Towards Understanding Deep Learning from Noisy Labels with Small-Loss Criterion

Xian-Jin Gui, Wei Wang and **Zhang-Hao Tian**
National Key Lab for Novel Software Technology, Nanjing University, China

Background

Deep neural networks need large amounts of labeled data to achieve good performance. In real-world applications, labels are usually collected from non-experts to save cost and thus are noisy. In the past few years, many deep learning methods based on the small-loss criterion for dealing with noisy labels have been developed. However, there are few theoretical analyses to explain why these methods could learn well from noisy labels.

Our Contribution

- We theoretically explain why the widely-used small-loss criterion works.
- Based on the explanation, we reformalize the vanilla small-loss criterion to better tackle noisy labels.
- The experimental results verify our theoretical explanation and also demonstrate the effectiveness of the reformalization.

Preliminaries

Neural network:

$g(\boldsymbol{x};\Theta): \mathcal{X} \to \mathbb{R}^c$ with output $[\hat{p}_1(\boldsymbol{x}),\dots,\hat{p}_c(\boldsymbol{x})]^\top \in \mathbb{R}^c$

where $\hat{p}_i(\boldsymbol{x}) = \frac{\exp(\boldsymbol{w}_i^\top \phi(\boldsymbol{x};\boldsymbol{\theta}))}{\sum_{j=1}^c \exp(\boldsymbol{w}_j^\top \phi(\boldsymbol{x};\boldsymbol{\theta}))}$

Loss function:

- 0-1 loss $\ell_{01}(f(\boldsymbol{x}),\tilde{y}) = \mathbb{I}[f(\boldsymbol{x}) \neq \tilde{y}]$
- Cross-entropy loss $\ell_{CE}(g(\boldsymbol{x};\Theta),\tilde{y}) = -\log(\hat{p}_{\tilde{y}}(\boldsymbol{x}))$

Noise transition matrix:

$$T_{ij} = p(\tilde{y} = j | y = i)$$

Learning process:

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}_{(\boldsymbol{x},\tilde{y})}[\ell_{CE}(g(\boldsymbol{x};\Theta),\tilde{y})]. \quad (1)$$

Our Work

Phenomenon:

The examples with correct labels will have smaller loss than the examples with incorrect labels.

Practical strategy:

For a warmed-up neural network g:

1. selects the examples with small loss values;
2. update the model parameter with these selected examples.

Lemma 1. *If T satisfies the row-diagonally dominant condition $T_{ii} > \max_{j\neq i} T_{ij}$, $\forall i$, then the target concept f^* has the minimum expected 0-1 loss on the noisy data, i.e., $\forall f \neq f^*$, $\mathbb{E}_{(\boldsymbol{x},\tilde{y})}[\ell_{01}(f^*(\boldsymbol{x}),\tilde{y})] \leq \mathbb{E}_{(\boldsymbol{x},\tilde{y})}[\ell_{01}(f(\boldsymbol{x}),\tilde{y})]$.*

With row-diagonally dominant condition, the target concept f^* has the minimum expected 0-1 loss on noisy data.

Lemma 2. *Let g^* denote the deep neural network minimizing the cross-entropy loss in Eq. (1), the induced classifier f_{g^*} satisfies $f_{g^*}(\boldsymbol{x}) = y$, $\forall \boldsymbol{x} \in \mathcal{X}$, if and only if T satisfies the row-diagonally dominant condition $T_{ii} > \max_{j\neq i} T_{ij}$, $\forall i$.*

With row-diagonally dominant condition, good neural network can be learned by minimizing the expected cross-entropy loss on noisy data.

Theorem 1. *Let g^* denote the deep neural network minimizing the cross-entropy loss in Eq. (1), $(\boldsymbol{x}_1,\tilde{y})$ and $(\boldsymbol{x}_2,\tilde{y})$ are any two examples with the same observed label $\tilde{y}$ in $\tilde{D}$ satisfying that $f^*(\boldsymbol{x}_1) = \tilde{y}$ and $f^*(\boldsymbol{x}_2) \neq \tilde{y}$, if T satisfies the diagonally-dominant condition $T_{ii} > \max\{\max_{j\neq i} T_{ij}, \max_{j\neq i} T_{ji}\}$, $\forall i$, then $\ell_{CE}(g^*(\boldsymbol{x}_1),\tilde{y}) < \ell_{CE}(g^*(\boldsymbol{x}_2),\tilde{y})$.*

With diagonally-dominant condition, for the g^* minimizing the expected cross-entropy loss on noisy data, the examples with correct labels will have smaller loss than that with incorrect labels.

Theorem 2. *Suppose g is ϵ-close to g^*, i.e., $\|g - g^*\|_\infty = \epsilon$, for two examples $(\boldsymbol{x}_1,\tilde{y})$ and $(\boldsymbol{x}_2,\tilde{y})$, assume $f^*(\boldsymbol{x}_1) = \tilde{y}$ and $f^*(\boldsymbol{x}_2) \neq \tilde{y}$, if T satisfies the diagonally-dominant condition $T_{ii} > \max\{\max_{j\neq i} T_{ij}, \max_{j\neq i} T_{ji}\}$, $\forall i$, and $\epsilon < \frac{1}{2}\cdot(T_{\tilde{y}\tilde{y}} - T_{f^*(\boldsymbol{x}_2)\tilde{y}})$, then $\ell_{CE}(g(\boldsymbol{x}_1),\tilde{y}) < \ell_{CE}(g(\boldsymbol{x}_2),\tilde{y})$.*

With diagonally-dominant condition, for a neural network g which is not far away from g^*, the examples with correct labels will have smaller loss than that with incorrect labels.

This explains why small-loss criterion works.

Algorithm 1 RSL: Reformalization of Small-Loss criterion

Input: Noisy dataset $\tilde{D}$, the initial model $g(\boldsymbol{x};\Theta^{(0)})$, epoch limit E

for $t = 1,\dots,E$ **do**
  Update $\Theta^{(t-1)}$ on $\tilde{D}$ with one epoch to get $\Theta^{(t)}$;
  Calculate each example's loss:
    $\forall(\boldsymbol{x},\tilde{y}) \in \tilde{D}$, $\ell_t(\boldsymbol{x},\tilde{y}) = \ell_{CE}(g(\boldsymbol{x};\Theta^{(t)}),\tilde{y})$;
end for
Calculate each example's mean loss:
  $\forall(\boldsymbol{x},\tilde{y}) \in \tilde{D}$, $\bar{\ell}(\boldsymbol{x},\tilde{y}) = \frac{1}{E}\sum_{t=1}^{E} \ell_t(\boldsymbol{x},\tilde{y})$;
for $i = 1,\dots,c$ **do**
  $\tilde{D}_i = \{(\boldsymbol{x},\tilde{y}) \in \tilde{D} | \tilde{y} = i\}$;
  Rank examples in $\tilde{D}_i$ by $\bar{\ell}(\boldsymbol{x},\tilde{y})$;
  Calculate $num(i)$ according to Eq. (2);
  Select $num(i)$ examples with smallest $\bar{\ell}(\boldsymbol{x},\tilde{y})$ as S_i;
end for
$D_{\text{sel}} = \cup_{i=1}^{c} S_i$;
Train $g(\boldsymbol{x};\Theta)$ with D_{sel};
Output: The final classifier $g(\boldsymbol{x};\Theta)$

Mean loss

Select class by class

Selection number $num(i)$:

Denote the noise rate by η_i and the number of examples for the i-th class by n_i:

- first introduce parameter $\beta \geq 0$ to make $prop(i)$ a little less than $1-\eta_i$:

$$prop(i) = max\{1-(1+\beta)\eta_i, (1-\beta)(1-\eta_i)\}$$

Issue: $[prop(1)\cdot n_1, \cdots prop(c)\cdot n_c]$ may seriously deviate from the true class distribution $[p_1,\cdots,p_c]$.

- set the selected data as $[p_1\cdot m,\dots,p_c\cdot m]$ to obey $[p_1,\cdots,p_c]$:

$$m = \min_{1\leq i\leq c}\{prop(i)\cdot n_i/p_i\} \text{ by constraints } p_i\cdot m \leq prop(i)\cdot n_i$$

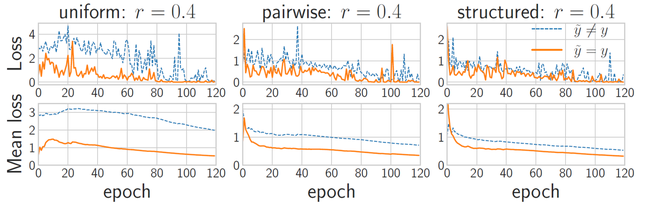
Issue: too many useful data may be wasted.

- thus additionally introduce parameter $\gamma \geq 1$:

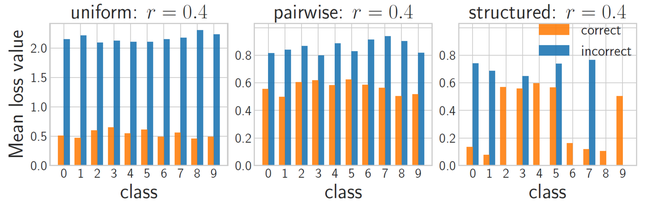
$$num(i) = \min\{\gamma\cdot p_i \times m, prop(i)\times n_i\}$$

Experiments

The stability of mean loss *vs.* single epoch's loss:



The necessity of class-wise sample selection:



More experimental results can be found in the paper.

Conclusion

- We establish the connection between noisy data distribution and the small-loss criterion.
- Then we theoretically explain why the widely-used small-loss criterion works and reformalize the vanilla small-loss criterion.

Our theoretical analysis gives the following insights:

- the empirically diagonally-dominant condition is theoretically justified.
- the loss value for examples with different labels are not comparable so the small-loss level should be determined class by class.
- the warm-up stage is necessary for the small-loss criterion.