Safe Deep Semi-Supervised Learning for Unseen-Class Unlabeled Data

Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, Zhi-Hua Zhou

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

{guolz, zhangzy, jiangy, liyf, zhouzh}@lamda.nju.edu.cn

Unseen Class Unlabeled Data

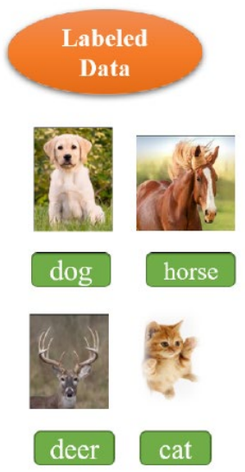
Semi-Supervised Learning

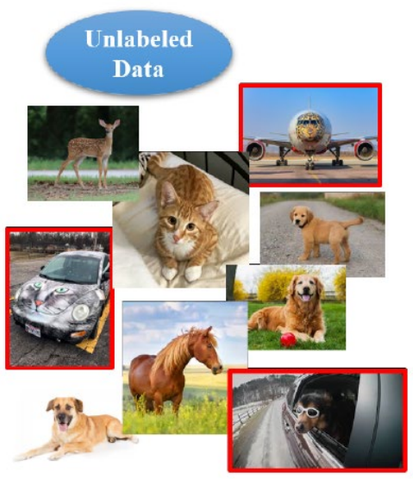
- **Learning with labeled and unlabeled data, e.g.,**
  - **Mean-Teacher** [Tarvainen &Valpola, 2017]
  - **VAT** [Miyato et al., 2018]
  - **Mix-Match** [Berthelot et al., 2019]

Assumption: labeled and unlabeled data are i.i.d sampled from the same distribution

However, in real application

Unlabeled data usually contains class not seen in labeled data and hurt SSL performance severely





Unseen class unlabeled data are bounded with red box

Safe Deep Semi-Supervised Learning (DS3L)

☐ **DSSL: Use all unlabeled equally**

$$\min_{\theta\in\Theta} \sum_{(x,y)\in\mathcal{D}_l} \ell(h(x;\theta),y) + \sum_{x\in\mathcal{D}_u} \Omega(x)$$

- $\Omega(x)$ refers to the regularization term, e.g., consistency regularization

☐ **DS3L: Use unlabeled data selectively**

$$\min_{\theta\in\Theta} \sum_{(x,y)\in\mathcal{D}_l} \ell(h(x;\theta),y) + \sum_{x\in\mathcal{D}_u} w(x;\alpha)\Omega(x)$$

- $w(x;\alpha)$ refers to the weight function

How to learn the weight function?

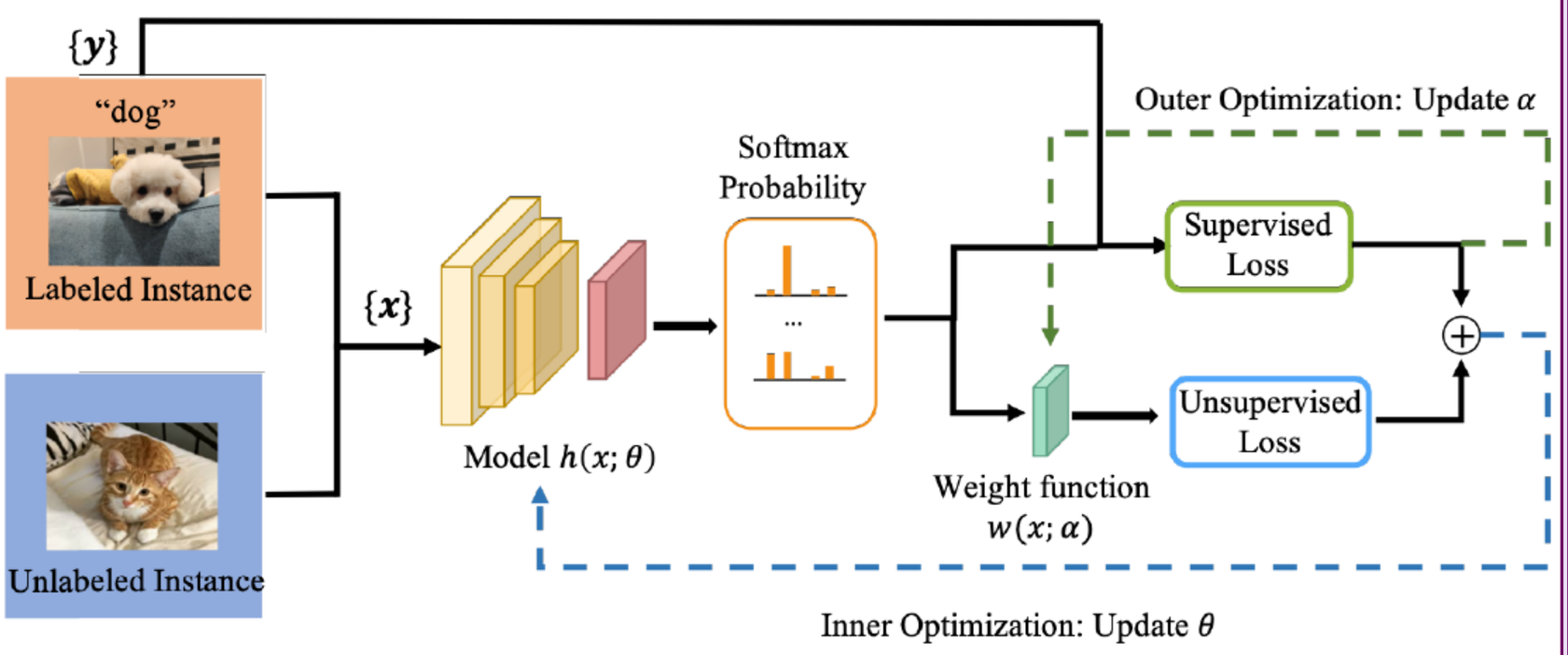
✓ **Idea: good weight realizes good generalization performance**

$$\min_{\alpha} \sum_{(x,y)\in\mathcal{D}_l} \ell(h(x;\hat{\theta}),y)$$

$$s.t.$$

$$\hat{\theta} = \operatorname{argmin}_{\theta\in\Theta} \sum_{(x,y)\in\mathcal{D}_l} \ell(h(x;\theta),y) + \sum_{x\in\mathcal{D}_u} w(x;\alpha)\Omega(x)$$

A novel bi-level formulation



Optimization

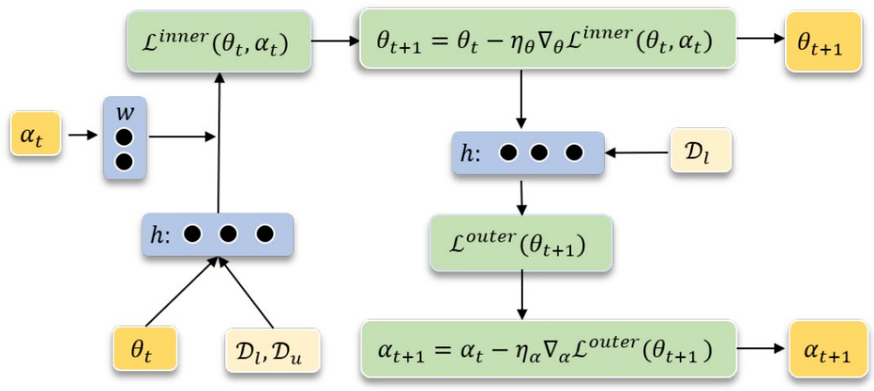
✓ **Adopting gradient-based methods, we can optimize θ and α iteratively**

$$\theta_{t+1} = \theta_t - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)$$

$$\alpha_{t+1} = \alpha_t - \eta_\alpha \nabla_\alpha \mathcal{L}^{outer}(\theta_{t+1})$$

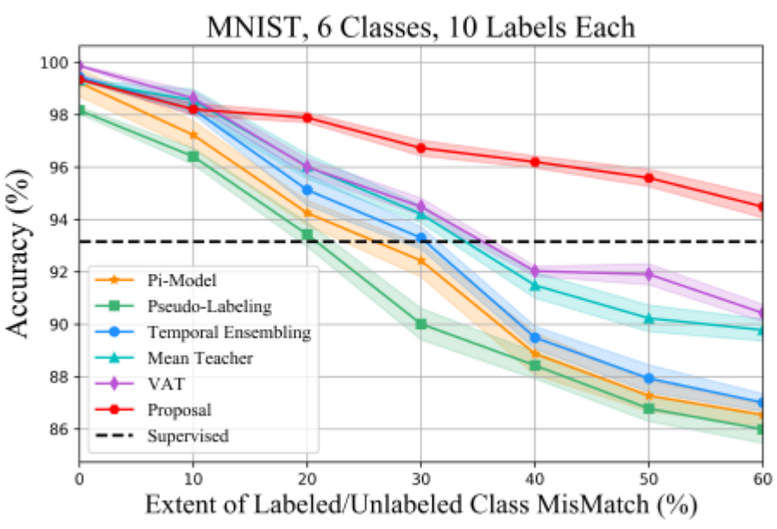
The bi-level gradient can be solved by the automatic differentiation techniques

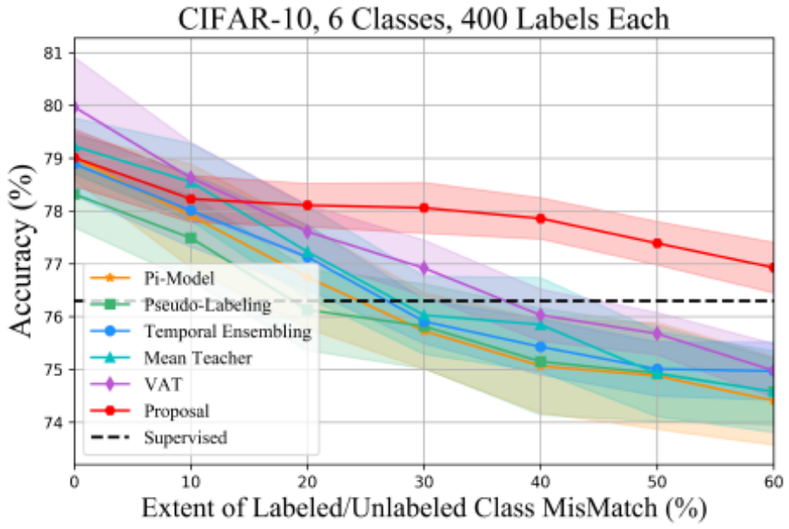
✓ **Main flowchart**



✓ **The optimization algorithm will converge and the convergence rate is $\mathcal{O}(\frac{C}{\sqrt{T}})$.**

Experiments





Our proposal DS3L achieves safe performance with varying extent of class mismatch, while other SSL methods suffer performance degradation problem

Analysis

➢ **Theorem 1: Empirical Safeness**

The empirical risk of $\hat{\theta}$ will **never worse than the supervised model** θ^{SL} that is learn from merely labeled data, i.e.,

$$\hat{R}(\hat{\theta}) \leq \hat{R}(\theta^{SL})$$

➢ **Theorem 2: Generalization**

DS3L approaches the optimal weight in order $\mathcal{O}\left(\sqrt{\frac{d \ln n}{n}}\right)$, i.e.,

$$R(\hat{\theta}(\alpha^*)) \leq R(\hat{\theta}(\hat{\alpha})) + \frac{(3\lambda + \sqrt{4d\ln(n) + 8\ln(2/\delta)})}{\sqrt{n}}$$

Take-Home Message

- **Unseen-class unlabeled data will hurt SSL performance**
- **We proposed a bi-level based reweight mechanism DS3L to use unlabeled data selectively**
- **The effectiveness can be demonstrated theoretically and empirically**

waitting for your feedback

Mail