Supplementary: Safe-Student for Safe Deep Semi-Supervised Learning with Unseen-Class Unlabeled Data

1. The Proof of Theorem 1

Proof. The change of outer-level objective from iteration t to $t+1$ is:

$$\begin{aligned}
&\mathcal{L}\left(\theta_s^{t+1}\right)-\mathcal{L}\left(\theta_s^{t}\right)\\
&=\mathcal{L}\left(\theta_s^{t}-\eta_{\theta_s}\nabla_{\theta_s}\mathcal{L}\left(\theta_s^{t}\right)\right)-\mathcal{L}\left(\theta_s^{t}\right)\\
&\leq\left\langle\nabla_{\theta_s}\mathcal{L}\left(\theta_s^{t}\right),-\eta_{\theta_s}\nabla_{\theta_s}\mathcal{L}\left(\theta_s^{t}\right)\right\rangle+\frac{L}{2}\left\|-\eta_{\theta_s}\nabla_{\theta_s}\mathcal{L}\left(\theta_s^{t}\right)\right\|^2\\
&\leq-\rho\eta_{\theta_s}\rho+\frac{L}{2}\eta_{\theta_s}^2\rho^2=\left(\frac{L\eta_{\theta_s}}{2}-1\right)\eta_{\theta_s}\rho^2\leq 0
\end{aligned}\tag{1}$$

For simplicity, let $\eta_{\theta_s}=\eta$. Suppose $\theta_s^a,\theta_s^b\in\mathcal{W}$. Then we can know $\|\theta_s^a-\theta_s^b\|\leq\Gamma$, and $\|\nabla_{\theta_s}\mathcal{L}\left(\theta_s\right)\|\leq\rho$. Let arbitrary $\theta_s\in\mathcal{W}$, and we can know

$$\begin{aligned}
&\mathcal{L}\left(\theta_s^{t}\right)-\mathcal{L}\left(\theta_s\right)\leq\left\langle\nabla_{\theta_s}\mathcal{L}\left(\theta_s^{t}\right),\theta_s^{t}-\theta_s\right\rangle\\
&=\frac{1}{\eta}\left\langle\theta_s^{t}-\hat{\theta}_s^{t+1},\theta_s^{t}-\theta_s\right\rangle\\
&=\frac{1}{2\eta}\left(\left\|\theta_s^{t}-\theta_s\right\|^2-\left\|\hat{\theta}_s^{t+1}-\theta_s\right\|^2+\left\|\theta_s^{t}-\hat{\theta}_s^{t+1}\right\|^2\right)\\
&=\frac{1}{2\eta}\left(\left\|\theta_s^{t}-\theta_s\right\|^2-\left\|\hat{\theta}_s^{t+1}-\theta_s\right\|^2\right)+\frac{\eta}{2}\left\|\nabla_{\theta_s}\mathcal{L}\left(\theta_s^{t}\right)\right\|^2\\
&\leq\frac{1}{2\eta}\left(\left\|\theta_s^{t}-\theta_s\right\|^2-\left\|\theta_s^{t+1}-\theta_s\right\|^2\right)+\frac{\eta}{2}\left\|\nabla_{\theta_s}\mathcal{L}\left(\theta_s^{t}\right)\right\|^2
\end{aligned}\tag{2}$$

And because loss function $\mathcal{L}$ is Lipschitz-smooth, we can know

$$\mathcal{L}\left(\theta_s^{t}\right)-\mathcal{L}\left(\theta_s\right)\leq\frac{1}{2\eta}\left(\left\|\theta_s^{t}-\theta_s\right\|^2-\left\|\theta_s^{t+1}-\theta_s\right\|^2\right)+\frac{\eta}{2}\rho^2\tag{3}$$

By summing Eq. (3) from $t=1$ to T, we can get

$$\begin{aligned}
&\sum_{t=1}^{T}\mathcal{L}\left(\theta_s^{t}\right)-T\mathcal{L}\left(\theta_s\right)\\
&\leq\frac{1}{2\eta}\left(\left\|\theta_s^{1}-\theta_s\right\|^2-\left\|\theta_s^{t+1}-\theta_s\right\|^2\right)+\frac{\eta T}{2}\rho^2\\
&\leq\frac{1}{2\eta}\left\|\theta_s^{1}-\theta_s\right\|^2+\frac{\eta I}{2}\rho^2\leq\frac{1}{2\eta}\Gamma^2+\frac{\eta T}{2}\rho^2
\end{aligned}\tag{4}$$

According Jensen inequation, we can know

$$\begin{aligned}
\mathcal{L}\left(\bar{\theta}_s^{T}\right)-\mathcal{L}\left(\theta_s\right)&=\mathcal{L}\left(\frac{1}{T}\sum_{i=1}^{T}\theta_s^{t}\right)-\mathcal{L}\left(\theta_s\right)\\
&\leq\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\theta_s^{t}\right)-\mathcal{L}\left(\theta_s\right)\leq\frac{\Gamma^2}{2\eta T}+\frac{\eta\rho^2}{2}
\end{aligned}\tag{5}$$

Then, we can get

$$\mathcal{L}\left(\bar{\theta}_s^{T}\right)-\min_{\theta_s\in\mathcal{W}}\mathcal{L}\left(\theta_s\right)\leq\frac{\Gamma^2}{2\eta T}+\frac{\eta\rho^2}{2}=\frac{\rho\Gamma}{\sqrt{T}}=\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)\tag{6}$$

2. The Further Analysis of Energy-Discrepancy and Energy

Through the body, we can know

$$\begin{aligned}
\mathrm{E}(\boldsymbol{x})&=-\log\left[e^{f_y(\boldsymbol{x})}\cdot\left(e^{f_1(\boldsymbol{x})-f_y(\boldsymbol{x})}+\cdots+e^{f_K(\boldsymbol{x})-f_y(\boldsymbol{x})}\right)\right]\\
&=-\log e^{f_y(\boldsymbol{x})}-\log\left(e^{f_1(\boldsymbol{x})-f_y(\boldsymbol{x})}+\cdots+e^{f_K(\boldsymbol{x})-f_y(\boldsymbol{x})}\right)\\
&\approx-f_y(\boldsymbol{x}),
\end{aligned}\tag{7}$$

and

$$\mathrm{ED}(\boldsymbol{x})=\left|\mathrm{E}(\boldsymbol{x})-\mathrm{E}^{'}(\boldsymbol{x})\right|\approx f_y(\boldsymbol{x})-f_{y'}(\boldsymbol{x}).\tag{8}$$

The overall loss is defined as follows,

$$\mathcal{L}=\mathcal{L}_{CE}+\lambda_1\mathcal{L}_{CBE}+\lambda_2\mathcal{L}_{UCD},\tag{9}$$

where the first two items can be approximatively considered as cross-entropy loss operating on seen classes, and the last item is a constraint on unseen classes. Suppose $(\boldsymbol{x}_i, y)$ is a seen-class instance. Then, we can know

$$\begin{aligned}
\mathcal{L}_{CE}(\boldsymbol{x}_i)&=-\log\frac{e^{f_y(\boldsymbol{x}_i)}}{\sum_{j=1}^{K}e^{f_j(\boldsymbol{x}_i)}}\\
&=-\log\frac{e^{f_y(\boldsymbol{x}_i)}}{e^{f_1(\boldsymbol{x}_i)}+e^{f_2(\boldsymbol{x}_i)}+\cdots+e^{f_K(\boldsymbol{x}_i)}}\\
&=\log\left(e^{f_1(\boldsymbol{x}_i)-f_y(\boldsymbol{x}_i)}+\cdots+e^{f_K(\boldsymbol{x}_i)-f_y(\boldsymbol{x}_i)}\right).
\end{aligned}\tag{10}$$

Minimizing $\mathcal{L}_{CE}$ is equivalent to enlarging $f_y(\boldsymbol{x}_i)$ and reducing the other logits, which enlarges ED scores of seen-class instances. Suppose $\boldsymbol{x}_i$ is an unseen-class instance. Then, we can know

$$\mathcal{L}_{UCD} = \frac{1}{|D_{uc}|} \sum_{\boldsymbol{x}_i^u \in D_{uc}} \omega(\boldsymbol{x}_i^u) \mathrm{KL}(\mathcal{U}(y) \| \tilde{\eta} \circ \Phi((\mathrm{aug}(\boldsymbol{x}_i^u)))), \tag{11}$$

For simplicity, we leave out $\omega(\boldsymbol{x}_i^u)$. Then, we can know,

$$\begin{aligned}
&\mathcal{L}_{UCD}(\boldsymbol{x}_i) = \mathrm{KL}(\mathcal{U}(y) \| \tilde{\eta} \circ \Phi((\mathrm{aug}(\boldsymbol{x}_i)))) \\
&= \frac{1}{K} \times \log \frac{\frac{1}{K}}{\frac{e^{f_1(\boldsymbol{x}_i)}}{\sum_{j=1}^{K} e^{f_j(\boldsymbol{x}_i)}}} + \cdots + \frac{1}{K} \times \log \frac{\frac{1}{K}}{\frac{e^{f_K(\boldsymbol{x}_i)}}{\sum_{j=1}^{K} e^{f_j(\boldsymbol{x}_i)}}} \\
&= \frac{1}{K} \times \left(K \log \frac{1}{K} - \sum_{t=1}^{K} \log \frac{e^{f_t(\boldsymbol{x}_i)}}{\sum_{j=1}^{K} e^{f_j(\boldsymbol{x}_i)}} \right).
\end{aligned} \tag{12}$$

Minimizing $\mathcal{L}_{UCD}$ is equivalent to enlarging $\sum_{t=1}^{K} \log \frac{e^{f_t(\boldsymbol{x}_i)}}{\sum_{j=1}^{K} e^{f_j(\boldsymbol{x}_i)}}$. Besides, we know that $\sum_{t=1}^{K} \frac{e^{f_t(\boldsymbol{x}_i)}}{\sum_{j=1}^{K} e^{f_j(\boldsymbol{x}_i)}} = 1$. $\sum_{t=1}^{K} \log \frac{e^{f_t(\boldsymbol{x}_i)}}{\sum_{j=1}^{K} e^{f_j(\boldsymbol{x}_i)}}$ can obtain maximum when $\frac{e^{f_t(\boldsymbol{x}_i)}}{\sum_{j=1}^{K} e^{f_j(\boldsymbol{x}_i)}} = \frac{1}{K}$, and $f_{t_1}(\boldsymbol{x}_i) - f_{t_2}(\boldsymbol{x}_i) = 0$, where $\forall t_1, t_2 \in \{1, \cdots, K\}$, and $t_1 \neq t_2$. Minimizing $\mathcal{L}_{UCD}$ will reduce ED scores of unseen-class instances.

In summary, the final loss function Eq. (9) is aimed at enlarging ED of seen classes by the first two items and minimizing ED of unseen classes by the last item, which verifies the consistency between the optimization objective and our proposed scoring function ED. On the contrary, energy focuses on the maximum logit, which is inconsistent with optimization objective.

3. Algorithm

As shown in Algorithm 1, SAFE-STUDENT contains five modules: *teacher pre-training module* to obtain a teacher model that serves as a mentor for the student model, *seen and unseen classes identification module* to select reliable seen-class and unseen-class instances, *seen-class learning module* to achieve the seen-class classification, *unseen-class label distribution learning module* to mitigate the adverse effects of unseen classes, *iterative optimization strategy* helps the teacher model improve the identification of unseen classes and helps the student model improve the performance of seen-class classification.

4. Datasets

We evaluate SAFE-STUDENT on image classification datasets: MNIST [3], CIFAR-10 [2], CIFAR-100 [2] and TinyImageNet (a subset of ImageNet [1]), with different ratios of class mismatch.

Algorithm 1: SAFE-STUDENT.

input : Labeled data set D_L, unlabeled data set D_U, student model θ_s, teacher model θ_t, max iterations I, max epochs E
output: θ_s, θ_t

/*Teacher Pre-training:*/
for e = 1 **to** E **do**
  compute $\mathcal{L}_{CE}$ in Eq. (1)
  update $\theta_t \leftarrow$ SGD with loss $\mathcal{L}_{CE}$
/*Iterative Optimization:*/
for $i = 1$ **to** I **do**
  for epoch = 1 **to** E **do**
    /*Seen and Unseen Classes Identification:*/
    collect energy-discrepancy ED($\boldsymbol{x}_i^u$) for $\boldsymbol{x}_i^u$ by θ_t and Eq. (4)
    obtain D_{sc}, D_{uc} by energy-discrepancy
    /*Seen-Class Learning:*/
    obtain pseudo label $\tilde{y}_i^u$ and probability distribution $\tilde{p}_i^u$ of instances $\boldsymbol{x}_i^u$ by θ_t
    compute $\mathcal{L}_{CBE}$ by $\tilde{p}_i^u$, $\tilde{y}_i^u$, D_{sc}, θ_s, and Eq. (9)
    compute $\mathcal{L}_{CE}$ in Eq. (1)
    /*Unseen-Class Label Distribution Learning:*/
    obtain uniform distribution $\mathcal{U}(y)$
    compute $\mathcal{L}_{UCD}$ by $D_{uc}, \mathcal{U}(y), \theta_s$, and Eq. (10)
    obtain final loss $\mathcal{L}$ in Eq. (12)
    update $\theta_s \leftarrow$ SGD with loss $\mathcal{L}$
  update $\theta_t \leftarrow \theta_s$

- **MNIST** includes 60,000 training images and 10,000 testing images of size 28×28, which contains 10 classes from digit '0' to digit '9'. Concretely, we respectively select ten images from digit '0' to digit '5' to construct the labeled data set D_L, i.e., a total of 60 labeled data, and select 30,000 images in total from digit '0' to digit '9' as unlabeled data D_U. We adjust the ratio of unseen-class images in the unlabeled data to modulate class distribution mismatch. For example, when the extent of labeled/unlabeled class mismatch ratio is 0%, all unlabeled data come from digit '0' to digit '5'.
- **CIFAR-10** includes 60,000 training images and 10,000 testing images of size 32×32 which contains ten categories: "airline", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "ship", and "trunk". Our experiment carries out six-class classification tasks. We consider animal categories (birds, cats, deer, dogs, frogs, and horses) as seen classes and the rest as unseen classes. We select 400 images from each seen category to construct the labeled data set D_L, i.e., 2400 labeled instances. Meanwhile, 20,000 images in total are randomly selected

Table 1. ACC(%) for λ_1 with different values on MNIST.

λ_1	0.1	0.5	1	5	10
ACC	96.2±0.2	96.7±0.3	97.2±0.3	95.1±0.1	92.1±0.2

as the unlabeled data set D_U from all the ten categories. We adjust the ratio of unseen-class images in the unlabeled data to modulate class distribution mismatch.

- **CIFAR-100** includes 50,000 training images and 10,000 testing images of size 32×32 which contains 100 categories. We use the first half categories (1-50) as seen classes, and the remaining classes as unseen classes. We select 100 images from each seen category to construct the labeled data set D_L, i.e., 5000 labeled instances. Meanwhile, 20,000 images in total are randomly selected as the unlabeled data set D_U from all the 100 categories with different ratios of unseen classes.
- **TinyImageNet** contains 200 categories which includes 500 training images and 50 testing images in each category. We resize all images to 32×32. We use the first 100 categories as seen classes, and the remaining classes as unseen classes. We select 100 images from each seen category to construct the labeled data set D_L, i.e., 10000 labeled instances. Meanwhile, 40,000 images in total are randomly selected as the unlabeled data set D_U from all the 200 categories with different ratios of unseen classes.

4.1. The Result in CIFAR-100

Fig. 1 reports the averaged accuracy on CIFAR-100 over five runs with the different class mismatches. Firstly, our proposed SAFE-STUDENT significantly outperforms existing deep SSL methods. For example, when the mismatch ratio is 40%, our method achieves **68.3%** averaged accuracy, about **9.7%** higher than the supervised learning method, about **9.5%** higher than Pi-Model, about **8.3%** higher than Pseudo-Labeling, about **6.9%** higher than VAT, and about **5.1%** higher than CL. These results verify that our method achieves the best performance compared with the deep SSL methods and the safe deep SSL methods on the SDU problem. These results verify the effectiveness of SAFE-STUDENT.

4.2. Sensitivity of Hyperparameter

λ_1, λ_2 are coefficients in final optimization goal. Table. 1 shows the results of λ_1 under different values. Table. 2 shows the results of λ_2 under different values. These experiments are based on the 50% ratio of class mismatch on MNIST. The results show that it is important to choose the appropriate parameters for λ_1, λ_2.

4.3. Unseen Class Identification

Fig. 2 exhibits the distributions of the four scoring functions on the MNIST under the 0% ratio of class mismatch in D_U by SAFE-STUDENT. In-distribution denotes seen classes, and out-of-distribution denotes unseen classes. Fig. 3 exhibits the distributions of the four scoring functions on the MNIST under the 60% ratio of class mismatch in D_U by SAFE-STUDENT. Fig. 4 exhibits the distributions of the four scoring functions on the MNIST under the 60% ratio of class mismatch in D_U by DS^{3}L. Fig. 5 exhibits the distributions of the four scoring functions on the MNIST under the 60% ratio of class mismatch in D_U by probability estimation method. These results prove that our proposed energy-discrepancy outperforms the other scoring functions and owns the greater ability to identify unseen classes.

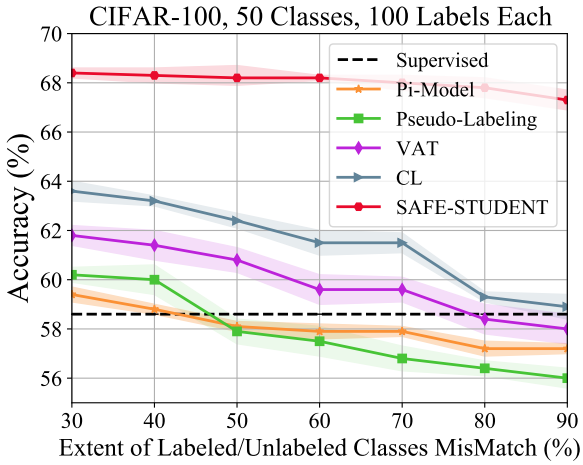

Figure 1. Seen-class classification accuracy (%) of SAFE-STUDENT and compared deep SSL methods on CIFAR-100 with different class mismatch ratios between labeled and unlabeled data. Shaded regions indicate standard deviation over five runs.

Table 2. ACC(%) for λ_2 with different values on MNIST.

λ_2	0.01	0.05	0.10	0.15	0.20	0.25
ACC	96.7±0.3	97.2±0.3	95.2±0.2	94.6±0.3	93.8±0.3	91.7±0.4

4.4. Energy and Energy-Discrepancy

Fig. 6 shows the accuracy on CIFAR-100 using energy or energy-discrepancy in the SUCI model. We can observe that energy-discrepancy more effectively identifies seen and unseen classes under different class distribution mismatch ratios.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[2] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

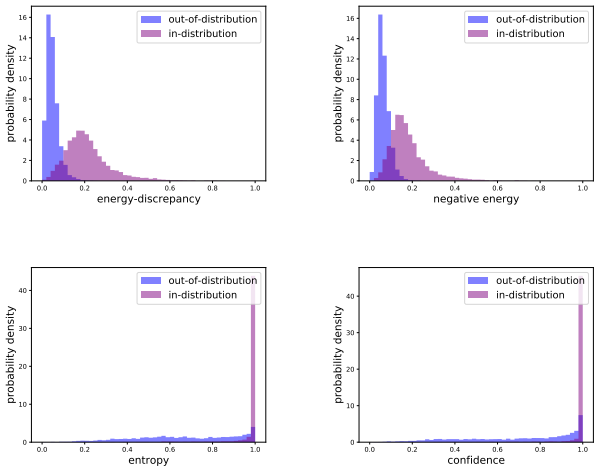

Figure 2. The distribution of four scoring functions under the 0% ratio of class mismatch in D_U by SAFE-STUDENT.

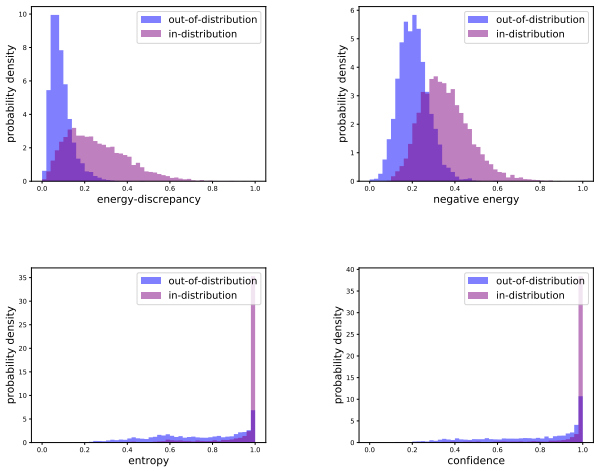

Figure 3. The distribution of four scoring functions under the 60% ratio of class mismatch in D_U by SAFE-STUDENT.

[3] Yann LeCun. The mnist database of handwritten digits. 1998. 2

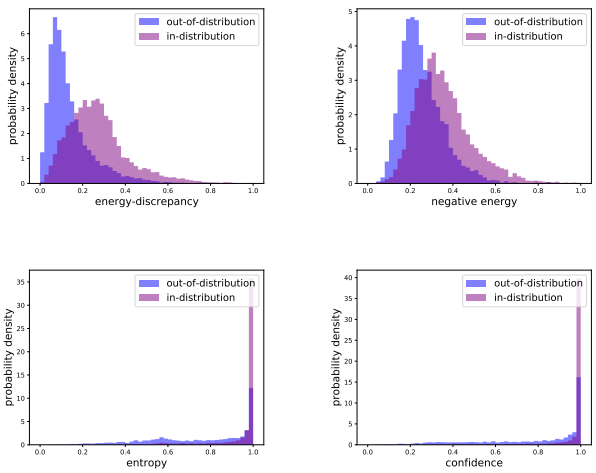

Figure 4. The distribution of the ED and negative energy under the 60% ratio of class mismatch in D_U by DS^{3}L.

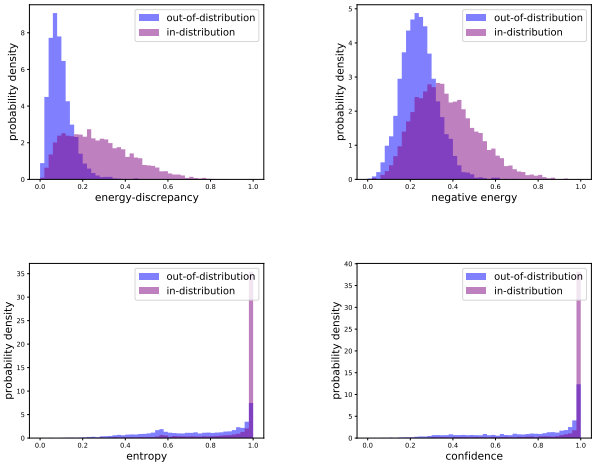

Figure 5. The distribution of four scoring functions by probability estimation method.

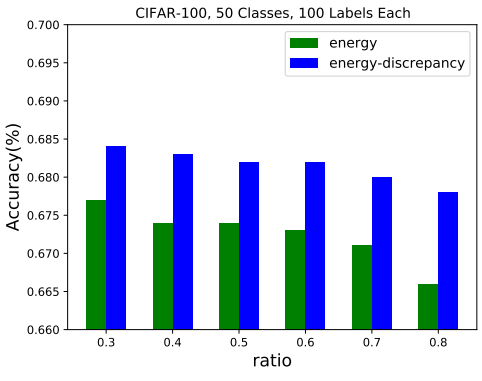

Figure 6. Seen-class classification accuracy (%) of SAFE-STUDENT on CIFAR-100 with energy or energy-discrepancy under different mismatch ratios between labeled and unlabeled data.