SoftMatch: Addressing the quantity-quality trade-off in Semi-supervised Learning

Problem: Threshold-based pseudo-labeling trains the model with pseudo-label whose prediction confidence is above a hard threshold.

Quantity-Quality Trade-off

- High confidence threshold: ensure the quality but discard unconfident yet correct
- Dynamically growing / Class-wise threshold: encourage more pseudo-labels but enroll those may mislead training

Problem Statement: C-class classification

$$D_L = \{x_i^l, y_i^l\}_{i=1}^{N_L} \quad \text{and} \quad D_U = \{x_i^u\}_{i=1}^{N_U}, \quad x_i^l, x_i^u \in \mathbb{R}^d$$

$p(y|x) \in \mathbb{R}^C$ denote the model's prediction

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_u$$

$$\mathcal{L}_s = \frac{1}{B_L} \sum_{i=1}^{B_L} \mathcal{H}(y_i, p(y|x_i^l))$$

$$\mathcal{L}_u = \frac{1}{B_u} \sum_{i=1}^{B_u} \lambda(p_i) \mathcal{H}(\hat{p}_i, p(y|\Omega(x_i^u)))$$

$\Omega(x^u)$: Strongly-augmented data;

p_i: $p(y|w(x_i^u))$, which $w(x_i^u)$ is weakly-augmented;

$\hat{p}_i$: one-hot pseudo-label $\arg\max(p_i)$;

$\lambda(p)$: the sampling weighting function with range $[0, \lambda_{max}]$.

A unified formulation of the confidence thresholding scheme (and other scheme) from the sample weighting perspective.

Quantity-Quality Trade-off from sample-weighting perspective:

Definition 2.1 (Quantity of pseudo-labels)**.** The quantity $f(\mathbf{p})$ of pseudo-labels enrolled in training is defined as the expectation of the sample weight $\lambda(\mathbf{p})$ over the unlabeled data:

$$f(\mathbf{p}) = \mathbb{E}_{\mathcal{D}_U}[\lambda(\mathbf{p})] \in [0, \lambda_{\max}]. \quad (3)$$

the ratio of unlabeled data enrolled in the weighted unsupervised loss

Definition 2.2 (Quality of pseudo labels)**.** The quality $g(\mathbf{p})$ is the expectation of the weighted 0/1 error of pseudo-labels, assuming the label $\mathbf{y}^u$ is given for $\mathbf{x}^u$ for only theoretical analysis purpose:

the ratio of correct pseudo-labels enrolled in

$$g(\mathbf{p}) = \sum_{i}^{N_U} \mathbb{1}(\hat{\mathbf{p}}_i = \mathbf{y}_i^u) \frac{\lambda(\mathbf{p}_i)}{\sum_j^{N_U} \lambda(\mathbf{p}_j)} = \mathbb{E}_{\bar{\lambda}(\mathbf{p})}[\mathbb{1}(\hat{\mathbf{p}} = \mathbf{y}^u)] \in [0, 1], \quad (4)$$

where $\bar{\lambda}(\mathbf{p}) = \lambda(\mathbf{p}) / \sum \lambda(\mathbf{p})$ is the probability mass function (PMF) of $\mathbf{p}$ being close to $\mathbf{y}^u$.

Based on the definitions of quality and quantity, we present the *quantity-quality trade-off* of SSL.

Definition 2.3 (The quantity-quality trade-off)**.** Due to the implicit assumptions of PMF $\bar{\lambda}(\mathbf{p})$ on the marginal distribution of model predictions, the lack of sophisticated design on it usually results in a trade-off in quantity and quality - when one of them increases, the other must decrease. Ideally, a well-defined $\lambda(\mathbf{p})$ should reflect the true distribution and lead to both high quantity and quality.

Table 1: Summary of different sample weighting function $\lambda(\mathbf{p})$, probability density function $\bar{\lambda}(\mathbf{p})$ of $\mathbf{p}$, quantity $f(\mathbf{p})$ and quality $g(\mathbf{p})$ of pseudo-labels used in previous methods and SoftMatch.

Scheme	Pseudo-Label	FixMatch	SoftMatch
$\lambda(\mathbf{p})$	$\lambda_{\max}$	$\begin{cases} \lambda_{\max}, & \text{if } \max(\mathbf{p}) \geq \tau, \\ 0.0, & \text{otherwise.} \end{cases}$	$\begin{cases} \lambda_{\max} \exp\left(-\frac{(\max(\mathbf{p}) - \mu_t)^2}{2\sigma_t^2}\right), & \text{if } \max(\mathbf{p}) < \mu_t, \\ \lambda_{\max}, & \text{otherwise.} \end{cases}$
$\bar{\lambda}(\mathbf{p})$	$1/N_U$	$\begin{cases} 1/\hat{N}_U^\tau, & \text{if } \max(\mathbf{p}) \geq \tau, \\ 0.0, & \text{otherwise.} \end{cases}$	$\begin{cases} \frac{\exp(-\frac{(\max(\mathbf{p}_i) - \hat{\mu}_t)^2}{2\hat{\sigma_t}^2})}{\frac{N_U}{2} + \sum_i^{\frac{N_U}{2}} \exp(-\frac{(\max(\mathbf{p}_i) - \hat{\mu}_t)^2}{2\hat{\sigma_t}^2})}, & \max(\mathbf{p}) < \mu_t \\ \frac{1}{\frac{N_U}{2} + \sum_i^{\frac{N_U}{2}} \exp(-\frac{(\max(\mathbf{p}_i) - \hat{\mu}_t)^2}{2\hat{\sigma_t}^2})}, & \max(\mathbf{p}) \geq \mu_t \end{cases}$
$[0, \lambda_{max}]$ $f(\mathbf{p})$	$\lambda_{\max}$	$\lambda_{\max} \hat{N}_U^\tau / N_U$	$\lambda_{\max}/2 + \lambda_{\max}/N_U \sum_i^{\frac{N_U}{2}} \exp(-\frac{(\max(\mathbf{p}_i) - \hat{\mu}_t)^2}{2\hat{\sigma_t}^2})$
$[0,1]$ $g(\mathbf{p})$	$\sum_i^{N_U} \mathbb{1}(\hat{\mathbf{p}} = \mathbf{y}^u)/N_U$	$\sum_i^{\hat{N}_U} \mathbb{1}(\hat{\mathbf{p}} = \mathbf{y}^u)/\hat{N}_U^\tau$	$\sum_j^{\hat{N}_U^{\mu_t}} \mathbb{1}(\hat{\mathbf{p}}_j = \mathbf{y}_j^u)/2\hat{N}_U + \sum_i^{N_U - \hat{N}_U^{\mu_t}} \mathbb{1}(\hat{\mathbf{p}}_i = \mathbf{y}_i^u) \exp(-\frac{(\max(\mathbf{p}_i) - \mu_t)^2}{\sigma_t^2})/2(N_U - \hat{N}_U^{\mu_t})$
Note	High Quantity Low Quality	Low Quantity High Quality	$\hat{N}_U = \sum_i^{N_U} \mathbb{1}(\max p_i \geq \mu_t)$ High Quantity High Quality

Pseudo-Label: the pseudo-labels are directly used to the model itself.

$\lambda(p) \equiv \lambda_{max}$, $\bar{\lambda}(p) = \frac{\lambda_{max}}{N_U \lambda_{max}} = \frac{1}{N_U}$, $f(p) = \sum_i^{N_U} \frac{\lambda_{max}}{N_U} = \lambda_{max}$, $g(p) = -$

FixMatch: prediction confidence $\max(p)$ is above the pre-defined threshold τ is fully enrolled during training, and others being ignored.

SoftMatch

$\lambda(p)$ of marginal distribution follows a dynamic and truncated Gaussian distribution of mean μ_t and variance σ_t at t-th training iteration.

$$\lambda(p) = \begin{cases} \lambda_{max} \exp\left(-\frac{(\max(p)-\mu_t)^2}{2\sigma_t^2}\right), & \text{if } \max(p) < \mu_t, \\ \lambda_{max}, & \text{otherwise} \end{cases}$$

Underlying true Gaussian parameters μ_t and σ_t are unknown.

We can estimate μ and σ^2 from the historical predictions of the model.

$$\hat{\mu}_b = \hat{\mathbb{E}}_{B_U}[\max(p)] = \frac{1}{B_U} \sum_{i=1}^{B_U} \max(p_i)$$

$$\hat{\sigma}_b = \hat{Var}_{B_U}[\max(p)] = \frac{1}{B_U} \sum_{i=1}^{B_U} (\max(p_i) - \hat{\mu}_b)^2.$$

Aggregate the batch statistics for a more stable estimation, using EMA:

$$\hat{\mu}_t = m\hat{\mu}_{t-1} + (1-m)\hat{\mu}_b,$$

$$\hat{\sigma}_t^2 = m\hat{\sigma}_{t-1}^2 + (1-m)\frac{B_U}{B_U-1}\hat{\sigma}_b^2$$

$$\hat{\mu}_0 = \frac{1}{C}, \quad \hat{\sigma}_0^2 = 1.0$$

Quantity $f(p) \in \left[\frac{\lambda_{max}}{2}\left(1+\exp\left(-\frac{(\frac{1}{C}-\hat{\mu}_t)^2}{2\hat{\sigma}_t^2}\right)\right), \lambda_{max}\right]$

guarantee at least $\lambda_{max}/2$ of quantity

Quality $g(p)$ at least $\sum_{j}^{\hat{N}_u} \frac{\mathbb{1}(\hat{p}_j = y_j^u)}{2\hat{N}_u}$

As $\hat{\mu}_t$ increases and $\hat{\sigma}_t$ decreases, the quantity maintains high, the quality of pseudo-labels also improves. ($\hat{N}_u$ increases)

Uniform Alignment For Fair Quantity:

Different classes exhibit different learning difficulties, generated pseudo-labels can have potentially imbalanced distribution.

Uniform Alignment (UA) encourages more uniform pseudo-labels of different classes.

$$UA(p) = \text{Normalize}\left(p \cdot \frac{u(C)}{\hat{\mathbb{E}}_{B_U}[p]}\right),$$

$C = 3,\ B_U = 1$

$P = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)^T$

$UA(p) = p \cdot \dfrac{\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^T}{\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)^T}$

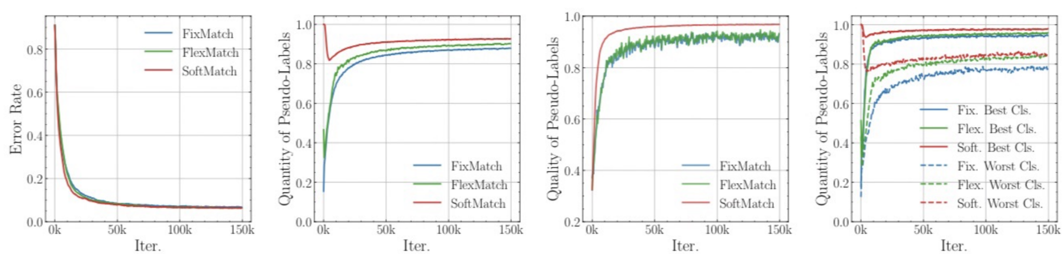
$u(C) \in \mathbb{R}^C$: a uniform distribution

$\text{Normalize}(\cdot) = (\cdot) / \sum(\cdot)$ ensuring the probability sums to 1, $= \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^T$

$$\lambda(p) = \begin{cases} \lambda_{max} \exp\left(-\dfrac{(\max(UA(p)) - \hat{\mu}_t)^2}{2\hat{\sigma}_t^2}\right), & \text{if } \max(UA(p)) < \hat{\mu}_t \\ \lambda_{max}, & \text{otherwise} \end{cases}$$

UA encourages larger weights to be assigned to less-predicted pseudo-labels and smaller weights to more-predicted pseudo-labels

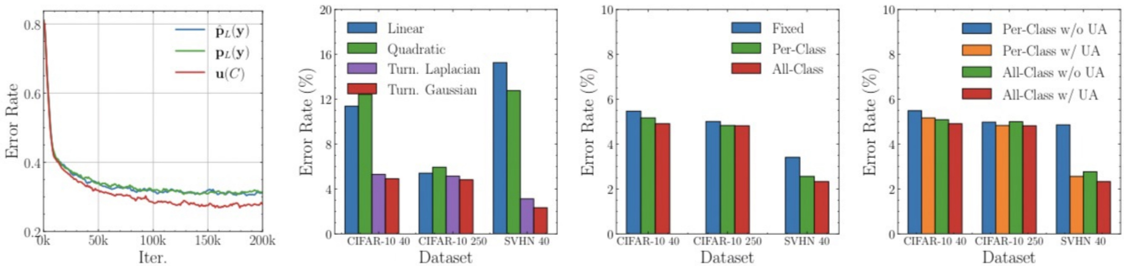
Experiments:



(a) Eval. Error (b) Quantity (c) Quality (d) Cls. Quality

Figure 2: Qualitative analysis of FixMatch, FlexMatch, and SoftMatch on CIFAR-10 with 250 labels. (a) Evaluation error; (b) Quantity of Pseudo-Labels; (c) Quality of Pseudo-Labels; (d) Quality of Pseudo-Labels from the best and worst learned class. Quality is computed according to the underlying ground truth labels. SoftMatch achieves significantly better performance.

Dataset	CIFAR-10-LT			CIFAR-100-LT		
Imbalance γ	50	100	150	20	50	100
FixMatch	18.46±0.30	25.11±1.20	29.62±0.88	50.42±0.78	57.89±0.33	62.40±0.48
FlexMatch	18.13±0.19	25.51±0.92	29.80±0.36	49.11±0.60	57.20±0.39	62.70±0.47
SoftMatch	**16.55**±**0.29**	**22.93**±**0.37**	**27.40**±**0.46**	**48.09**±**0.55**	**56.24**±**0.51**	**61.08**±**0.81**



(a) L.T. UA (b) Weight. Func. (c) Gau. Param. (d) UA

Figure 3: Ablation study of SoftMatch. (a) Target distributions for Uniform Alignment (UA) on long-tailed setting; (b) Error rate of different sample functions; (c) Error rate of different Gaussian parameter estimation, with UA enabled; (d) Ablation on UA with Gaussian parameter estimation;