Semi-Supervised Learning based on Pseudo-labeling

Shengjie Niu

23 Summer Study - Week3

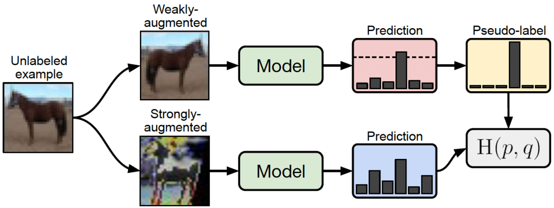
niusj03@gmail.com

July 5, 2023

Overview

Semi-Supervised Learning (SSL)

- Leverage unlabeled data to improve the performance when labeled data are limited.
- Recent state-of-the-art SSL types
  - **Pseudo labeling (PL)**
  - Consistency regularization
  - Entropy minimization
  - Combination of above
- Related Works
  - Data Augmentations
  - Active Learning
  - Curriculum Learning
  - Learning etc.



e.g. FixMatch, K Sohn et al. (2020)

$\min_{\theta\in\Theta} L(\mathcal{D}_L, \theta) + \Omega(\mathcal{D}_U, \theta)$

Why Semi-Supervised Learning (SSL)

- Machine learning algorithms are data-driven.
- Acquiring large amounts of labeled data can be a expensive, labor-intensive and time-consuming process.
- SSL is a hybrid approach that lies between supervised learning and unsupervised learning.
- Deep SSL has demonstrated highly competitive performance compared to supervised learning models.

Hypothesis of SSL

Primary Hypothesis, link:

- Smoothness Hypothesis
- Cluster Hypothesis
- Low-density Separation Hypothesis
- Mainfold Hypothesis

Additionally Hypothesis (Impractical Scenarios):

- Homogeneous Hypothesis - Open-set SSL
- Uniform Hypothesis - Imbalance SSL

Problem Formulation

Training Data

- Training data: $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$.
- $\mathcal{D}_l = \{(\boldsymbol{x}_l, \boldsymbol{y}_l)\}_{l=1}^{B}$, $\mathcal{D}_u = \{(\boldsymbol{x}_u)\}_{u=1}^{\mu B}$, where $\mu \gg 1$ determining the relative size of $\mathcal{D}_l$ and $\mathcal{D}_u$.
- $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^D, \boldsymbol{y} \in \mathcal{Y} = \{1, \cdots, C\}$ where D is the input dimension and C is the number of output class in labeled data.

Objective

Train a model $p_m(\boldsymbol{x}; \theta) : \{\mathcal{X}; \Theta\} \rightarrow \mathcal{Y}$ from training data to minimize the generalization risk $R(p_m) = \mathbb{E}_{(X,Y)}[l(p_m(X; \theta), Y)]$.

Pseudo-Labeling (PL)

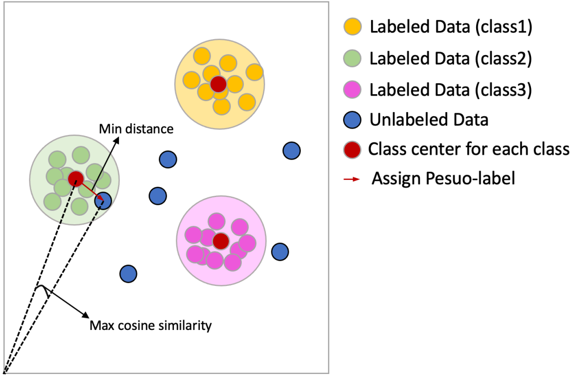
It acts like self-training by utilizing the model itself to obtain artificial labels for unlabeled data D Lee et al. (2014)

Label Guesser:

- Linear PL
- Semantic PL
- Wasserstein PL etc.

Loss form:

$$L_u = \frac{1}{\mu B} \sum_{u=1}^{\mu B} \mathbb{I}(\text{con}) \Omega(\hat{\boldsymbol{p}}_u, p_m(\boldsymbol{y}|\boldsymbol{x}_u)), \qquad (1)$$



Deformable Template Feature-level Pool

Consistency regularization

It leverages unlabeled data based on a primary assumption that model should produce similar predictions for perturbed versions of the same image.
encourages a model to produce the same prediction when the input is perturbed.

$$\Omega(\boldsymbol{x}; \theta) = \mathcal{H}(p_m(\boldsymbol{y}|\boldsymbol{x}^w), p_m(\boldsymbol{y}|\boldsymbol{x}^s)), \tag{2}$$

Augmentations:

- Weak augmentations: small translations, rotations, flips etc.
  Make model more robust to small variations in the input without changing its semantic meaning
- Strong augmentations: RandAugment

Objective Function

A popular form of unsupervised objective combining data augmentation, consistency regularization and PL is formulated as follow:

$$L_U = \frac{1}{\mu B} \sum_{u=1}^{\mu B} \mathbb{I}(\text{con}) \mathcal{H}(\hat{\boldsymbol{p}}_u, p_m(\boldsymbol{y}|\boldsymbol{x}_u^s)), \tag{3}$$

and the supervised objective is formulated as:

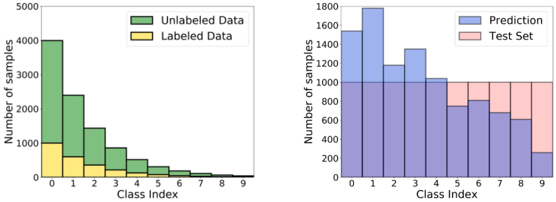
$$L_S = \frac{1}{B} \sum_{l=1}^{B} \mathcal{H}(\boldsymbol{y}_l, p_m(\boldsymbol{y}|\boldsymbol{x}_l^s)). \tag{4}$$

The objective function is

$$L = L_S + L_U \tag{5}$$

Imbalanced SSL

- Datasets of real-world exhibit class imbalanced, or long tailed distributions.
- Classifiers are biased toward the majority classes
- Objective: produce debiased pseudo-labels with class-imbalanced data
- Some Techniques
    - Re-sampling
    - Re-weighting
    - Adaptive Thresholding
    - Re-balancing
    - decouple learning representation and classifier etc.



Source: ABC, H Lee et al. (2021)

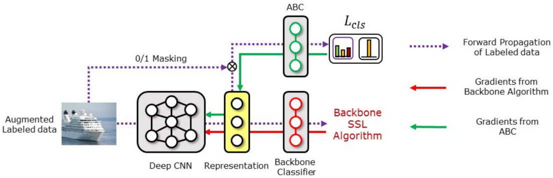
The degree of imbalance for each dataset is characterized by the imbalance ratio, γ_l, γ_u, where $\gamma_l = \frac{\max_k N_k}{\min_k N_k}$

Supervised Loss: (N_L refers to # of minority class)

$$L_S = \frac{1}{B}\sum_{l=1}^{B} M(\boldsymbol{x}_l)\mathcal{H}(\boldsymbol{y}_l, p_m(\boldsymbol{y}|\boldsymbol{x}_l^s)),$$
$$M(\boldsymbol{x}_l) = \mathcal{B}(\frac{N_L}{N_{\boldsymbol{y}_l}}) \tag{6}$$

Unsupervised Loss:

$$L_U = \frac{1}{\mu B}\sum_{u=1}^{\mu B} M(\boldsymbol{x}_u)\mathbb{I}(\text{con})\mathcal{H}(\hat{\boldsymbol{p}}_u, p_m(\boldsymbol{y}|\boldsymbol{x}_u^s)),$$
$$M(\boldsymbol{x}_u) = \mathcal{B}(\frac{N_L}{N_{\hat{\boldsymbol{y}}_l}}) \tag{7}$$



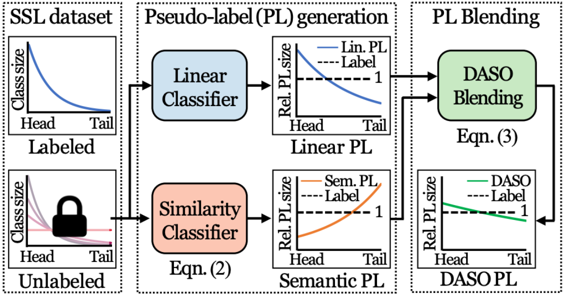
Source: ABC, H Lee et al. (2021)

Attach the ABC to the backbone's representation layer to utilize the high-quality representations.

- Blend two complementary PLs from different classifiers.
  - Linear: low recall, high precision in minority classes
  - Semantic: high recall, low precision in minority classes
  - Trade-offs between Linear and Semantic
- Distribution-Aware Blending:

$$\hat{p}_D = (1 - v_{k'})\hat{p}_L + v_{k'}\hat{p}_S, \qquad (8)$$

where $v_k = \frac{1}{\max_k \hat{m}_k^{1/T}}(\hat{m}_k^{1/T})$



Source: DASO, Y Oh et al. (2022)

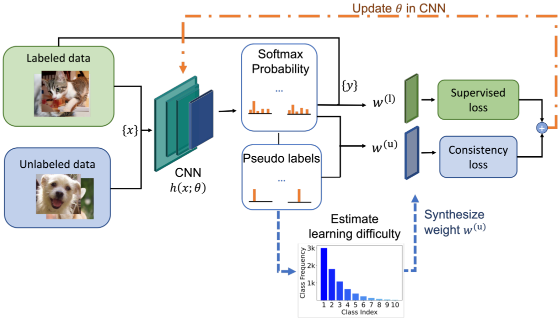
Smoothed Adaptive Weighting - 23ICML

Supervised & unsupervised Loss:

$$L_S = \frac{1}{B}\sum_{l=1}^{B} w_k \mathcal{H}(\boldsymbol{y}_l, p_m(\boldsymbol{y}|\boldsymbol{x}_l^s)),$$
$$L_U = \frac{1}{\mu B}\sum_{u=1}^{\mu B} w_k \mathbb{I}(\mathrm{con})\mathcal{H}(\hat{\boldsymbol{p}}_u, p_m(\boldsymbol{y}|\boldsymbol{x}_u^s)). \tag{9}$$

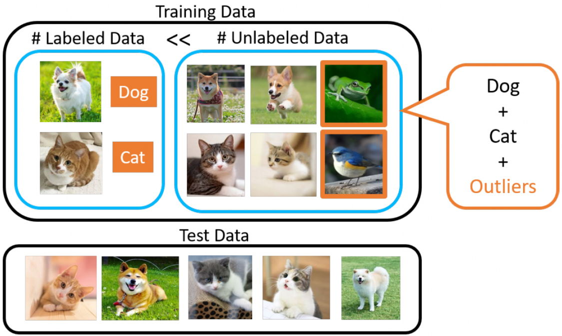
Weighting Function:

$$w_k \propto 1/E_k, E_k = (1-\beta^{n_k})/(1-\beta)$$
$$n_k = \sum_{u=1}^{N_U} p(\boldsymbol{x}_u, \theta)_k \tag{10}$$



Source: SAW, Z Lai et al. (2022)

Open-set SSL

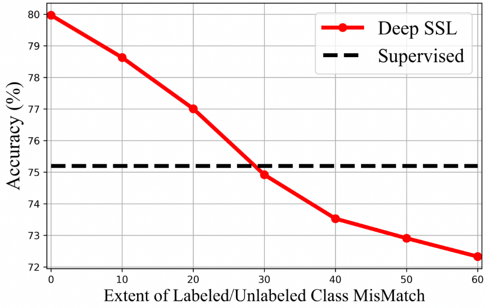


Source: MTC, Q Yu et al. (2020)

- Promising result in SSL are based on homogeneous hypothesis - easily violated in practical applications. (MisMatch, OOD samples)
- Outliers: do not belong to the classes of labeled data, exist in the unlabeled data.
- Deep SSL no longer works well and accompanies with severe performance degradation.

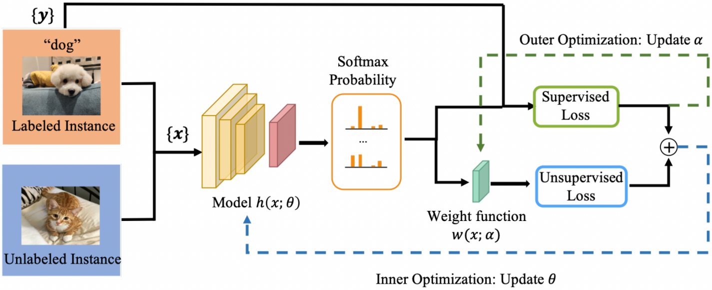
Open-set SSL

- Deep SSL is even worse than a simple SL model.
- Objective: the model should be trained by eliminating the effect of these outliers.
- Existing methodlogies:
  - D3SL-20ICML, link.
  - MTC-20CVPR, link.
  - UASD-20AAAI, link.
  - OpenMatch-21CVPR, link.



Source: D3SL, L-Z Guo et al. (2020)

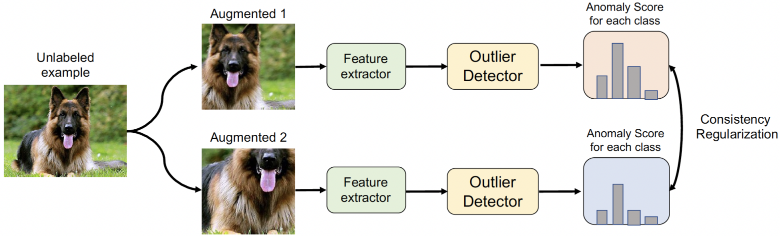
D3SL - 20ICML



Source: D3SL, L-Z Guo et al. (2020)

$$
\begin{aligned}
\hat{\theta}(\alpha) &= min_{\theta\in\Theta}\sum_i l(\boldsymbol{x}_i, \boldsymbol{y}_i; \theta) + \sum_u w(\boldsymbol{x}_u; \alpha)\Omega(\boldsymbol{x}_i; \theta) \\
\hat{\alpha} &= argmin_{\alpha}\sum_i l(\boldsymbol{x}_i, \boldsymbol{y}_i; \hat{\theta}(\alpha))
\end{aligned} \tag{11}
$$

OpenMatch - 21CVPR

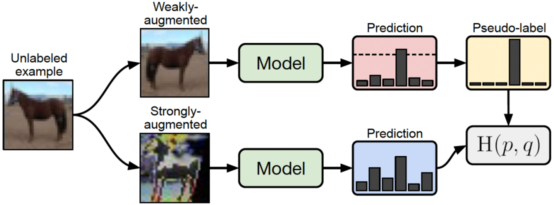


Source: OpenMatch, K Saito et al. (2023)

- One-Vs-All (OVA) network that can learn a threshold to distinguish outliers from inliers.
- Soft open-set consistency regularization (SOCR) for more effective representations.

Dynamic Thresholding Schemes

- Leverage unlabeled data to improve the performance when labeled data are limited.
- Recent state-of-the-art SSL types:
  - **Pseudo labeling (PL)**
  - Consistency regularization
  - Entropy minimization
  - Combination of above
- Some Techniques
  - Data Augmentations
  - Active Learning
  - Curriculum Learning
  - Learning etc.



e.g. FixMatch, K Sohn et al. (2020)

$\min_{\theta\in\Theta} L(\mathcal{D}_L, \theta) + \Omega(\mathcal{D}_U, \theta)$

Adjust class-wise thresholds:

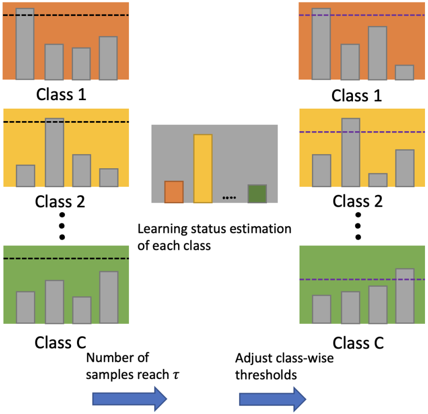
$$\mathcal{T}_t(c) = \sigma_t(c) \cdot \tau(c \in [1, 2, \cdots, \mathcal{C}]), \qquad (12)$$

Learning status evaluation:

$$\sigma_t(c) = \sum_{u=1}^{\mu B} \mathbb{I}(\max(\boldsymbol{p}_u) \geq \tau \ \& \ m = n = c). \qquad (13)$$

MaxNorm Scaling:

$$\mathcal{T}_t(c) = \frac{\sigma_t(c)}{\mathrm{max_c}\sigma_\mathrm{t}(\mathrm{c})} \cdot \tau \qquad (14)$$



Source: FreeMatch, Y Wang et al. (2023)

Adjust global threshold τ_t:

$$\tau_t(c) = \sigma_t(c) \cdot \tau_t, (c \in [1, 2, \cdots, \mathcal{C}]), \tag{15}$$

A expected global threshold should

- reflect the overall learning status
- progressively increase

Global threshold τ_t with EMA:

$$\tau_b = \frac{1}{\mu B} \sum_{u=1}^{\mu B} \max(\boldsymbol{p}_u), \tag{16}$$

$$\tau_t = \lambda \tau_{t-1} + \lambda \tau_b$$

Hard v.s. Soft Thresholding Scheme:

$$L_U = \frac{1}{\mu B}\sum_{u=1}^{\mu B} \mathcal{I}(\text{con})\mathcal{H}(\hat{\boldsymbol{p}}_u, \boldsymbol{q}_u^s),$$
$$L_U = \frac{1}{\mu B}\sum_{u=1}^{\mu B} \lambda(\boldsymbol{f}_u)\mathcal{H}(\hat{\boldsymbol{p}}_u, \boldsymbol{q}_u^s), \tag{17}$$

where $\lambda(\boldsymbol{f}) \in [0, \lambda_{\max}]$ refers to sample weighting function.

Quantity of pseudo-labels: Expectation of the weighting function $\lambda(\boldsymbol{f})$ over the unlabeled data:

$$Q_1 = \mathbb{E}_{\mathcal{D}_U}[\lambda(\boldsymbol{f})] \in [0, \lambda_{\max}]. \tag{18}$$

Quality of pseudo-labels: Expectation of the weighted 0/1 errors of pseudo-labels:

$$Q_2 = \sum_{u=1}^{N_U} \mathbb{I}(\boldsymbol{y}_u = \hat{\boldsymbol{p}}_u)\frac{\lambda(\boldsymbol{f}_u)}{\sum_{i=1}^{N_U}\lambda(\boldsymbol{f}_i)} \in [0, 1]. \tag{19}$$

Unlabeled Weighting Function

Specifically, I assume it follows a dynamic and truncated Gaussian distribution with mean μ_t and variance σ_t:

$$\lambda(\boldsymbol{f}) = \begin{cases} \lambda_{\max} \exp(-\frac{(\max(\boldsymbol{p}_u) - \mu_t)}{2\sigma_t^2}), & \text{if } \max(\boldsymbol{p}_u) < \mu_t \\ \lambda_{\max} & \text{otherwise} \end{cases}, \tag{20}$$

where the empirical mean and variance can be computed as:

$$\begin{aligned} \hat{\mu}_b &= \hat{\mathbb{E}}_{\mu B}[\max(\boldsymbol{p}_u)] = \frac{1}{\mu B} \sum_{u=1}^{\mu B} \max(\boldsymbol{p}_u), \\ \hat{\sigma}_b^2 &= \hat{Var}_{\mu B}[\max(\boldsymbol{p}_u)] = \frac{1}{\mu B} \sum_{u=1}^{\mu B} (\max(\boldsymbol{p}_u) - \hat{\mu}_b)^2, \end{aligned} \tag{21}$$

Quantity-Quality Trade-Off

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

Source: SoftMatch, H Chen et al. (2023)

Conclusion

- SSL: Leverage numerous of cheap unlabeled data to enhance model performance.
- Hypothesis for SSL in the literature (two additional and impractical assumptions).
- Intro to Open-set SSL & Imbalance SSL
- PL-related SSL (Dynamic thresholding), Soft-thresholding, Quantity-Quality Trade=off.