Adaptive Sample Selection for Robust Learning under Label Noise

Deep Patel and P S Sastry

Department of Electrical Engineering, Indian Institute of Science, Bengaluru, INDIA

WACV WAIKOLOA HAWAII JAN 3-7 • 2023

Learning under Label Noise

- **Given:** Noisily labelled training set, $S_\eta = \{(x_i, y_i)\}_{i=1}^m$, drawn from $\mathcal{D}_\eta$
- ★ **Want:** Learn a classifier to perform well on clean training data, $S = \{(x_i, y_i^{cl})\}_{i=1}^m$, drawn from $\mathcal{D}$
- **Modelling Label Noise:** y_i's (noisy labels) are random variables dependent on the clean labels, y_i^{cl} as:
  - ★ Non-Uniform Noise (NULN):

$$\eta_{\mathbf{x},ij} = \mathbb{P}[y_\mathbf{x} = j | \mathbf{x}, y_\mathbf{x}^{cl} = i] \; \forall i \neq j \in [K], \; \forall \mathbf{x} \tag{1}$$

  - ★ Case #1: Symmetric Noise (SLN):

$$\eta_{\mathbf{x},ij} = \frac{\eta}{K-1}, \; \eta_{\mathbf{x},ii} = 1 - \eta \; \forall j \neq i, \; \forall \mathbf{x} \tag{2}$$

  - ★ Case #2: Class-conditional Noise (CCLN):

$$\eta_{\mathbf{x},ij} = \eta_{ij} = \mathbb{P}[y_\mathbf{x} = j | y_\mathbf{x}^{cl} = i] \; \forall j \neq i, \; \forall \mathbf{x} \tag{3}$$

Sample Reweighting: A popular approach

- Arpit et al. [1] show that neural nets learn from clean data before overfitting to label noise.
- Similarity with *curriculum learning* $\implies$ 'cleanly labelled' $\equiv$ 'Easy' & 'noisily labelled' $\equiv$ 'Hard'
- Typically, this 'difficulty' is modelled via sample weights.

★ **Idea:** Binary/Real-valued Weighted Risk Minimization to reduce overfitting

★ **Popular heuristic:** Sample weight = *function*(loss value of that sample)

Proposed Framework: An adaptive curriculum

★ **Observation:** Loss value of a sample depends on current state of learning & changes differently per-class throughout training

★ **Idea:** Use statistics of loss values of samples in a minibatch to adaptively infer whether a sample has clean labels.

An Adaptive Curriculum

General Curriculum:

$$\min_{\theta, \mathbf{w} \in [0,1]^m} \mathcal{L}_{\text{wtd}}(\theta, \mathbf{w}) = \sum_{i=1}^m w_i \underbrace{\mathcal{L}(f(\mathbf{x}_i; \theta), y_i)}_{\ell_i} + G(\mathbf{w}) \tag{4}$$

Self-Paced Learning [3]:

Use $G(\mathbf{w}) = -\lambda ||\mathbf{w}||_1, \; \lambda > 0$. For a fixed θ, $\forall i \in [m]$, we have

$$w_i^* = \arg \min_{w_i \in [0,1]} \sum_{i=1}^m w_i \ell_i - \lambda w_i \tag{5}$$

$$\implies w_i^* = \begin{cases} 1, & \text{if } \ell_i \leq \lambda \\ 0, & \text{else} \end{cases} \tag{6}$$

Our Formulation:

Use $G(\mathbf{w}) = -||\boldsymbol{\lambda} \odot \mathbf{w}||_1$. For a fixed θ, $\forall i \in [m]$, we have

$$w_i^* = \arg \min_{w_i \in [0,1]} \sum_{i=1}^m w_i (\ell_i - \underbrace{\lambda(y_i, \theta, \{\mathbf{x}_j\}_{j=1}^m)}_{\lambda_{y_i}}) \tag{7}$$

$$\implies w_i^* = \begin{cases} 1, & \text{if } \ell_i \leq \lambda_{y_i} \text{ or } f_{y_i}(\mathbf{x}_i; \theta) \geq \lambda_{y_i} \\ 0, & \text{else} \end{cases} \tag{8}$$

We make the thresholds λ_{y_i} dependent on current state of learning via:

$$\implies w_i^* = \begin{cases} 1, & \text{if } f_{y_i}(\mathbf{x}_i; \theta) \geq \lambda_{y_i} = \mu_{y_i} + \kappa \cdot \sigma_{y_i} \\ 0, & \text{else} \end{cases} \tag{9}$$

where $\forall p \in [K]$

$$\mu_p := \frac{1}{|\mathcal{S}_p|} \sum_{s \in \mathcal{S}_p} f_p(\mathbf{x}_s; \theta) \text{ and } \sigma_p^2 := \frac{1}{|\mathcal{S}_p|} \sum_{s \in \mathcal{S}_p} (f_p(\mathbf{x}_s; \theta) - \mu_p)^2 \tag{10}$$

Note: $\mathcal{S}_p := \{k \in [m] | y_k = e_p\} \; \forall p \in [K]$ where m is the mini-batch size.

Algorithm:

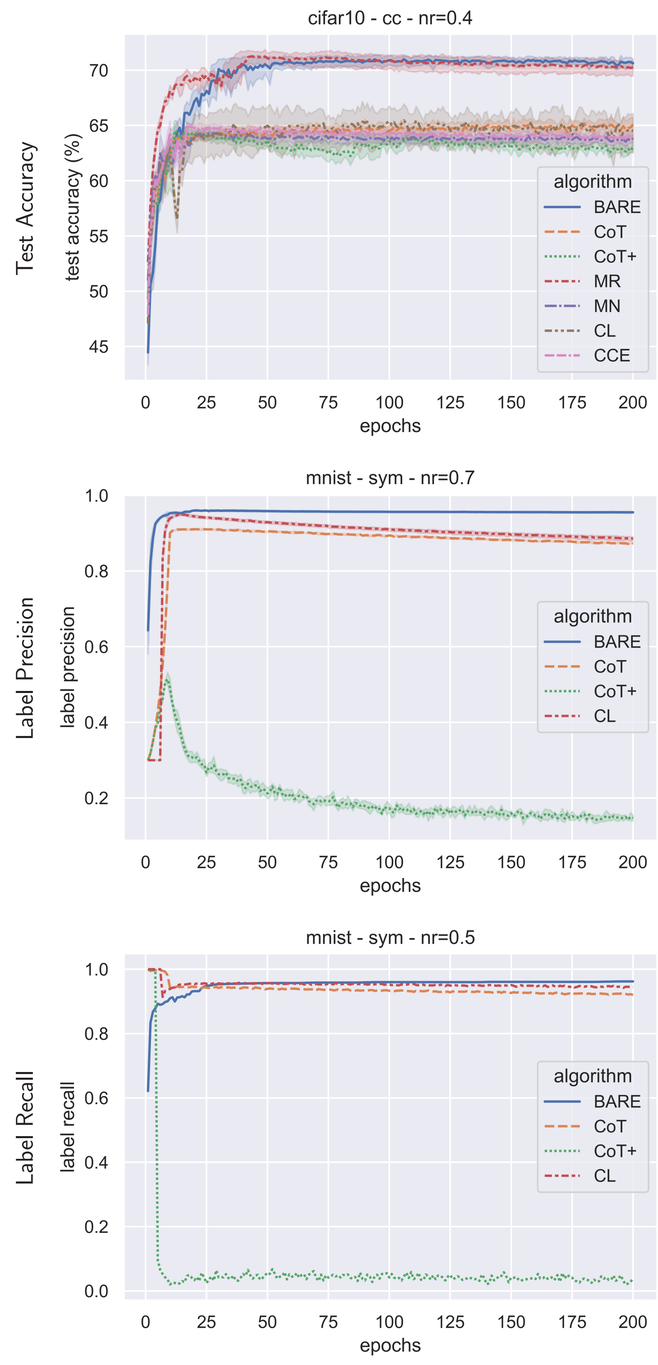
BAtch **RE**weighting (**BARE**)

Input: Noisy training set S_η, # classes K, # epochs T_{max}, mini-batch size m, and κ
Initialize: network parameters, θ_0, for classifier $f(\cdot; \theta)$
while $t \leq T_{max}$ **do**
  Shuffle training set $\mathcal{D}_\eta$
  for each mini-batch $\mathcal{M}$ from $\mathcal{D}_\eta$ **do**
    for each class $p \in [K]$ **do**
      Compute loss statistics (Equation 10) for $\mathcal{S}_p = \{k \in [m] | y_k = e_p\}$
    end
    $\mathcal{R} := \{(\mathbf{x}, y_\mathbf{x}) \mid w_{y_\mathbf{x}}^* = 1$ as per Equation 9$\}$
    $\theta \leftarrow \theta - \alpha \nabla_\theta \left(\frac{1}{|\mathcal{R}|} \sum_{(\mathbf{x}, y_\mathbf{x}) \in \mathcal{R}} \mathcal{L}(\mathbf{x}, y_\mathbf{x}; \theta) \right)$
  end
end

Experimental Setup

- Performance Metrics:
  - ✠ Test Accuracy (on a separate test set with clean labels)
  - ✠ Label Precision (# clean labels selected / # selected labels)
  - ✠ Label Recall (# clean labels selected / # clean labels)
- Datasets:
  - ■ MNIST (SLN/CCLN)
  - ■ CIFAR-10 (SLN/CCLN)
  - ■ Clothing-1M [4] (NULN)
  - ■ Food-101N [2] (NULN)
- Architectures:
  - ♦ MLPs
  - ♦ Standard CNNs
  - ♦ ResNets

Results



BARE has comparable or better test accuracies (first row), label precision (second row), and label recall (third row) compared to baselines

Conclusion

We propose BARE, an adaptive sample selection method, for robust learning under label noise with neural networks.

- Use class-wise statistics of loss values of samples.
- No knowledge of noise rates, extra clean data or training of multiple nets needed.
- On par or improved network performance in terms of test accuracy, label precision, and label recall.

Acknowledgments

Deep Patel is supported by a Prime Minister's Research Fellowship, Government of India. We thank NVIDIA for providing the Titan X Pascal GPU.

References

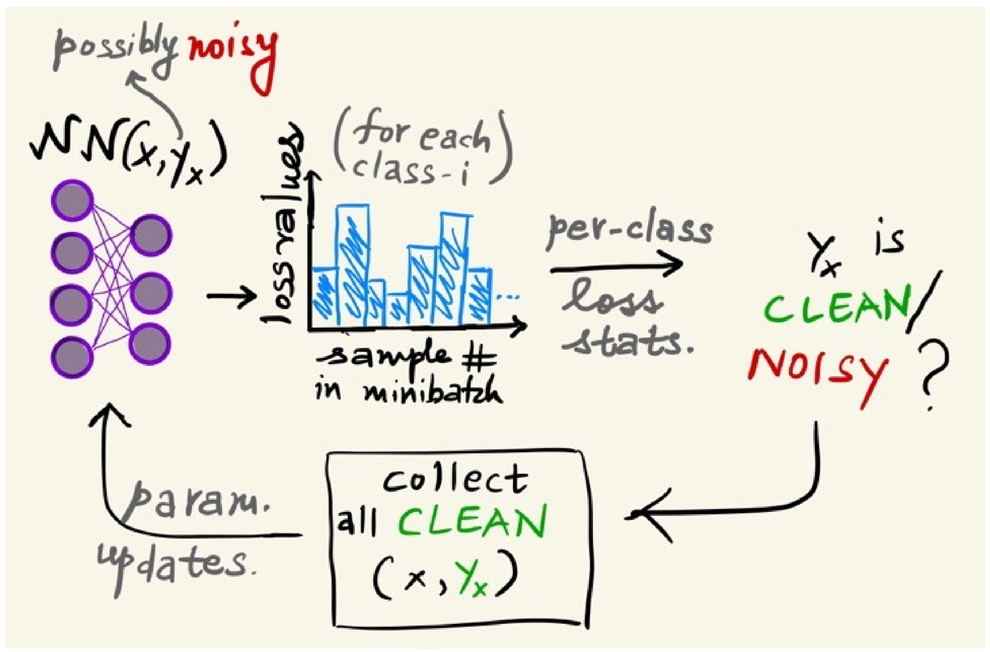
[1] Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International Conference on Machine Learning. pp. 233–242. PMLR (2017)

[2] Bossard, L., Guillaumin, M., Gool, L.V.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)

[3] Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 1. pp. 1189–1197 (2010)

[4] Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2691–2699 (2015)

Outline of the proposed algorithm



Links for Paper and Code

{deeppatel,sastry}@iisc.ac.in

https://github.com/dbp1994/bare-wacv-2023