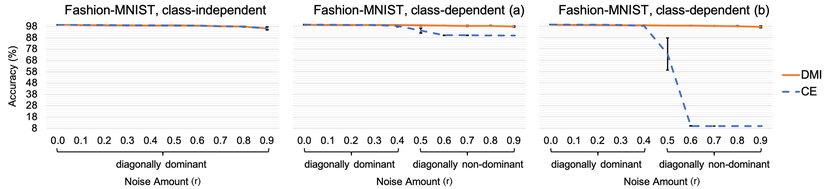
$\mathcal{L}_{\mathrm{DMI}}$: A Novel Information-theoretic Loss Function for Training Deep Nets Robust to Label Noise

Yilun Xu*, Peng Cao*, Yuqing Kong, Yizhou Wang {xuyilun, caopeng2016, yuqing.kong, Yizhou.Wang}@pku.edu.cn

◀ Learning with Noisy Label ▶

- Annotating large scale dataset is notoriously time- and money- consuming
- Crowdsourcing platforms -> acquire low-quality-annotated dataset cheaper and faster
- However, noisy labels damage the performance of DNN without particular treatment

- Previous method: only handle limited kinds of noise patterns, require auxiliary information or steps (e.g., knowing or estimating the noise transition matrix), or lack theoretical justification.
- *Distance-based* loss: diagonally non-dominant noise -> prefer "meaningless classifier"
- *Information-theoretic* loss: "meaningless classifier" has no information about anything



- However, eliminating the "meaningless classifier" is not enough.
- For every two classifiers, in fact, we want an information measure I to satisfy:
  I(classifier 1's output; *noisy* labels) > I(classifier 2's output; *noisy* labels) <=>
  I(classifier 1's output; *clean* labels) > I(classifier 2's output; *clean* labels).
- Unfortunately, Shannon mutual information (MI) *does not* satisfy this formula.

Definition 1 (Determinant based Mutual Information [1]). Given two discrete random variables W_1, W_2, we define the Determinant based Mutual Information (DMI) between W_1 and W_2 as

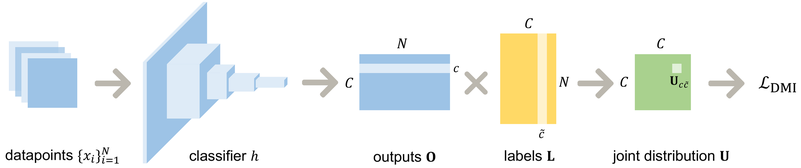
$$\mathrm{DMI}(W_1, W_2) = |\det(\mathbf{Q}_{W_1,W_2})|$$

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

[1] Yuqing Kong. Dominantly truthful multi-task peer prediction, with constant number of tasks. ACM-SIAM Symposium on Discrete Algorithms (SODA20)

- We find a generalized information measure, DMI, which satisfies the formula.
- We propose a noise-robust loss function based on DMI:
  $\mathcal{L}_{\mathrm{DMI}}(\text{data; classifier}) := -\log[\mathrm{DMI}(\text{classifier's output; labels})]$.
- *With $\mathcal{L}_{\mathrm{DMI}}$, 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*, i.e.,
  $\mathcal{L}_{\mathrm{DMI}}(\text{noisy data; classifier}) = \mathcal{L}_{\mathrm{DMI}}(\text{clean data; classifier}) + \text{noise amount}$,
  where the noise amount is a constant given the dataset.

◀ Method Overview ▶



Based on DMI, our loss function is defined as

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

where $Q_{h(X),\tilde{Y}}$ is the joint distribution over $h(X), \tilde{Y}$ and $\mathbf{Q}_{h(X),\tilde{Y}}$ is the $C \times C$ matrix format of $Q_{h(X),\tilde{Y}}$.

In each step of iteration, we sample a batch of datapoints and their noisy labels $\{(x_i, \tilde{y}_i)\}_{i=1}^N$. We denote the outputs of the classifier by a matrix $\mathbf{O}$ and the noisy labels by a 0-1 matrix $\mathbf{L}$. We define $\mathbf{U} = \frac{1}{N}\mathbf{OL}$. The empirical loss function is defined as

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

◀ Theoretical justification ▶

Theorem 3 (Main Theorem). If every datapoint X is independent of its noisy label $\tilde{Y}$ conditioning on the ground truth label Y, and the transition matrix between the noise labels and ground truth labels $\mathbf{T}_{Y\to\tilde{Y}}$ is invertible, i.e., $\det(\mathbf{T}_{Y\to\tilde{Y}}) \neq 0$, $\mathcal{L}_{\mathrm{DMI}}$ has following properties:

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

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

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 $\boldsymbol{\mathcal{H}}$,

$$\mathrm{argmin}_{h\in\mathcal{H}}\mathcal{L}_{\mathrm{DMI}}(Q_{h(X),\tilde{Y}}) = \mathrm{argmin}_{h\in\mathcal{H}}\mathcal{L}_{\mathrm{DMI}}(Q_{h(X),Y})$$

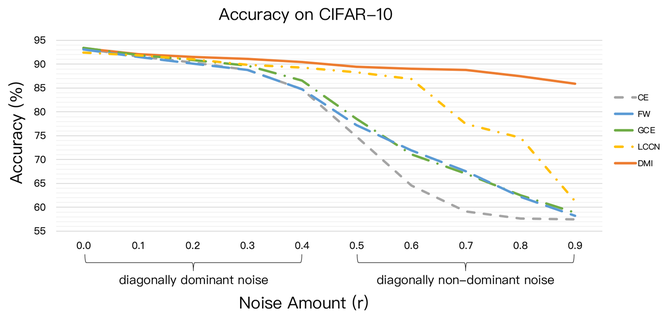
and in fact, training using noisy labels is the same as training using clean labels in the dataset except a constant shift,

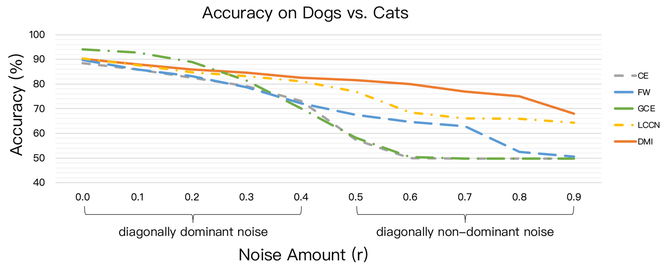
$$\mathcal{L}_{\mathrm{DMI}}(Q_{h(X),\tilde{Y}}) = \mathcal{L}_{\mathrm{DMI}}(Q_{h(X),Y}) + \alpha;$$

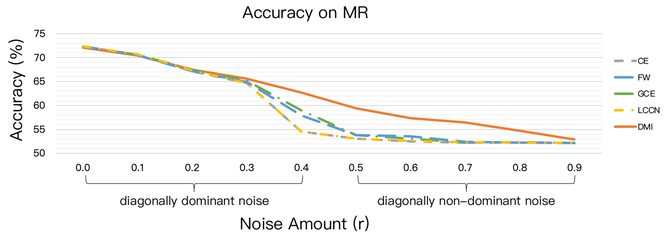
Information-monotone: for every two classifiers $\boldsymbol{h}, \boldsymbol{h}'$, if $\boldsymbol{h}'(\boldsymbol{X})$ is less informative for $\boldsymbol{Y}$ than $\boldsymbol{h}(\boldsymbol{X})$, i.e. $\boldsymbol{h}'(\boldsymbol{X})$ is independent of $\boldsymbol{Y}$ conditioning on $\boldsymbol{h}(\boldsymbol{X})$,

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

◀ New State-of-the-art Results ▶







Accuracy on Clothing1M

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

code link

paper link