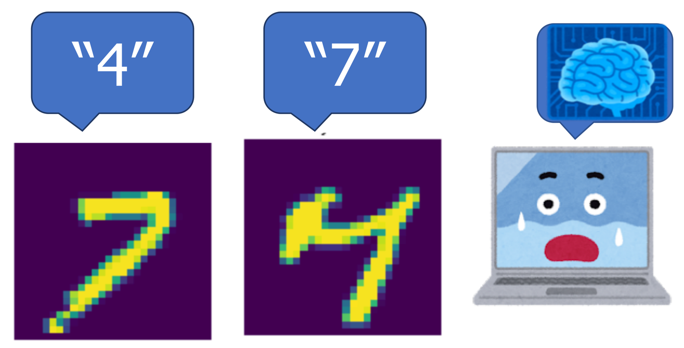
Data Cleansing for Models Trained with SGD

Satoshi Hara (Osaka Univ.), Atsushi Nitanda (Tokyo Univ./RIKEN AIP), Takanori Maehara (RIKEN AIP)

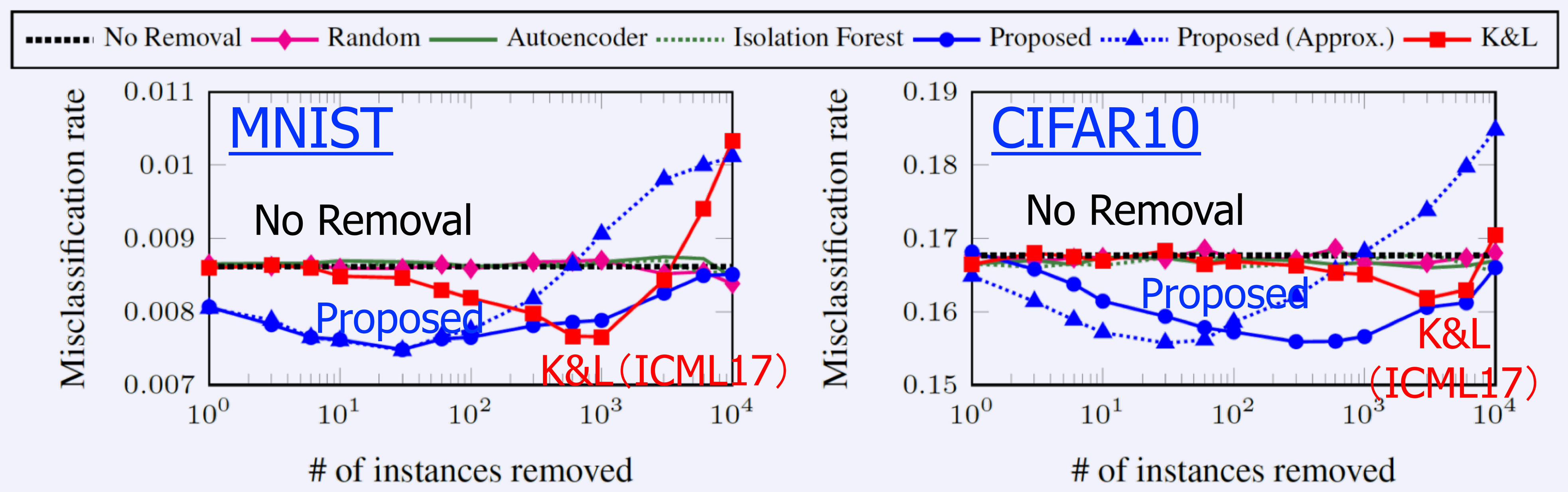
Data Cleansing

Remove "harmful" training instances,
and improve the model's accuracy.

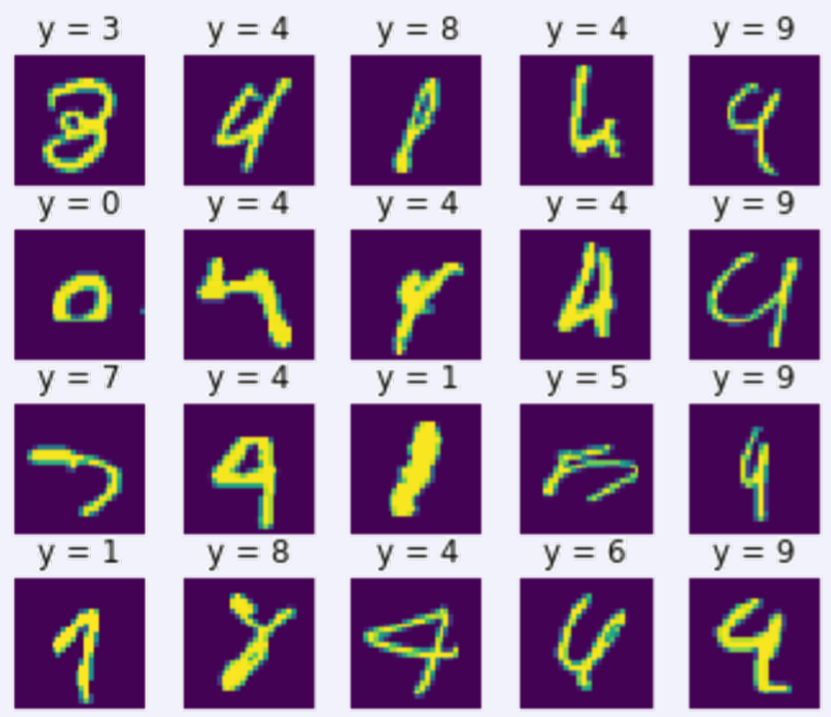


The Results on MNIST/CIFAR10

By removing "harmful" instances, the trained CNN became more accurate.



Top-20 "harmful" instances



Algorithm

estimated validation loss

for $k = 1, 2, \ldots, N$
  Compute $r_k = \langle u, \Delta\theta_{-k} \rangle$
Identify "harmful" instance:
  $\hat{k} = \operatorname{argmin}_k r_k$
Remove "harmful" instance:
  $D \leftarrow D \setminus \{z_{\hat{k}}\}$

Theoretical Analysis

Convex Case

Our Estimator

$$\left\| \left(\theta_{-k}^{[T]} - \theta^{[T]}\right) - \Delta\theta_{-k} \right\| \le \sqrt{2(h_k(\lambda)^2 + h_k(\Lambda)^2)}$$

$$h_k(a) := \frac{\eta_{\pi(k)}}{|S_{\pi(k)}|} \prod_{t=\pi(k)+1}^{T-1} (1 - \eta_t a) \|\nabla_\theta \ell(z_k; \theta^{[\pi(k)]})\|$$

Assumption
- The loss ℓ is strongly convex, smooth, and twice differentiable.
- $\exists \lambda, \Lambda > 0$ such that $\lambda I \preccurlyeq \nabla_\theta^2 \ell(z; \theta) \preccurlyeq \Lambda I$.

Non-Convex Case

Our Estimator

$$\left\| \left(\theta_{-k}^{[T]} - \theta^{[T]}\right) - \Delta\theta_{-k} \right\| \le \frac{\gamma^2 T G^2 L}{\Lambda} \exp^{O(\gamma \Lambda \sqrt{T})}$$

Assumption
- The Hessian of the loss ℓ is Lipschitz.
- $\|\nabla_\theta \ell(z; \theta)\| \le G, \nabla_\theta^2 \ell(z; \theta) \preccurlyeq \Lambda I$.
- $\eta = O(\gamma/\sqrt{T})$

The Proposed Estimator

Estimation of "harmful" instances

Training Data D

$$\hat{\theta} = \operatorname*{argmin}_\theta \sum_{z \in D} \ell(z; \theta)$$

2% error!

One instance missing $D \setminus \{z_k\}$

$$\hat{\theta}_{-k} = \operatorname*{argmin}_\theta \sum_{z \in D \setminus \{z_k\}} \ell(z; \theta)$$

1% error!!

z_k is "harmful".

Problem
Find z_k that minimizes the validation loss.

$$\sum_{z \in D_V} \left(\ell(z; \hat{\theta}_{-k}) - \ell(z; \hat{\theta})\right) \approx \langle u, \hat{\theta}_{-k} - \hat{\theta} \rangle$$

Small validation loss ≈ Small inner product

- $u := \sum_{z \in D_V} \nabla_\theta \ell(z; \hat{\theta})$ for the val. set D_V.
- No model retraining.

Our Contributions

- To estimate $\hat{\theta}_{-k} - \hat{\theta}$, the current methods require "the optimal $\hat{\theta}$" and "convex loss ℓ".
  → Too restrictive for modern ML.
- Our estimator assumes that "the model is trained with SGD".
  → More appropriate for modern ML.

The Proposed Estimator

- $\theta^{[t+1]} \leftarrow \theta^{[t]} - \frac{\eta_t}{|S_t|} \sum_{z \in S_t} \nabla_\theta \ell(z; \theta^{[t]}) =: h_t(\theta^{[t]})$ — Ordinary SGD
- $\theta_{-k}^{[t+1]} \leftarrow \theta_{-k}^{[t]} - \frac{\eta_t}{|S_t|} \sum_{z \in S_t \setminus \{z_k\}} \nabla_\theta \ell(z; \theta^{[t]})$ — SGD with missing z_k

Our Estimator $\Delta\theta_{-k} \approx \theta_{-k}^{[T]} - \theta^{[T]}$

"Backprop" through SGD

$$\theta^{[t]} \xrightarrow{h_t} \theta^{[t+1]} \xrightarrow{h_{t+1}} \cdots \xrightarrow{h_{T-2}} \theta^{[T-1]} \xrightarrow{h_{T-1}} \theta^{[T]}$$

$$\frac{\eta_t}{|S_t|} \nabla_\theta \ell(z_k; \theta^{[t]})$$

$$\theta_{-k}^{[T]} - \theta^{[T]} \approx Z_{T-1}\left(\theta_{-k}^{[T-1]} - \theta^{[T-1]}\right) = Z_{T-1} Z_{T-2} \ldots Z_{t+1}\left(\theta_{-k}^{[t+1]} - \theta^{[t+1]}\right)$$

$$Z_{t+1} := \frac{\partial \theta^{[t+2]}}{\partial \theta^{[t+1]}}$$

Computing $r_k = \langle u, \Delta\theta_{-k} \rangle$ for all $k = 1, 2, \ldots, N$.

for $t = T-1, T-2, \ldots, 1$
  $r_j \leftarrow r_j + \left\langle u, \frac{\eta_t}{|S_t|} \nabla_\theta \ell(z_j; \theta^{[t]}) \right\rangle, \forall j \in S_t$
  $u \leftarrow Z_t u$

$O(|S_t| T)$ time in total