NEURAL INFORMATION PROCESSING SYSTEMS

Reinforcement Learning Guided Semi-Supervised Learning

Marzi Heidari[1], Hanping Zhang[1], Yuhong Guo[1,2]

[1]Carleton University, Ottawa, Canada
[2]CIFAR AI Chair, Amii, Canada

Carleton University

Introduction

- **Semi-Supervised Learning (SSL):** Uses a small labeled dataset with a large unlabeled pool.
- **Limitations**: Heuristics or predefined rules for pseudo-labeling methods are often suboptimal.
- **Challenge**: How can we better leverage unlabeled data to guide the learning process?

Reinforcement Learning Guided Semi-Supervised Learning (RLGSSL)

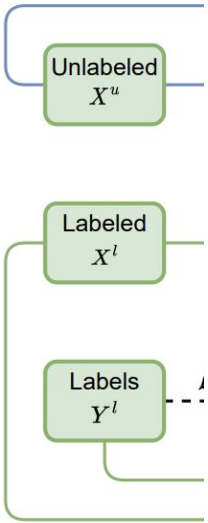
- Treats SSL as a **one-armed bandit problem**.
- **Dynamically** adapt and respond to the data.
- **Beyond standard norm** for SSL.
- Potential to **transform** SSL frameworks.

Framework

- Frames SSL as **one-armed bandit problem**.

Framework

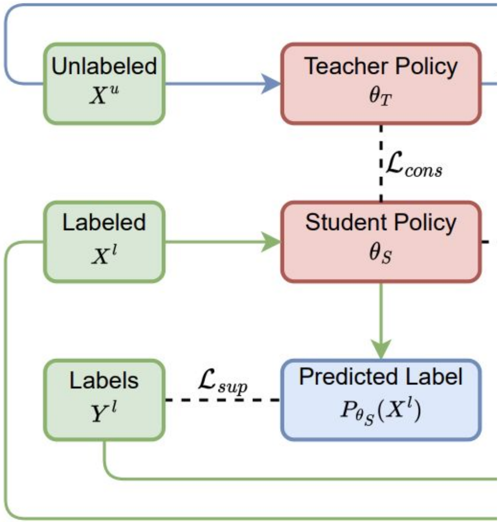
- Frames SSL as **one-armed bandit problem**.
- **State**: $s = (X^l, Y^l, X^u)$.



Carleton University

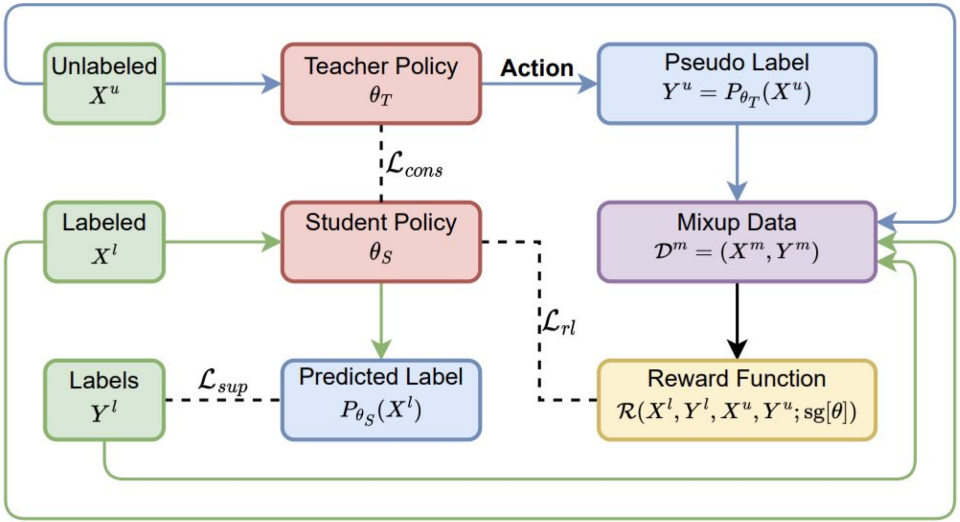
Framework

- Frames SSL as **one-armed bandit problem**.
- **State**: $s = (X^l, Y^l, X^u)$.
- **Policy**: The prediction networks ($\pi_\theta(\cdot) = P_\theta(\cdot)$)
- **Action:** Model's predictions (pseudo-labels).



Framework

- Frames SSL as **one-armed bandit problem**.
- **State**: $s = (X^l, Y^l, X^u)$.
- **Policy**: The prediction networks.($\pi_\theta(\cdot) = P_\theta(\cdot)$)
- **Action:** Model's predictions (pseudo-labels).
- **Reward**: Generalization measured via label mixup between labeled and pseudo-labeled samples.



Carleton University

Reward Function

- **Balanced** utilization of labeled and unlabeled data by Inter-mixup:

$$x_i^{\mathrm{m}} = \mu\, x_i^u + (1-\mu)\, x_i^l, \qquad \mathbf{y}_i^{\mathrm{m}} = \mu\, \mathbf{y}_i^u + (1-\mu)\, \mathbf{y}_i^l$$

- **Reward: Negative disagreement** between model predictions and mixup labels:

$$\mathcal{R}(s, a; \mathrm{sg}[\theta]) = \mathcal{R}(X^l, Y^l, X^u, Y^u; \mathrm{sg}[\theta]) = -\frac{1}{C \cdot N^{\mathrm{m}}} \sum_{i=1}^{N^{\mathrm{m}}} ||P_\theta(x_i^{\mathrm{m}}) - \mathbf{y}_i^{\mathrm{m}}||_2^2$$

Carleton University

Reinforcement Learning Loss

- **One-Armed Bandit Principle:** Optimize one-time reward based on the policy output.
- Exploits **non-differentiable reward.**
- Enables policy gradient with a **deterministic policy**.
- **KL-Divergence Weighted Negative Reward:**

$$\mathcal{L}_{\text{rl}}(\theta) = -\mathbb{E}_{\mathbf{y}_i^u \sim \pi_\theta} \mathrm{KL}(\mathbf{e}, \mathbf{y}_i^u) \mathcal{R}(s, a; \mathrm{sg}[\theta]) = -\mathbb{E}_{x_i^u \in \mathcal{D}_u} \mathrm{KL}(\mathbf{e}, P_\theta(x_i^u)) \mathcal{R}(s, a; \mathrm{sg}[\theta])$$

- Measures distance between label predictions and a uniform distribution vector *e=1/C*

Carleton University

Teacher Student Framework

- Teacher parameter update via WMA:

$$\theta_T = \beta\,\theta_T + (1-\beta)\,\theta_S$$

- **Supervised Loss** on labeled data**:**

$$\mathcal{L}_{\text{sup}}(\theta_S) = \mathbb{E}_{(x^l,\mathbf{y}^l)\in\mathcal{D}^l}\left[\ell_{CE}\left(P_{\theta_S}(x^l),\mathbf{y}^l\right)\right]$$

- **Consistency Loss** between student and teacher on unlabeled data:

$$\mathcal{L}^{\text{cons}} = \mathbb{E}_{x^u\in\mathcal{D}^u}\left[\ell_{\text{KL}}\left(P_{\theta_S}(x^u), P_{\theta_T}(x^u)\right)\right]$$

- **Learning objective:** $\mathcal{L}(\theta_S) = \mathcal{L}_{\text{rl}} + \lambda_1\mathcal{L}_{\text{sup}} + \lambda_2\mathcal{L}_{\text{cons}}$

Experimental Results

Table 1: Performance of RLGSSL and state-of-the-art SSL algorithms with the CNN-13 network. We report the average test errors and the standard deviations of 5 trials.

Dataset	CIFAR-10			CIFAR-100	
Number of Labeled Samples	1000	2000	4000	4000	10000
Supervised	$39.95_{(0.75)}$	$27.67_{(0.12)}$	$20.42_{(0.21)}$	$58.31_{(0.89)}$	$44.56_{(0.30)}$
Supervised + MixUp [40]	$31.83_{(0.65)}$	$24.22_{(0.15)}$	$17.37_{(0.35)}$	$54.87_{(0.07)}$	$40.97_{(0.47)}$
Π-model [6]	$28.74_{(0.48)}$	$17.57_{(0.44)}$	$12.36_{(0.17)}$	$55.39_{(0.55)}$	$38.06_{(0.37)}$
Temp-ensemble [6]	$25.15_{(1.46)}$	$15.78_{(0.44)}$	$11.90_{(0.25)}$	-	$38.65_{(0.51)}$
Mean Teacher[8]	$21.55_{(0.53)}$	$15.73_{(0.31)}$	$12.31_{(0.28)}$	$45.36_{(0.49)}$	$35.96_{(0.77)}$
VAT [5]	$18.12_{(0.82)}$	$13.93_{(0.33)}$	$11.10_{(0.24)}$	-	-
SNTG [15]	$18.41_{(0.52)}$	$13.64_{(0.32)}$	$10.93_{(0.14)}$	-	$37.97_{(0.29)}$
Learning to Reweight [41]	$11.74_{(0.12)}$	-	$9.44_{(0.17)}$	$46.62_{(0.29)}$	$37.31_{(0.47)}$
MT + Fast SWA [14]	$15.58_{(0.12)}$	$11.02_{(0.23)}$	$9.05_{(0.21)}$	-	$33.62_{(0.54)}$
ICT [16]	$12.44_{(0.57)}$	$8.69_{(0.15)}$	$7.18_{(0.24)}$	$40.07_{(0.38)}$	$32.24_{(0.16)}$
RLGSSL (Ours)	$\mathbf{9.15}_{(0.57)}$	$\mathbf{6.90}_{(0.11)}$	$\mathbf{6.11}_{(0.10)}$	$\mathbf{36.92}_{(0.45)}$	$\mathbf{29.12}_{(0.20)}$

Thank You