and Akhil Mathur. Orchestra: Unsupervised federated learning via globally consistent clustering. *arXiv preprint arXiv:2205.11506*, 2022. 1

[30] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967. 2, 6

[31] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y 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 Bissacco, 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 classification. *arXiv preprint arXiv:2106.06042*, 2021. 1, 3, 4

[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

[36] Doug Semenick. Tests and measurements: The t-test. *Strength & Conditioning Journal*, 12(1):36–37, 1990. 4

[37] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 2, 7, 8

[38] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *ICCV*, pages 5972–5981, 2019. 3

[39] Helmuth Spath. *Cluster analysis algorithms for data reduction 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 for image classification. *arXiv preprint arXiv:2009.14448*, 2020. 6

[41] Dan Wang and Yi Shang. A new active labeling method for 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 Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020. 11

[43] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. In *AAAI*, volume 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 1954–1963, 2015. 3

[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 data. *arXiv preprint arXiv:1806.00582*, 2018. 4

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

[48] Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Divergence-aware federated self-supervised learning. *arXiv preprint arXiv:2204.04385*, 2022. 1, 3