$\mathcal{L}_{\text{DMI}}$:

A Novel Information-theoretic Loss Function for Training Deep Nets Robust to Label Noise

Yilun Xu*, Peng Cao*, Yuqing Kong, Yizhou Wang

Peking Univeristy

Deep learning with noisy Label

Crowdsourcing platforms: A potential way to get annotations cheaper and faster.

However, the collected labels are usually very noisy.

Unfortunately, noisy labels hamper performance.

More unfortunately, previous works proposed distance-based loss (e.g. 0-1 loss) are not robust to certain noise patterns.

Information-theoretic loss

Information-theoretic loss: Choose the classifier whose outputs have the highest mutual information with the labels has the lowest loss.

Remark

The meaningless classifier has no information $\rightarrow$ easily eliminated

Information-monotonicity $\rightarrow$ weak classifier less preferred

However, what we actually want with an information measure I:

$$\begin{aligned} &\mathrm{I}(\text{classifier1's output}; \text{noisy labels}) > \mathrm{I}(\text{classifier2's output}; \text{noisy labels}) \\ \Leftrightarrow &\mathrm{I}(\text{classifier1's output}; \text{clean labels}) > \mathrm{I}(\text{classifier2's output}; \text{clean labels}). \end{aligned}$$

Unfortunately, the traditional Shannon mutual information does not satisfy this formula.

Determinant based Mutual Information

Definition (DMI)

Given two discrete random variables W_1, W_2, we define the Determinant based Mutual Information between W_1 and W_2 as

$$\mathrm{DMI}(W_1, W_2) = |\det(\mathbf{Q}_{W_1,W_2})|$$

where $\mathbf{Q}_{W_1,W_2}$ is the joint distribution matrix over W_1 and W_2.

Theorem (Properties of DMI)

DMI is non-negative, symmetric and information-monotone. Moreover, it is relatively invariant: for all random variables W_1, W_2, W_3, when W_3 is less informative for W_2 than W_1, i.e., *W_3 is independent of W_2 conditioning W_1,*

$$\mathrm{DMI}(W_2, W_3) = \mathrm{DMI}(W_2, W_1)|\det(\mathbf{T}_{W_1 \to W_3})|$$

where $\mathbf{T}_{W_1 \to W_3}$ is the matrix $T_{W_1 \to W_3}(w_1, w_3) = \Pr[W_3 = w_3 | W_1 = W_1]$.

DMI satisfies the formula!

The measurement based on noisy labels $\mathrm{DMI}(h(X), \tilde{Y})$ is consistent with the measurement based on clean labels $\mathrm{DMI}(h(X), Y)$, *i.e.*, for every two classifiers h and h',

$$\mathrm{DMI}(h(X), Y) > \mathrm{DMI}(h'(X), Y) \Leftrightarrow \mathrm{DMI}(h(X), \tilde{Y}) > \mathrm{DMI}(h'(X), \tilde{Y}).$$

Definition:

$$\mathcal{L}_{\text{DMI}}(Q_{h(X),\tilde{Y}}) := -\log(\text{DMI}(h(X),\tilde{Y})) = -\log(|\det(\mathbf{Q}_{h(X),\tilde{Y}})|)$$

Property:

$\mathcal{L}_{\text{DMI}}(\text{noisy data; classifier}) = \mathcal{L}_{\text{DMI}}(\text{clean data; classifier}) + \text{noise amount},$

Remark

With $\mathcal{L}_{\text{DMI}}$, training with the noisy labels is theoretically equivalent with training with the clean labels in the dataset, regardless of the noise patterns, including the noise amounts.

Main Theorem

Theorem (Main Theorem)

With two assumptions, $\mathcal{L}_{\text{DMI}}$ *is*

legal *if there exists a ground truth classifier* h^* *such that* $h^*(X) = Y$*, then it must have the lowest loss,* i.e., *for all classifier* h*,*

$$\mathcal{L}_{\text{DMI}}(Q_{h^*(X),\tilde{Y}}) \leq \mathcal{L}_{\text{DMI}}(Q_{h(X),\tilde{Y}})$$

and the inequality is strict when $h(X)$ *is not a permutation of* $h^*(X)$, i.e., *there does not exist a permutation* $\pi : \mathcal{C} \mapsto \mathcal{C}$ s.t. $h(x) = \pi(h^*(x)), \forall x \in \mathcal{X}$*;*

noise-robust *for the set of all possible classifiers* $\mathcal{H}$*,*

$$\arg\min_{h \in \mathcal{H}} \mathcal{L}_{\text{DMI}}(Q_{h(X),\tilde{Y}}) = \arg\min_{h \in \mathcal{H}} \mathcal{L}_{\text{DMI}}(Q_{h(X),Y})$$

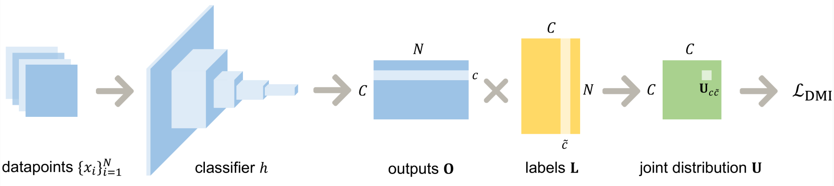
and in fact, training using noisy labels is the same as training using clean labels in the dataset except a constant shift,

$$\mathcal{L}_{\text{DMI}}(Q_{h(X),\tilde{Y}}) = \mathcal{L}_{\text{DMI}}(Q_{h(X),Y}) + \alpha;$$

information-monotone *for every two classifiers* h, h'*, if* $h'(X)$ *is less informative for* Y *than* $h(X)$*, i.e.* $h'(X)$ *is independent of* Y *conditioning on* $h(X)$*, then*

$$\mathcal{L}_{\text{DMI}}(Q_{h(X),\tilde{Y}}) \leq \mathcal{L}_{\text{DMI}}(Q_{h(X),Y}).$$

Implementation



$$\mathcal{L}_{\mathrm{DMI}}(\{(x_i, \tilde{y}_i)\}_{i=1}^N; h) = -\log(|\det(\mathbf{U})|)$$

Experiment

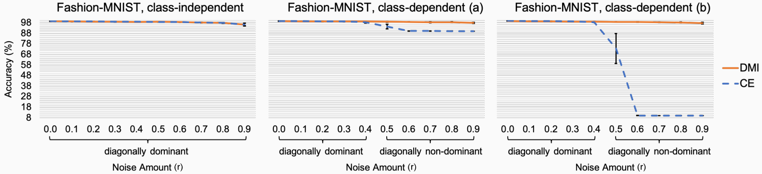


Figure 1: Comparison of distance-based and information-theoretic losses

Table 1: Test accuracy (mean) on real-world dataset Clothing1M

Method	CE	FW	GCE	LCCN	DMI
Accuracy	68.94	70.83	69.09	71.63	**72.46**

Experiment

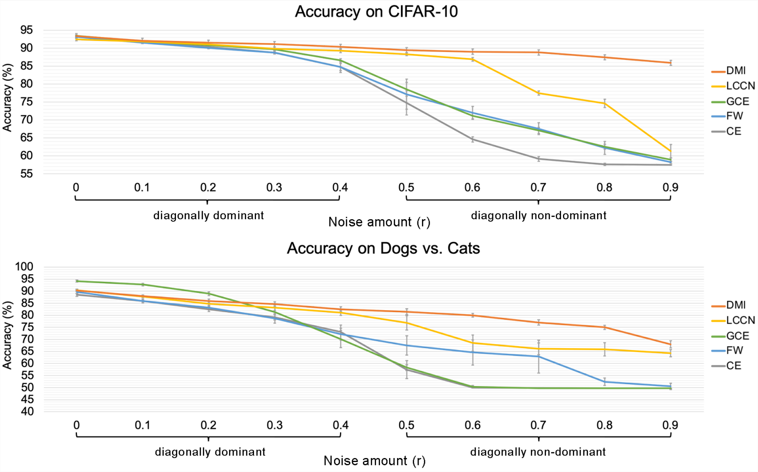


Figure 2: Test accuracy synthesized datasets

Thank you!