Towards Cost-Effective Learning: A Synergy of Semi-Supervised and Active Learning

Supplementary Material

Detailed Experimental Settings(Appendix A)

We conducted experiments on two commonly used datasets, CIFAR-100 and TinyImagenet. We conducted experiments on two commonly used datasets, CIFAR-100 and Tiny-ImageNet, with a total of four different settings: CIFAR-100 #100, CIFAR-100 #200, Tiny-ImageNet #200, and Tiny-ImageNet #400. CIFAR-100 contains 100 classes, while Tiny-ImageNet contains 200 classes. CIFAR-100 #100 and Tiny-ImageNet #200 indicate that one sample was selected from each class, whereas CIFAR-100 #200 and Tiny-ImageNet #400 indicate that two samples were selected from each class.

Our code is built upon the open-source semi-supervised learning framework USB[1] and adopts its defined strong and weak augmentation methods. Detailed experimental settings and the adopted data augmentation methods are shown in the Table below.

Experiment Settings	Values
Model	vit-tiny-patch32
Optimizer	AdamW
LR	0.0005
Layer Decay	0.5
Momentum	0.9
Weight Decay	0.0005
Batch Size	64
Iteration Number	20480
Weak Augmentation	RandomCrop RandomHorizontalFlip
Strong Augmentation	RandomCropInterpolation RandomHorizontalFlip RandAug (AutoContrast, Brightness, Color, Contrast, Equalize, Identity, Posterize, Rotate, Sharpness, ShearX, ShearY, Solarize, TranslateX, TranslateY)

Table 3. The list of experimental settings.

Theoretical Proof(Appendix B)

In Section 3.3 of the original paper, we conducted a theoretical analysis and here elaborate on the derivation process from Equation 5 to Equation 10 in detail.

1. From Equation 4 to Equation 5.

Equation 4:

$$|\underbrace{\frac{1}{|D^*|}\sum_{(x,y)\in D^*}\ell(x,y;f_D)-\frac{1}{|D|}\sum_{(x,y)\in D}\ell(x,y;f_D)}_{Semi-supervised\ Error}|. \tag{20}$$

Equation 5:

$$\sum_{(x,y)\in U^*}\ell(x,y;f_D)-\sum_{(x,\hat{y})\in U}\ell(x,\hat{y};f_D). \tag{21}$$

Proof. Equation 5 represents the simplified semi-supervised error , which is derived from the last line of Equation 4.

$$\begin{aligned}
&|\sum_{(x,y)\in D^*}\ell(x,y;f_D)-\sum_{(x,y)\in D}\ell(x,\hat{y};f_D)|\\
=&|\{\sum_{(x,y)\in D^*/L}\ell(x,y;f_D)+\sum_{(x,y)\in L}\ell(x,y;f_D)\}-\\
&\{\sum_{(x,\hat{y})\in D/L}\ell(x,\hat{y};f_D)+\sum_{(x,y)\in L}\ell(x,y;f_D)\}|\\
=&|\sum_{(x,y)\in U^*}\ell(x,y;f_D)-\sum_{(x,\hat{y})\in U}\ell(x,\hat{y};f_D)|.
\end{aligned} \tag{22}$$

Here, D/L denotes the dataset obtained by removing L from D.

2. From Equation 5 to Equation 6.

Equation 5:

$$\sum_{(x,y)\in U^*}\ell(x,y;f_D)-\sum_{(x,\hat{y})\in U}\ell(x,\hat{y};f_D). \tag{23}$$

Equation 6:

$$\sum_{(x,y)\in U^*}\sum_{k\in\mathcal{Y}}\mathcal{P}(x_w,k;D)\cdot\mathcal{M}(x_s,k;D). \tag{24}$$

Proof. We introduce the cross-entropy loss function and adopt both strong augmentations $\mathcal{A}$ and weak augmentations α, denoting $\mathcal{A}(x)$ and $\alpha(x)$ as x_s and x_w. Conse-

[1] https://www.github.com/microsoft/Semi-supervised-learning

quently, we can derive:

$$
\begin{aligned}
&|\sum_{(x,y)\in U^*}\ell(x,y;f_D)-\sum_{(x,\hat{y})\in U}\ell(x,\hat{y};f_D)| \\
&=\sum_{(x,y)\in U^*}|\ell(f_D(x_s),y)-\ell(f_D(x_s),h_D(x_w))| \\
&\overset{(a)}{=}\sum_{(x,y)\in U^*}\sum_{k\in\mathcal{Y}}\{|\mathbb{I}(y=k)-\mathbb{I}(h_D(x_w)=k)|\cdot \\
&\qquad\qquad(\log\sum_{c\in\mathcal{Y}}\exp f_D^{(c)}(x_s)-f_D^{(k)}(x_s))\} \\
&\approx\sum_{(x,y)\in U^*}\sum_{k\in\mathcal{Y}}\{|\mathbb{I}(y=k)-\mathbb{I}(h_D(x_w)=k)|\cdot \\
&\qquad\qquad(f_D^{(max)}(x_s)-f_D^{(k)}(x_s))\} \\
&=\sum_{(x,y)\in U^*}\sum_{k\in\mathcal{Y}}\mathcal{P}(x_w,k;D)\cdot\mathcal{M}(x_s,k;D).
\end{aligned}
\tag{25}
$$

Here $\mathcal{P}(x_w,k;D)=|\mathbb{I}(y=k)-\mathbb{I}(h_D(x_w)=k)|$ is termed the *pseudo error*, and $\mathcal{M}(x_s,k;D)=f_D^{(max)}(x_s)-f_D^{(k)}(x_s)$ is referred to as the *margin error*. The notation (a) indicates the introduction of the cross-entropy loss function, which can be expressed as:

$$
\begin{aligned}
&\ell(f_D(x_s),y) \\
&=-\sum_{k\in\mathcal{Y}}y_k\log\frac{\exp(f_D^{(k)}(x_s))}{\sum_{c\in\mathcal{Y}}\exp(f_D^{(c)}(x_s))} \\
&=\sum_{k\in\mathcal{Y}}\mathbb{I}(y=k)(\log\sum_{c\in\mathcal{Y}}f_D^{(c)}(x_s)-f_D^{(k)}(x_s))
\end{aligned}
\tag{26}
$$

3. Equation 8.

Equation 8:

$$
\begin{aligned}
&\mathbb{E}_{k\in\mathcal{Y}}|\mathbb{I}(y_j^u=k)-\mathbb{I}(h_D(\alpha(x_j^u))=k)| \\
&\propto(1-\text{sim}(z_j^u,z_*^l)).
\end{aligned}
\tag{27}
$$

Proof. Next, we will conduct a detailed derivation of the "pseudo error" term $\mathcal{P}(x_w,k;D)$.

$$
\begin{aligned}
&\mathbb{E}_{k\in\mathcal{Y}}|\mathbb{I}(y_j^u=k)-\mathbb{I}(h_D(\alpha(x_j^u))=k)| \\
&=\sum_{k\in\mathcal{Y}}P_{y_i^u\sim\eta_k(z_j^u)}(y_i^u=k)\mathbb{I}(h_D(\alpha(x_j^u))\neq y_i^u) \\
&\overset{(a)}{\leq}\sum_{k\in\mathcal{Y}}P_{y_i^u\sim\eta_k(z_*^l)}(y_i^u=k)\mathbb{I}(h_D(\alpha(x_j^u))\neq y_i^u) \\
&\quad+\sum_{k\in\mathcal{Y}}|\eta_k(z_j^u)-\eta_k(z_*^l)|\mathbb{I}(h_D(\alpha(x_j^u))\neq y_i^u).
\end{aligned}
\tag{28}
$$

The notation (a) in the above equation refers to the inference presented in the study [1]. And this reasoning has also been adopted and utilized by other research studies[2, 3]. Assuming that the higher the feature similarity between samples, the greater the probability that they belong to the same category, we can derive two conclusions from the above equation based on this assumption:

$$
\begin{aligned}
&\sum_{k\in\mathcal{Y}}P_{y_i^u\sim\eta_k(z_*^l)}(y_i^u=k)\mathbb{I}(h_D(\alpha(x_j^u))\neq y_i^u) \\
&=\sum_{k\in\mathcal{Y}}P_{y_i^u\sim\eta_k(z_*^l)}(y_i^u=k)\mathbb{I}(h_D(\alpha(x_j^u))\neq y_*^l) \\
&=\mathcal{O}_1(1-\text{sim}(z_j^u,z_*^l)),
\end{aligned}
\tag{29}
$$

and

$$
\begin{aligned}
&\sum_{k\in\mathcal{Y}}|\eta_k(z_j^u)-\eta_k(z_*^l)|\mathbb{I}(h_D(\alpha(x_j^u))\neq y_i^u) \\
&\leq\sum_{k\in\mathcal{Y}}|\eta_k(z_j^u)-\eta_k(z_*^l)|=\mathcal{O}_2(1-\text{sim}(z_j^u,z_*^l)).
\end{aligned}
\tag{30}
$$

From this, we can infer that the "pseudo error" $\mathcal{P}(x_w,k;D)$ is negatively correlated with $\text{sim}(z_j^u,z_*^l)$, meaning that as the similarity increases, the "pseudo error" decreases. Therefore, $\mathcal{P}(x_w,k;D)\propto(1-\text{sim}(z_j^u,z_*^l))$.

4. Equation 10.

Equation 10:

$$
\begin{aligned}
&\sum_{k\in\mathcal{Y}}\mathcal{P}(x_w,k;D)\cdot\mathcal{M}(x_s,k;D) \\
&=f_D^{(\hat{y})}(x_s)-f_D^{(y)}(x_s).
\end{aligned}
\tag{31}
$$

Proof. In this section, we will conduct a detailed derivation of the "margin error" $\mathcal{M}(x_s,k;D)$:

$$
\sum_{k\in\mathcal{Y}}\mathcal{M}(x_s,k;D)=\sum_{k\in\mathcal{Y}}f_D^{(max)}(x_s)-f_D^{(k)}(x_s),
\tag{32}
$$

where $f_D^{(max)}(x_s)$ represents the network output for the predicted class under the strong augmentation $h_D(x_s)$. In the semi-supervised learning framework, as the training process progresses, the prediction results under strong and weak augmented views gradually converge to consistency ($h_D(x_s)=h_D(x_w)$). Consequently, the Equation 32 can be adjusted to:

$$
\sum_{k\in\mathcal{Y}}\mathcal{M}(x_s,k;D)=\sum_{k\in\mathcal{Y}}f_D^{(\hat{y})}(x_s)-f_D^{(k)}(x_s),
\tag{33}
$$

where $\hat{y}$ is the predicted category on weak augmentation ($h_D(x_w)$). The semi-supervised error can be formulated as:

$$
\begin{aligned}
&\sum_{k\in\mathcal{Y}}\mathcal{P}(x_w,k;D)\cdot\mathcal{M}(x_s,k;D) \\
&=\sum_{k\in\mathcal{Y}}|\mathbb{I}(y=k)-\mathbb{I}(h_D(x_w)=k)|(f_D^{(\hat{y})}(x_s)-f_D^{(k)}(x_s)).
\end{aligned}
\tag{34}
$$

In this scenario, we can divide it into two cases for handling: The first case is when $h_D(x_w)$ equals true label y, in which case $\mathbb{I}(y = k) - \mathbb{I}(h_D(x_w) = k) = 0$, and we need not consider the margin error anymore. The second case is when $h_D(x_w)$ is not equal to y, at which point Equation 34 can be simplified as:

$$\begin{aligned} &\sum_{k \in \mathcal{Y}} \mathcal{P}(x_w, k; D) \cdot \mathcal{M}(x_s, k; D) \\ =&f_D^{(max)}(x_s) - f_D^{(y)}(x_s) + f_D^{(max)}(x_s) - f_D^{(\hat{y})}(x_s) \\ =&f_D^{(\hat{y})}(x_s) - f_D^{(y)}(x_s). \end{aligned} \tag{35}$$

In other words, the margin error is closely related to the output boundary value between the predicted class and the true class under strong augmentation. To reduce the margin error across the entire dataset, we need to identify the sample x^* with the largest boundary values from the unlabeled dataset:

$$x^* = \arg\max_{x \in U} f_D^{(\hat{y})}(x_s) - f_D^{(y)}(x_s). \tag{36}$$

References

[1] Berlind, Christopher, and Ruth Urner. "Active nearest neighbors in changing environments." International conference on machine learning. 2015.

[2] Sener, Ozan, and Silvio Savarese. "Active Learning for Convolutional Neural Networks: A Core-Set Approach." International Conference on Learning Representations. 2018.

[3] Du, Pan, et al. "Contrastive active learning under class distribution mismatch." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.4 (2022): 4260-4273.