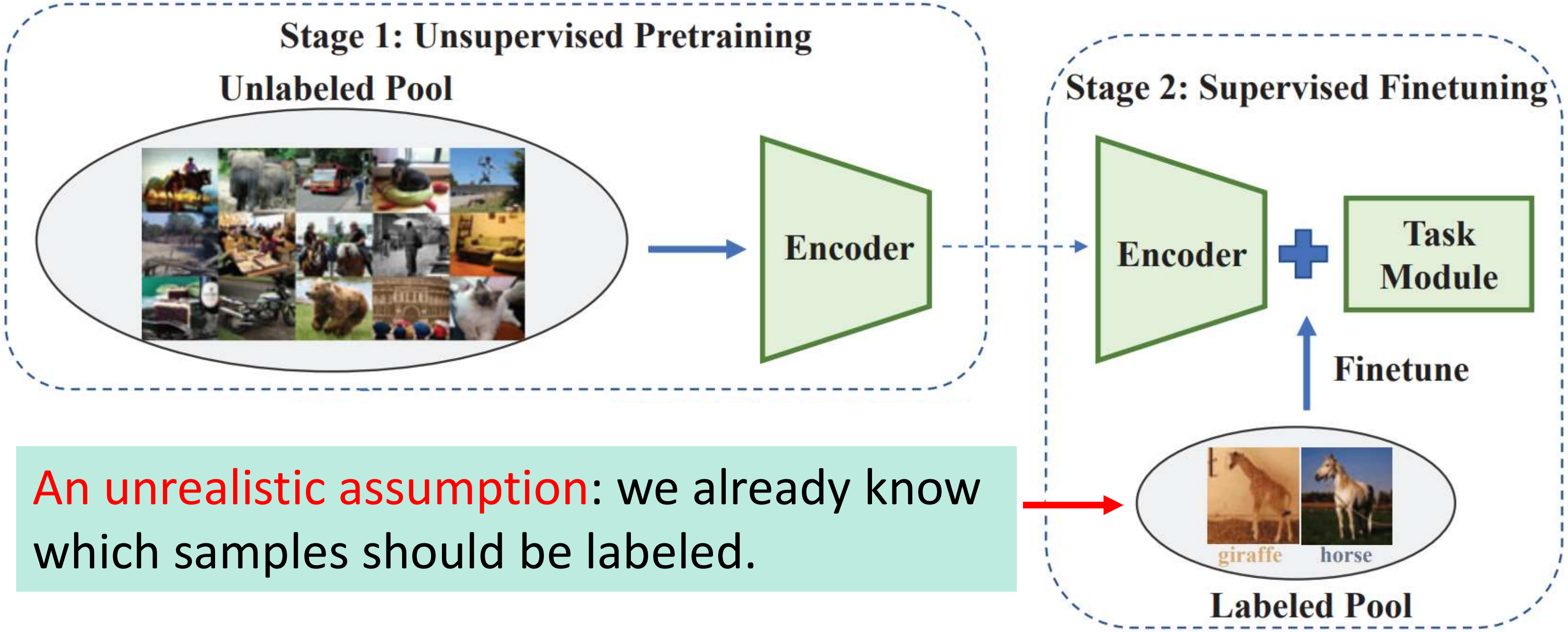
南京航空航天大學
Nanjing University of Aeronautics and Astronautics

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

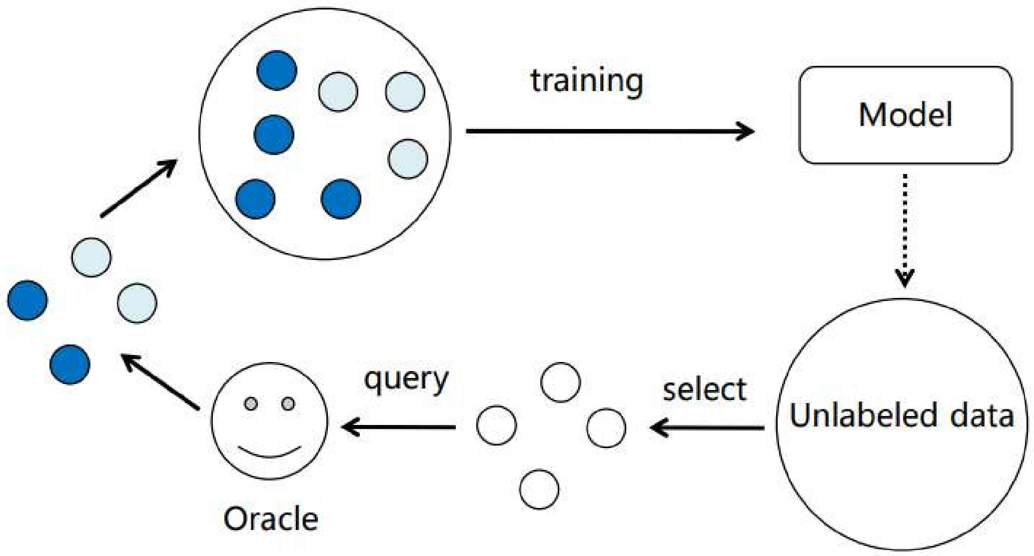
CVPR 2023

Pretraining-Finetuning Paradigm



An unrealistic assumption: we already know which samples should be labeled.

Active Learning



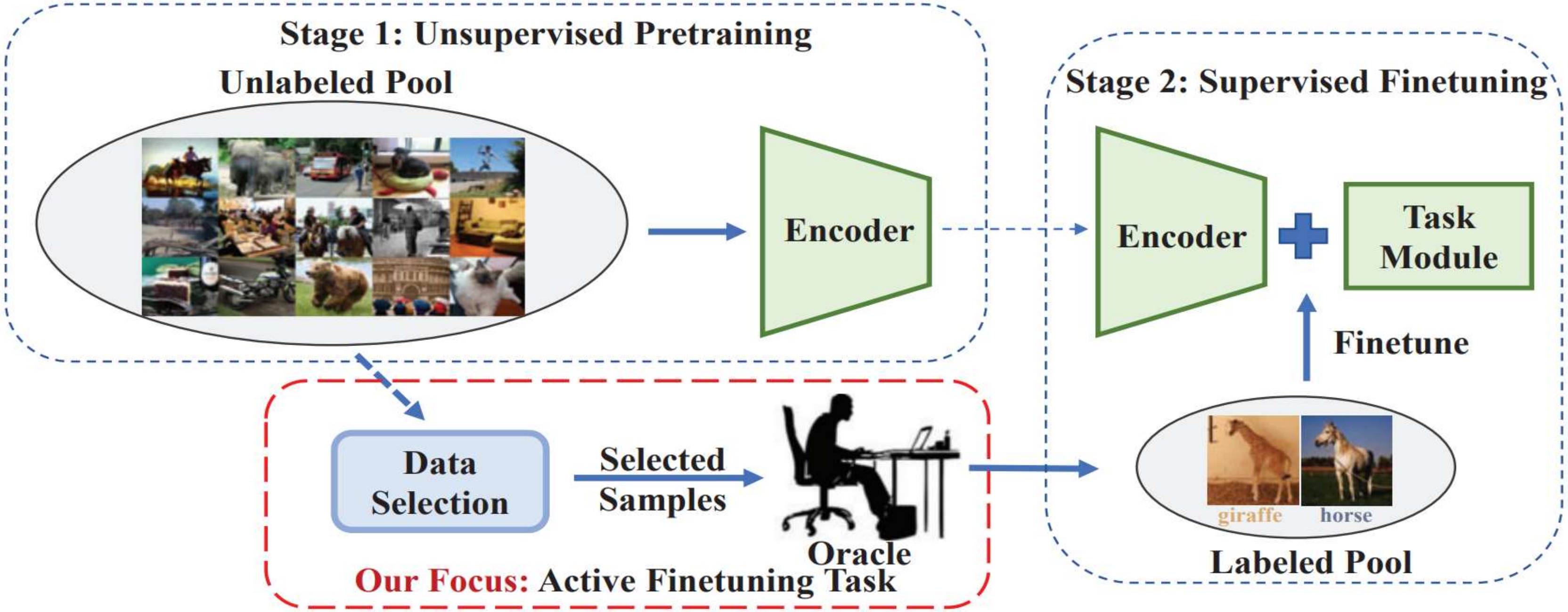
Goal: query less for more.

- Uncertainty-based sampling
  - Least-confidence
  - Margin
  - Entropy
  - MC-dropout
- Diversity-based sampling
  - CoreSet

✓ Challenges: more limited annotation proportion (e.g., <10%).

- *From-scratch training* does not fit the pretraining-finetuning paradigm.
- *Batch-selection strategy* leads to harmful bias inside the selection process.

Active Finetuning



✓ Differences:

- The size of the sampled subset is relatively small.
- Access to the pretrained model is available.
- No access to any labels before data selection.
- Selected samples are directly applied to the finetuning of the pretrained model.

Data Selection with Parametric Model

$\mathcal{X}$: is the data space $\qquad \mathbb{R}^C$: is the normalized high dimensional feature space

$f(\cdot; w_0): \mathcal{X} \rightarrow \mathbb{R}^C$: is a DNN model with pretrained weight w_0 is given

$\mathcal{P}^u = \{x_i\}_{i \in [N]} \sim p_u$: is a large unlabeled data pool

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

$\mathcal{P}^u_{\mathcal{S}} = \{x_{s_j}\}_{j \in [B]}$: is the target subset $\longrightarrow$ $\{y_{s_j}\}_{j \in [B]} \in \mathcal{Y}$: are the corresponding labels

$\mathcal{P}^l_{\mathcal{S}} = \{x_{s_j}, y_{s_j}\}_{j \in [B]}$: is the labeled data pool for supervised finetuning $\longrightarrow$ $w_{\mathcal{S}}$

The goal of active finetuning is to find $\mathcal{S}_{opt}$ minimizing the expectation model error.

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

Data Selection with Parametric Model

- Samples are selected under the guidance two basic intuitions:
  - Bringing close the distributions between the selected subset $\mathcal{P}_S^u$ and the original pool $\mathcal{P}^u \sim p_u$
  - Maintaining the diversity of $\mathcal{P}_S^u$

$p_u(x) \rightarrow p_{f_u}(x) \longrightarrow f_i = f(x_i; w_0)$: is the normalized feature $\longrightarrow \mathcal{F}^u = \{f_i\}_{i \in [N]}$

- Our goal is to find the optimal selection strategy $\mathcal{S}$ as:

$$\mathcal{S}_{opt} = \arg\min_{\mathcal{S}} D(p_{f_u}, p_{f_S}) - \lambda \mathcal{R}(\mathcal{F}_S^u) \longrightarrow \mathcal{S} \text{ is discrete selection strategy}$$

$$p_{f_S}(x) \rightarrow p_{\theta_S}(x),\ f_S = \{f_{S_i}\}_{i \in [B]},\ \theta_S = \left\{\theta_S^j\right\}_{j \in [B]}$$

- The goal is written as:

$$\theta_{\mathcal{S},opt} = \arg\min_{\theta_S} D(p_{f_u}, p_{\theta_S}) - \lambda \mathcal{R}(\theta_{\mathcal{S}})\ s.t.\ ||\theta_{\mathcal{S}}^j||_2 = 1$$

Parametric Model Optimization

- Consider the parametric model p_{θ_S} as a mixture model with B components:

$$\boxed{p_{\theta_{\mathcal{S}}}(\mathbf{f}) = \sum_{j=1}^{B} \phi_j p(\mathbf{f}|\theta_{\mathcal{S}}^j)} \qquad \sum_{j=1}^{B} \phi_j = 1 \qquad p(\mathbf{f}|\theta_{\mathcal{S}}^j) = \frac{\exp(sim(\mathbf{f}, \theta_{\mathcal{S}}^j)/\tau)}{Z_j}$$

- For each $f_i \in \mathcal{F}^u$, there exists a $\theta_S^{c_i}$ most similar (and closest) to f_i:

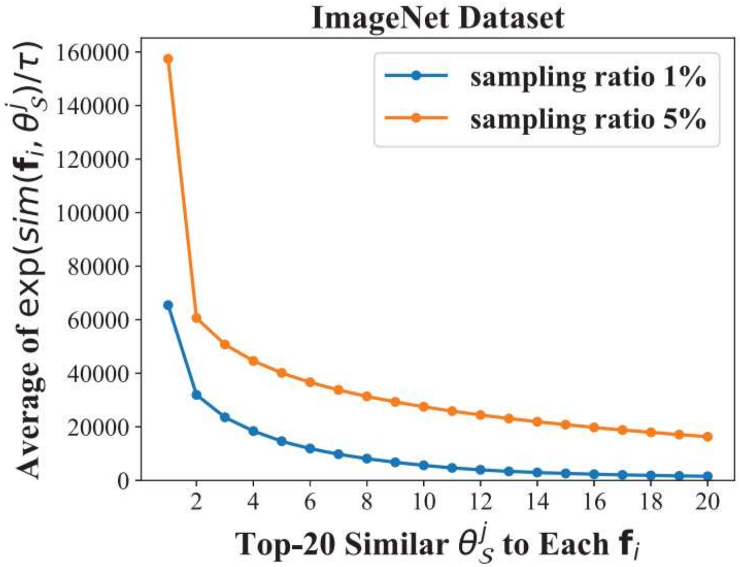
$$c_i = \arg\max_{j\in[B]} sim(\mathbf{f}_i, \theta_{\mathcal{S}}^j)$$

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

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

$$\Downarrow$$

$$p(\mathbf{f}_i|\theta_{\mathcal{S}}^{c_i}) \gg p(\mathbf{f}_i|\theta_{\mathcal{S}}^j), j \neq c_i, j \in [B]$$



Parametric Model Optimization

- The parametric model p_{θ_S} for $f_i \in \mathcal{F}^u$ can be approximated as:

$$p_{\theta_S}(\mathbf{f}) = \sum_{j=1}^{B} \phi_j p(\mathbf{f}|\theta_S^j) \longrightarrow p_{\theta_S}(\mathbf{f}_i) \approx \phi_{c_i} p(\mathbf{f}_i|\theta_S^{c_i}) = \frac{\exp(sim(\mathbf{f}_i, \theta_S^{c_i})/\tau)}{Z_{c_i}/\phi_{c_i}} = \frac{\exp(sim(\mathbf{f}_i, \theta_S^{c_i})/\tau)}{\tilde{Z}_{c_i}}$$

$$p_{\theta_S}(\mathbf{f}_i) \propto \exp(sim(\mathbf{f}_i, \theta_S^{c_i})/\tau)$$

- The two distributions p_{f_u}, p_{θ_S} can be brought close by minimizing the KL-divergence :

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

- Therefore, the first term in $\theta_{S,\,opt}$ is derived as:

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

An extra regularization term is needed to ensure the diversity of selected subset.

Parametric Model Optimization

- The second term in $\theta_{S,\,opt}$ is derived as :

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

- At this point, we can solve $\theta_{S,\,opt}$ by optimizing the following loss function continuously:

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

- Find feature $\{f_{S_i}\}_{i\in[B]}$ with the highest similarity to θ_S^j :

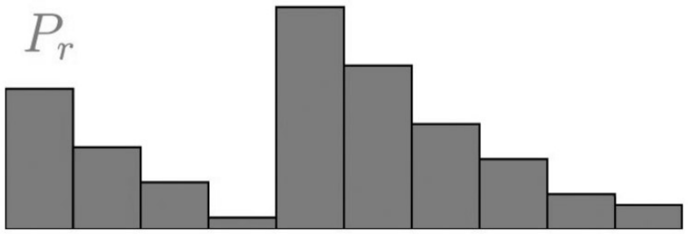
$$\mathbf{f}_{s_j} = \arg \max_{\mathbf{f}_k \in \mathcal{F}^u} sim(\mathbf{f}_k, \theta_{\mathcal{S}}^{j})$$

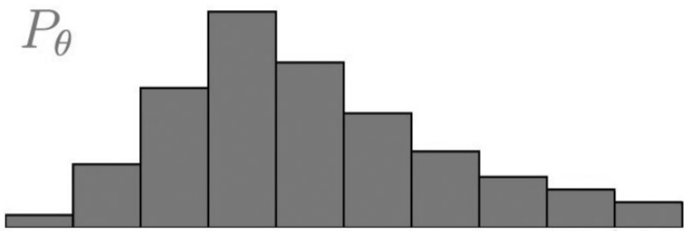
$\{\mathbf{x}_{s_j}\}_{j\in[B]}$

Relation to Earth Mover’s Distance

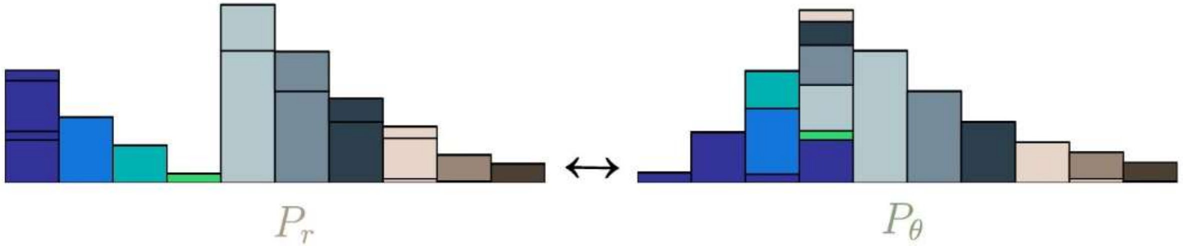
推土机距离

- 如果我们将分布想象为两个有一定存土量的土堆，每个土堆维度为 N，那么 EMD 就是将一个土堆转换为另一个土堆所需的最小总工作量。工作量的定义是单位泥土 的总量乘以它移动的距离。两个离散的土堆分布记作 P_r 和 P_θ ，以以下两个任意的分布为例。

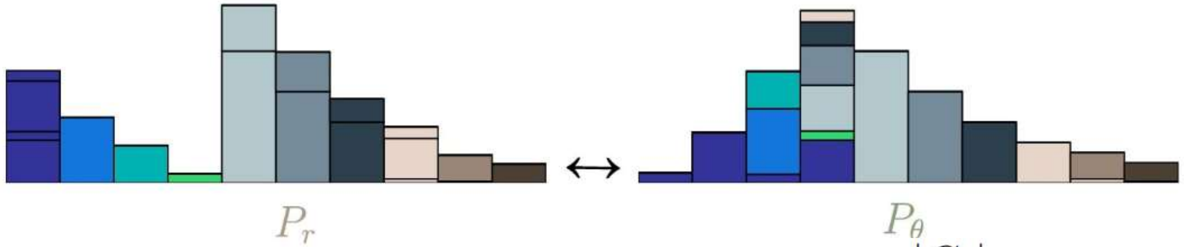




- **直观理解**，土堆 P_r 和 P_θ其中的单个柱条记为，$P_r(x)$ 和 $P_\theta(y)$，每一个 $P_r(x)$就是当前 x 位置土的存量， $P_\theta(x)$ 指的是最终 x 位置要存放的土量。
    - 如果 $P_r(x) > P_\theta(x)$：就要将 x 处多余的一部分$P_r(x) - P_\theta(x)$土方搬运到别处；
    - 如果 $P_r(x) < P_\theta(x)$：就要从其他处搬一部分将部分到 x 处，使得 x 处的土方存量为$P_r(x)$。



Relation to Earth Mover's Distance



$$p_{f_u}(\mathbf{f}_i) = \frac{1}{N}, \mathbf{f}_i \in \mathcal{F}^u \quad \longleftrightarrow \quad p_{f_S}(\mathbf{f}_{s_j}) = \frac{|C_j|}{N}, C_j = \{\mathbf{f}_i | c_i = j\}, \mathbf{f}_{s_j} \in \mathcal{F}_S^u$$

- The EMD between p_{f_u} , p_{f_S} is written as :

$$EMD(p_{f_u}, p_{f_S}) = \inf_{\gamma \in \Pi(p_{f_u}, p_{f_S})} \mathop{E}_{(\mathbf{f}_i, \mathbf{f}_{s_j}) \sim \gamma} [||\mathbf{f}_i - \mathbf{f}_{s_j}||_2]$$

$$-sim(\mathbf{f}_i, \theta_S^{c_i})$$

- It is intuitive to come up with the infimum, i.e. each $f_i \sim p_{f_u}$ transports to their closest $f_{s_j} \sim p_{f_S}$:

$$\gamma_{f_u, f_S}(\mathbf{f}_i, \mathbf{f}_{s_j}) = \begin{cases} \frac{1}{N} & \mathbf{f}_i \in \mathcal{F}^u, \mathbf{f}_{s_j} \in \mathcal{F}_S^u, c_i = j \\ 0 & otherwise \end{cases}$$

$$EMD(p_{f_u}, p_{f_S}) = \mathop{E}_{(\mathbf{f}_i, \mathbf{f}_{c_i}) \sim \gamma} [||\mathbf{f}_i - \mathbf{f}_{s_{c_i}}||_2]$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left[\sqrt{2 - 2 \boxed{sim(\mathbf{f}_i, \mathbf{f}_{s_{c_i}})}} \right]$$

Method: ActiveFT

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

Experiments

南京航空航天大学
Nanjing University of Aeronautics and Astronautics

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.

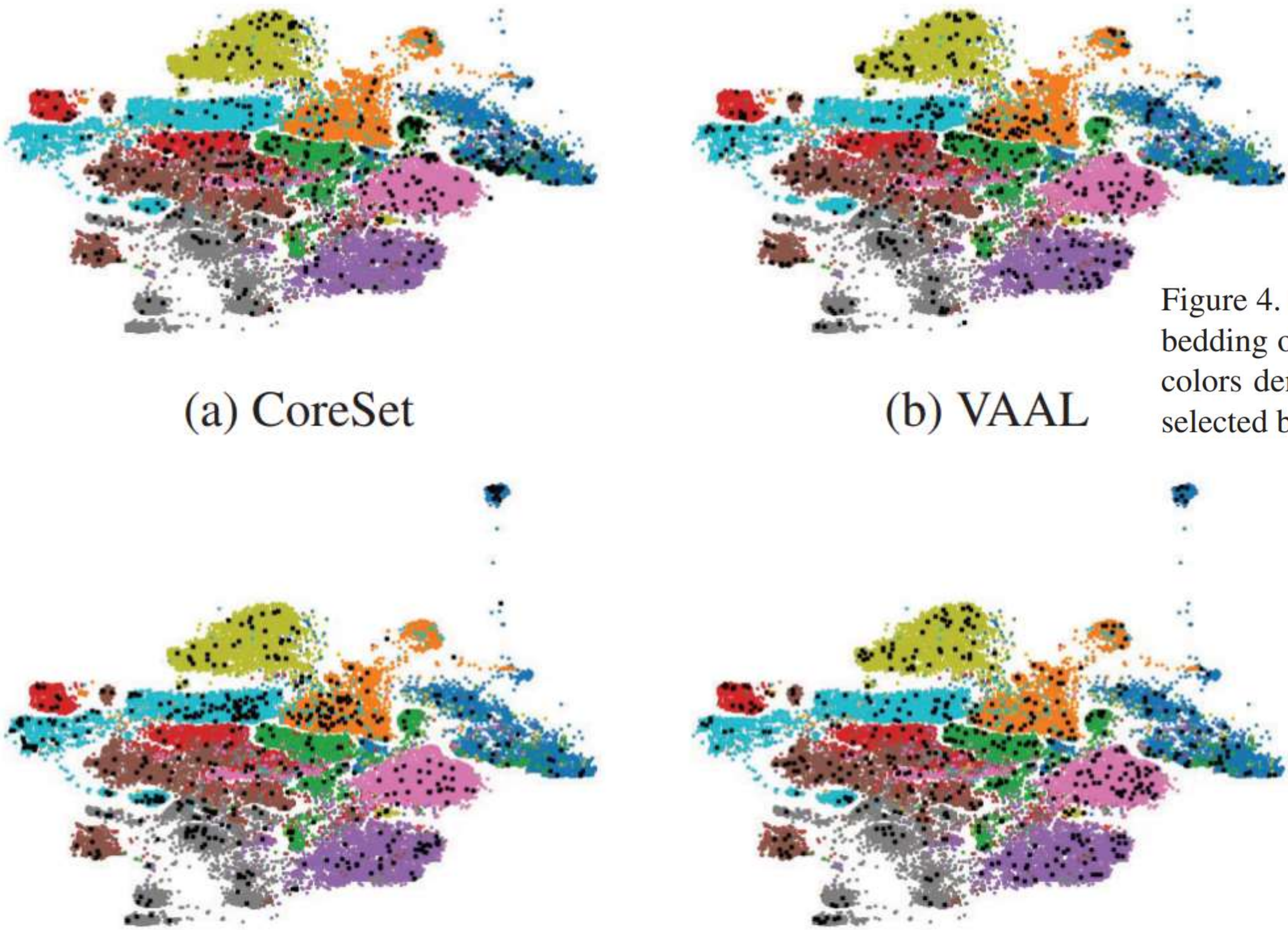
Sel. Ratio	Random	FDS	K-Means	ActiveFT (ours)
5%	14.54	6.74	13.62	**15.37**±0.11
10%	20.27	12.65	19.12	**21.60**±0.40

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**

$$L = -\underset{\mathbf{f}_i \in \mathcal{F}^u}{E}\left[\log \frac{\exp(\mathbf{f}_i^T \theta_S^{c_i}/\tau)}{\sum_{k\in[N]} \exp(sim(\mathbf{f}_k^T \theta_S^{c_i}/\tau)}\right]$$

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