Active Finetuning: Exploiting Annotation Budget in the Pretraining-Finetuning Paradigm

Yichen Xie[1], Han Lu[2], Junchi Yan[2], Xiaokang Yang[2], Masayoshi Tomizuka[1], Wei Zhan[1]
[1] University of California, Berkeley [2] Shanghai Jiao Tong University
{yichen_xie,tomizuka,wzhan}@berkeley.edu,{sjtu_luhan,yanjunchi,xkyang}@sjtu.edu.cn

Introduction

- recent success of deep learning heavily relies on abundant training data.

- the annotation of large-scale datasets often requires intensive human labor.

- inspires a popular pretraining-finetuning paradigm where models are pretrained on a large amount of data in an unsupervised manner and finetuned on a small labeled subset

Introduction

- A possible explanation comes from the batch-selection strategy of most current active learning methods. Starting from a random initial set, this strategy repeats the model training and data selection processes multiple times until the annotation budget runs out. Despite their success in from-scratch training, it does not fit this pretraining-finetuning paradigm well due to the typically low annotation budget, where too few samples in each batch lead to harmful bias inside the selection process.

Motivation

- To fill in this gap in the pretraining-finetuning paradigm,we formulate a new task called active finetuning, concentrating on the sample selection for supervised finetuning. In this paper, a novel method, ActiveFT, is proposed to deal with this task. Starting from purely unlabeled data, ActiveFT fetches a proper data subset for supervised finetuning in a negligible time. Without any redundant heuristics, we directly bring close the distributions between the selected subset and the entire unlabeled pool while ensuring the diversity of the selected subset. This goal is achieved by continuous optimization in the high-dimensional feature space, which is mapped with the pretrained model.

Methodology

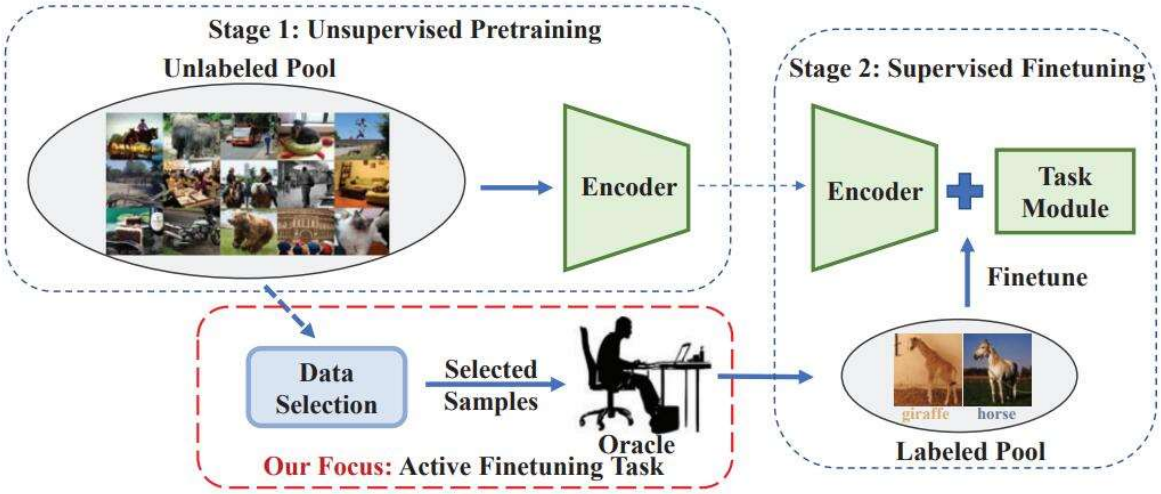


Figure 1. **Pretraining-Finetuning Paradigm:** We focus on the selection strategy of a small subset from a large unlabeled data pool for annotation, named as active finetuning task, which is under-explored for a long time.

$$f(\cdot; w_0) : \mathcal{X} \to \mathbb{R}^C$$

$$\mathcal{P}^u = \{\mathbf{x}_i\}_{i\in[N]} \sim p_u$$

$$\mathcal{S} = \{s_j \in [N]\}_{j\in[B]}$$

$$\mathcal{P}^u_{\mathcal{S}} = \{\mathbf{x}_{s_j}\}_{j\in[B]} \subset \mathcal{P}^u$$

$$\mathcal{P}^l_{\mathcal{S}} = \{\mathbf{x}_{s_j}, \mathbf{y}_{s_j}\}_{j\in[B]}$$

The goal of active finetuning is to find the sampling strategy minimizing the expected model error

$$\mathcal{S}_{opt} = \arg\min_{\mathcal{S}} \underset{\mathbf{x},\mathbf{y}\in\mathcal{X}\times\mathcal{Y}}{E} [error(f(\mathbf{x}; w_{\mathcal{S}}), \mathbf{y})] \quad (1)$$

Methodology

The difference with traditional active learning :

- 1) We have access to the pretrained model f, which will be finetuned, before data selection.
- 2)The selected samples are applied to the finetuning of the pretrained model f instead of from-scratch training.
- 3) The sampled subset size is relatively small, less than 10% in most cases.
- 4) We have no access to any labels such as a random initial labeled set before data selection.

Methodology

Select Samples:

- 1) bringing close the distributions between the selected subset $\mathcal{P}^u_{\mathcal{S}}$ and the original pool $\mathcal{P}^u \sim p_u$.
- 2) maintaining the diversity of $\mathcal{P}^u_{\mathcal{S}}$

$$\mathcal{S}_{opt} = \arg\min_{\mathcal{S}} \underset{\mathbf{x},\mathbf{y}\in\mathcal{X}\times\mathcal{Y}}{E} [error(f(\mathbf{x}; w_{\mathcal{S}}), \mathbf{y})] \quad (1)$$

$$\mathcal{S}_{opt} = \arg\min_{\mathcal{S}} D(p_{f_u}, p_{f_{\mathcal{S}}}) - \lambda R(\mathcal{F}^u_{\mathcal{S}}) \quad (2)$$

$$\theta_{\mathcal{S},opt} = \arg\min_{\theta_{\mathcal{S}}} D(p_{f_u}, p_{\theta_{\mathcal{S}}}) - \lambda R(\theta_{\mathcal{S}}) \; s.t. \; ||\theta^j_{\mathcal{S}}||_2 = 1 \quad \theta_{\mathcal{S}} = \{\theta^j_{\mathcal{S}}\}_{j\in[B]} \quad (3)$$

Methodology

Parametric Model Optimization:

$$p_{\theta_S}(\mathbf{f}) = \sum_{j=1}^{B} \phi_j p(\mathbf{f}|\theta_S^j) \tag{4}$$

$$p(\mathbf{f}|\theta_S^j) = \frac{\exp(sim(\mathbf{f}, \theta_S^j)/\tau)}{Z_j} \tag{5}$$

$$c_i = \arg\max_{j \in [B]} sim(\mathbf{f}_i, \theta_S^j) \tag{6}$$

Assumption 1 *$\forall i \in [N], j \in [B]$, if τ is small, the following far-more-than relationship holds that*

$$\exp(sim(\mathbf{f}_i, \theta_S^{c_i})/\tau) \gg \exp(sim(\mathbf{f}_i, \theta_S^j)/\tau), j \neq c_i$$

$$\begin{aligned} p_{\theta_S}(\mathbf{f}_i) &\approx \phi_{c_i} p(\mathbf{f}_i|\theta_S^{c_i}) \\ &= \frac{\exp(sim(\mathbf{f}_i, \theta_S^{c_i})/\tau)}{Z_{c_i}/\phi_{c_i}} \\ &= \frac{\exp(sim(\mathbf{f}_i, \theta_S^{c_i})/\tau)}{\tilde{Z}_{c_i}} \end{aligned} \tag{7}$$

$$\begin{aligned} KL(p_{f_u}|p_{\theta_S}) &= \sum_{\mathbf{f}_i \in \mathcal{F}^u} p_{f_u}(\mathbf{f}_i) \log \frac{p_{f_u}(\mathbf{f}_i)}{p_{\theta_S}(\mathbf{f}_i)} \\ &= \underset{\mathbf{f}_i \in \mathcal{F}^u}{E} [\log p_{f_u}(\mathbf{f}_i)] - \underset{\mathbf{f}_i \in \mathcal{F}^u}{E} [\log p_{\theta_S}(\mathbf{f}_i)] \end{aligned} \tag{8}$$

$$D(p_{f_u}, p_{\theta_S}) = - \underset{\mathbf{f}_i \in \mathcal{F}^u}{E} [sim(\mathbf{f}_i, \theta_S^{c_i})/\tau] \tag{9}$$

$$R(\theta_S) = - \underset{j \in [B]}{E} \left[\log \sum_{k \neq j, k \in [B]} \exp\left(sim(\theta_S^j, \theta_S^k)/\tau \right) \right] \tag{10}$$

$$\begin{aligned} L &= D(p_{f_u}, p_{\theta_S}) - \lambda \cdot R(\theta_S) \\ &= - \underset{\mathbf{f}_i \in \mathcal{F}^u}{E} [sim(\mathbf{f}_i, \theta_S^{c_i})/\tau] + \underset{j \in [B]}{E} \left[\log \sum_{k \neq j, k \in [B]} \exp\left(sim(\theta_S^j, \theta_S^k)/\tau \right) \right] \end{aligned} \tag{11}$$

$$\mathbf{f}_{s_j} = \arg\max_{\mathbf{f}_k \in \mathcal{F}^u} sim(\mathbf{f}_k, \theta_S^j) \tag{12}$$

Methodology

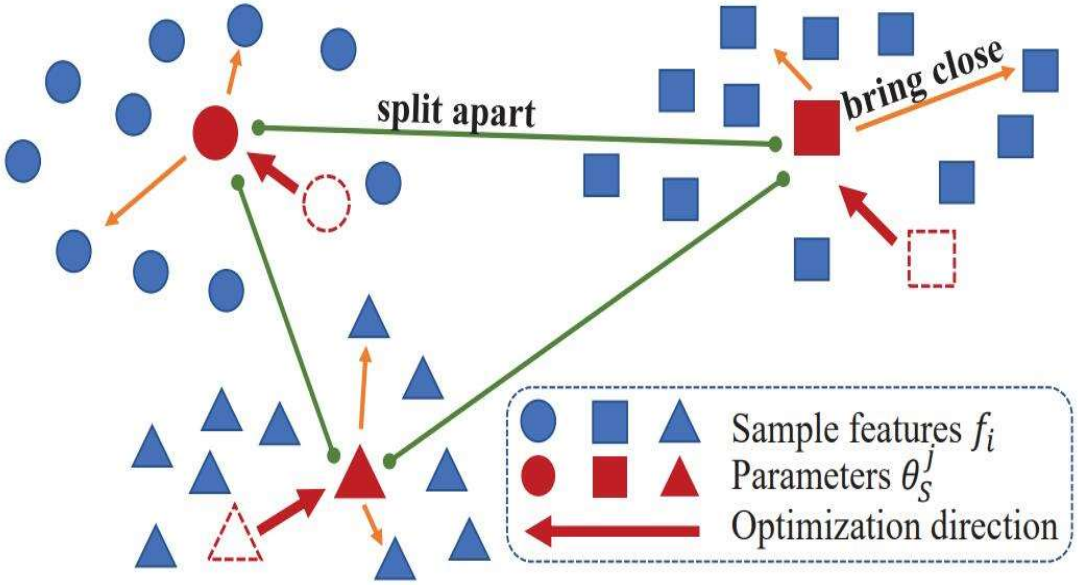


Figure 2. **Parametric Model Optimization Process:** By optimizing the loss in Eq. 11, each parameter θ_S^j is appealed by nearby sample features (orange in the figure, Eq. 9) and repelled by other parameters $\theta_S^k, k \neq j$ (green in the figure, Eq. 10).

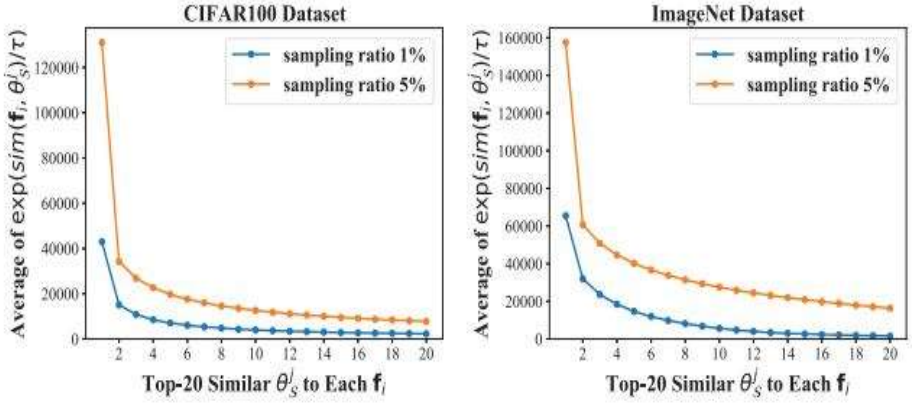


Figure 3. **Similarity between Features and Parameters:** On CIFAR100 and ImageNet, we find the Top-20 most similar parameters θ_S^j with each sample feature $\mathbf{f}_i$, and calculate the average exponential similarity $E_{i\in[N]}[\exp(sim(\mathbf{f}_i, \theta_S^j)/\tau]$. Here $\theta_S = \{\theta_S^j\}_{j\in[B]}$ is randomly sampled following the distribution p_{f_u}. The model $f(\cdot; w_0)$ is DeiT-Small [42] pretrained on ImageNet [37] with DINO framework [6]. The results verify Assumption 1 that the Top-1 similarity is significantly larger than others.

Methodology

Implementation as a Learning Model

Algorithm 1: Pseudo-code for ActiveFT

Input: Unlabeled data pool $\{\mathbf{x}_i\}_{i\in[N]}$, pretrained model $f(\cdot; w_0)$, annotation budget B, iteration number T for optimization

Output: Optimal selection strategy $\mathcal{S} = \{s_j \in [N]\}_{j\in[B]}$

for $i \in [N]$ **do**
  $\mathbf{f}_i = f(\mathbf{x}_i; w_0)$
/* Construct $\mathcal{F}^u = \{\mathbf{f}_i\}_{i\in[N]}$ based on $\mathcal{P}^u$, normalized to $||\mathbf{f}_i||_2 = 1$ */
Uniformly random sample $\{s_j^0 \in [N]\}_{j\in[B]}$, and initialize $\theta_{\mathcal{S}}^j = \mathbf{f}_{s_j^0}$
/* Initialize the parameter $\theta_{\mathcal{S}} = \{\theta_{\mathcal{S}}^j\}_{j\in[B]}$ based on $\mathcal{F}^u$ */
for $iter \in [T]$ **do**
  Calculate the similarity between $\{\mathbf{f}_i\}_{i\in[N]}$ and $\{\theta_{\mathcal{S}}^j\}_{j\in[B]}$: $Sim_{i,j} = \mathbf{f}_i^\top \theta_{\mathcal{S}}^j / \tau$
  $MaxSim_i = \max_{j\in[B]} Sim_{i,j} = Sim_{i,c_i}$
  /* The Top-1 similarity between $\mathbf{f}_i$ and $\theta_{\mathcal{S}}^j, j \in [B]$ */
  Calculate the similarity between $\theta_{\mathcal{S}}^j$ and $\theta_{\mathcal{S}}^k, k \neq j$ for regularization: $RegSim_{j,k} = \exp(\theta_{\mathcal{S}}^{j\top} \theta_{\mathcal{S}}^k / \tau), k \neq j$
  $Loss = -\frac{1}{N}\sum_{i\in[N]} MaxSim_i + \frac{1}{B}\sum_{j\in[B]} \log\left(\sum_{k\neq j} RegSim_{j,k}\right)$
  /* Calculate the loss function in Eq. 11 */
  $\theta_{\mathcal{S}} = \theta_{\mathcal{S}} - lr \cdot \nabla_{\theta_{\mathcal{S}}} Loss$
  /* Optimize the parameter through gradient descent */
  $\theta_{\mathcal{S}}^j = \theta_{\mathcal{S}}^j / ||\theta_{\mathcal{S}}^j||_2, j \in [B]$
  /* Normalize the parameters to ensure $||\theta_{\mathcal{S}}^j||_2 = 1$ */
Find $\mathbf{f}_{s_j}$ closest to $\theta_{\mathcal{S}}^j$: $s_j = \arg\max_{k\in[N]} \mathbf{f}_k^\top \theta_{\mathcal{S}}^j$ for each $j \in [B]$
Return the selection strategy $\mathcal{S} = \{s_j\}_{j\in[B]}$

Methodology

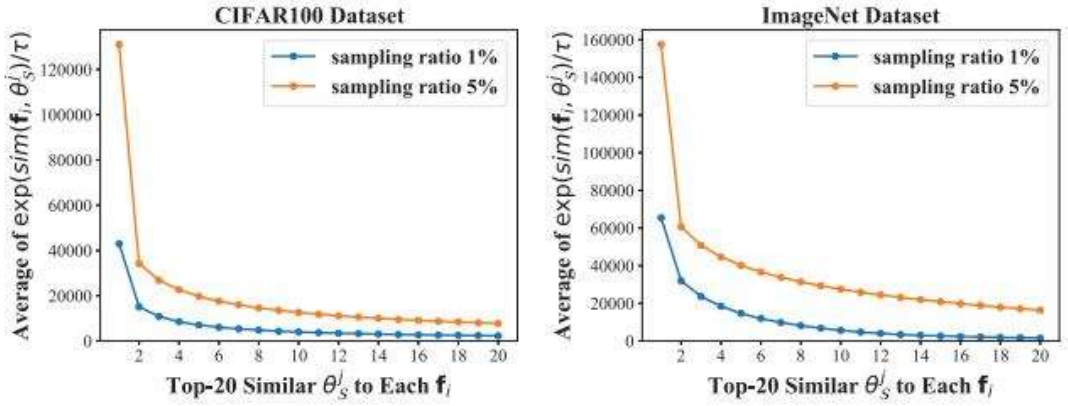


Figure 3. **Similarity between Features and Parameters:** On CIFAR100 and ImageNet, we find the Top-20 most similar parameters $\theta_{\mathcal{S}}^j$ with each sample feature $\mathbf{f}_i$, and calculate the average exponential similarity $E_{i\in[N]}[\exp(sim(\mathbf{f}_i, \theta_{\mathcal{S}}^j)/\tau]$. Here $\theta_{\mathcal{S}} = \{\theta_{\mathcal{S}}^j\}_{j\in[B]}$ is randomly sampled following the distribution p_{f_u}. The model $f(\cdot; w_0)$ is DeiT-Small [42] pretrained on ImageNet [37] with DINO framework [6]. The results verify Assumption 1 that the Top-1 similarity is significantly larger than others.

Experiments

Table 1. **Image Classification Results:** Experiments are conducted on natural images with different sampling ratios. We report the mean and std over three trials. Explanation of N/A results ("-") is in our supplementary materials.

Methods	CIFAR10			CIFAR100				ImageNet	
	0.5%	1%	2%	1%	2%	5%	10%	1%	5%
Random	77.3±2.6	82.2±1.9	88.9±0.4	14.9±1.9	24.3±2.0	50.8±3.4	69.3±0.7	45.1±0.8	64.3±0.3
FDS	64.5±1.5	73.2±1.2	81.4±0.7	8.1±0.6	12.8±0.3	16.9±1.4	52.3±1.9	26.7±0.6	55.5±0.1
K-Means	83.0±3.5	85.9±0.8	89.6±0.6	17.6±1.1	31.9±0.1	42.4±1.0	70.7±0.3	-	-
CoreSet [38]	-	81.6±0.3	88.4±0.2	-	30.6±0.4	48.3±0.5	62.9±0.6	-	61.7±0.2
VAAL [39]	-	80.9±0.5	88.8±0.3	-	24.6±1.1	46.4±0.8	70.1±0.4	-	64.0±0.3
LearnLoss [48]	-	81.6±0.6	86.7±0.4	-	19.2±2.2	38.2±2.8	65.7±1.1	-	63.2±0.4
TA-VAAL [21]	-	82.6±0.4	88.7±0.2	-	34.7±0.7	46.4±1.1	66.8±0.5	-	64.3±0.2
ALFA-Mix [33]	-	83.4±0.3	89.6±0.2	-	35.3±0.8	50.4±0.9	69.9±0.6	-	64.5±0.2
ActiveFT (ours)	**85.0**±0.4	**88.2**±0.4	**90.1**±0.2	**26.1**±2.6	**40.7**±0.9	**54.6**±2.3	**71.0**±0.5	**50.1**±0.3	**65.3**±0.1

Experiments

Table 2. **Semantic Segmentation Results:** experiments are conducted on ADE20k with sampling ratios $5\%, 10\%$. Results are averaged over three trials.

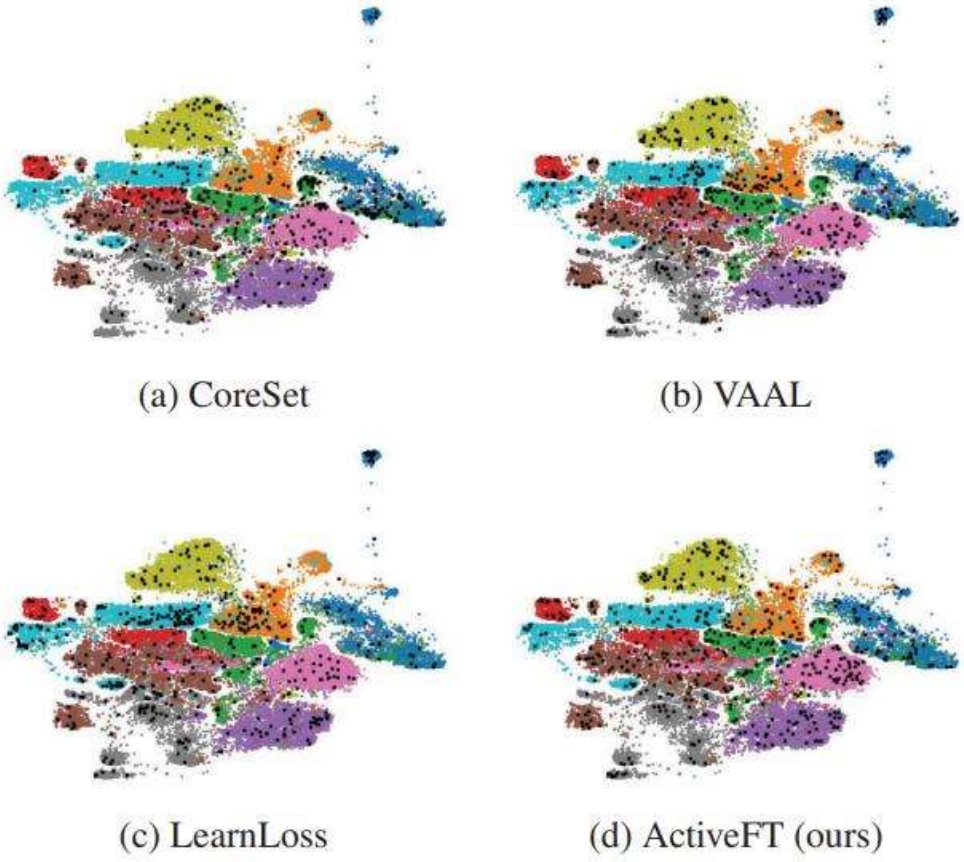
Sel. Ratio	Random	FDS	K-Means	ActiveFT (ours)
5%	14.54	6.74	13.62	**15.37**$\pm$0.11
10%	20.27	12.65	19.12	**21.60**$\pm$0.40

Experiments

Table 3. **Data Selection Efficiency:** We compare the time cost to select different percentages of samples from the CIFAR100 training set.

Sel. Ratio	K-Means	CoreSet	VAAL	LearnLoss	ours
2%	16.6s	1h57m	7h52m	20m	**12.6s**
5%	37.0s	7h44m	12h13m	1h37m	**21.9s**
10%	70.2s	20h38m	36h24m	9h09m	**37.3s**

Experiments



(a) CoreSet (b) VAAL

(c) LearnLoss (d) ActiveFT (ours)

Figure 4. **tSNE Embeddings of CIFAR10:** We visualize the embedding of selected samples using different algorithms. Different colors denote categories, and the black dots are the 1% samples selected by our method.

Experiments

Table 4. **Generality on Pretraining Frameworks and Model Architectures:** We examine the performance of ActiveFT on different pretraining frameworks and models on CIFAR-10.

(a) **Performance on DeiT-Small Pretrained with iBOT**

Methods	0.5%	1%	2%
Random	81.7	83.0	89.8
CoreSet [38]	-	82.8	89.2
LearnLoss [48]	-	83.6	89.2
VAAL [39]	-	85.1	89.3
ActiveFT (ours)	**87.6**±0.8	**88.3**±0.2	**90.9**±0.2

(b) **Performance on ResNet-50 Pretrained with DINO**

Methods	0.5%	1%	2%
Random	64.8	76.2	83.7
CoreSet [38]	-	70.4	83.2
LearnLoss [48]	-	71.7	81.3
VAAL [39]	-	75.0	83.3
ActiveFT (ours)	**68.5** ±0.4	**78.6** ±0.7	**84.9** ±0.3

Experiments

Table 5. **Ablation Study:** We examine the effect of two modules in our method. Experiments are conducted on CIFAR100 with pretrained DeiT-Small model.

(a) c_i **Update Manner**

Ratio	**No-Update**	**Update**
2%	20.6	**40.7**
5%	52.8	**54.6**

(b) **Regularization Design**

Ratio	**S1**	**S2**	**ours**
2%	33.1	26.8	**40.7**
5%	51.5	46.9	**54.6**

Experiments

Table 6. **Effect of Temperatures:** We try different temperatures in our method. Experiments are conducted on CIFAR10 with pre-trained DeiT-Small model.

Ratio	$\tau = 0.04$	$\tau = 0.07$	$\tau = 0.2$	$\tau = 0.5$
0.5%	**85.6**	85.0	84.1	83.5
1%	87.4	**88.2**	85.3	86.1
2%	**90.3**	90.1	89.6	89.0

Thanks