Data Cleansing for Models Trained with SGD

NIPS2019

Introduction

Problem 1 (Data Cleansing)**.** Find a subset of the training instances such that the trained model obtained after removing the subset has a better accuracy.

SGD Let $g(z,\theta) \coloneqq \nabla_\theta l(z;\theta)$.

$\theta^{[t+1]} \leftarrow \theta^{[t]} - \frac{\eta_{\mathrm{t}}}{|S_t|} \sum_{i \in S_t} g(z_i; \theta^{[t]})$,

where S_t denotes the set of instance indices used in the t-th step, and η_t is the learning rate.

Introduction

Definition 3 (SGD-Influence)**.** We refer to the parameter difference $\theta_{-j}^{[t]} - \theta^{[t]}$ as the *SGD-influence* of the instance $z_j \in D$ at step t.

$$\theta_{-j}^{[t+1]} \leftarrow \theta_{-j}^{[t]} - \frac{\eta_t}{|S_t|} \sum_{i \in S_t \setminus \{j\}} g(z_i; \theta_{-j}^{[t]})$$

Influence Function[ICML2017 Understanding black-box predictions via influence functions]

$$\hat{\theta}_{-j} - \hat{\theta} \approx \frac{1}{N} \hat{H}^{-1} \nabla_\theta \ell(z_j; \hat{\theta}) \quad \hat{\theta}_{-j} = \text{argmin}_\theta \sum_{n=1; n \neq j}^{N} \ell(z; \theta)$$

Problem 4 (Linear Influence Estimation (LIE))**.** For a given query vector $u \in \mathbb{R}^p$, estimate the *linear influence* $L_{-j}^{[T]}(u) := \langle u, \theta_{-j}^{[T]} - \theta^{[T]} \rangle$.

Proposed Method

One epoch SGD-Influence

$$\theta_{-j}^{[t]} - \theta^{[t]} = (\theta_{-j}^{[t-1]} - \theta^{[t-1]}) - \frac{\eta_{t-1}}{|S_{t-1}|} \sum_{i \in S_{t-1}} (\nabla_\theta \ell(z_i; \theta_{-j}^{[t-1]}) - \nabla_\theta \ell(z_i; \theta^{[t-1]}))$$
$$\approx (I - \eta_{t-1} H^{[t-1]})(\theta_{-j}^{[t-1]} - \theta^{[t-1]}).$$

$$\text{where} \quad H^{[t]} := \frac{1}{|S_t|} \sum_{i \in S_t} \nabla_\theta^2 \ell(z_i; \theta^{[t]})$$

Proposed Method

Let $Z_t := I - \eta_t H^{[t]}$ and $\pi(j)$ be the SGD step where the instance z_j is used.

$$\theta_{-j}^{[\pi(j)+1]} - \theta^{[\pi(j)+1]} = \frac{\eta_{\pi(j)}}{|S_{\pi(j)}|} g(z_j; \theta^{[\pi(j)]})$$

$$\theta_{-j}^{[T]} - \theta^{[T]} \approx \frac{\eta_{\pi(j)}}{|S_{\pi(j)}|} Z_{T-1} Z_{T-2} \cdots Z_{\pi(j)+1} g(z_j; \theta^{[\pi(j)]}) =: \Delta\theta_{-j}.$$

Proposed Method

K epoch SGD-Influence

$$\Delta\theta_{-j} = \sum_{k=1}^{K} \left(\prod_{s=1}^{T-\pi_k(j)-1} Z_{T-s}\right) \frac{\eta_{\pi_k(j)}}{|S_{\pi_k(j)}|} g(z_j; \theta^{[\pi_k(j)]})$$

Query vector

$$u = \frac{1}{|D'|} \sum_{z' \in D'} \nabla_\theta \ell(z'; \theta^{[T]})$$

LIE

$$\text{Let } u^{[t]} := Z_{t+1} Z_{t+2} \dots Z_{T-1} u.$$

$$\langle u, \Delta\theta_{-j} \rangle = \sum_{k=1}^{K} \langle u^{[\pi_k(j)]}, \frac{\eta_{\pi_k(j)}}{|S_{\pi_k(j)}|} g(z_j; \theta^{[\pi_k(j)]}) \rangle$$

$$u^{[t]} \leftarrow Z_{t+1} u^{[t+1]} = u^{[t+1]} - \eta_{t+1} H_{\theta^{[t+1]}} u^{[t+1]}$$

Proposed Method

Algorithm 1 LIE for SGD: Training Phase

Initialize the parameter $\theta^{[1]}$
Initialize the sequence as null: $A \leftarrow \emptyset$
for $t = 1, 2, \ldots, T-1$ **do**
  $A[t] \leftarrow (S_t, \eta_t, \theta^{[t]})$ *// store information*
  $\theta^{[t+1]} \leftarrow \theta^{[t]} - \frac{\eta_t}{|S_t|} \sum_{i \in S_t} g(z_i; \theta^{[t]})$
end for

Algorithm 2 LIE for SGD: Inference Phase

Require: $u \in \mathbb{R}^p$
Initialize the influence: $\hat{L}_{-j}^{[T]}(u) \leftarrow 0, \forall j$
for $t = T-1, T-2, \ldots, 1$ **do**
  $(S_t, \eta_t, \theta^{[t]}) \leftarrow A[t]$ *// load information*
  // update the linear influence of z_j
  $\hat{L}_{-j}^{[T]}(u) \mathrel{+}= \langle u, \frac{\eta_t}{|S_t|} g(z_j; \theta^{[t]}) \rangle, \forall j \in S_t$
  $u \mathrel{-}= \eta_t H^{[t]} u$ *// update* u
end for

Experiments

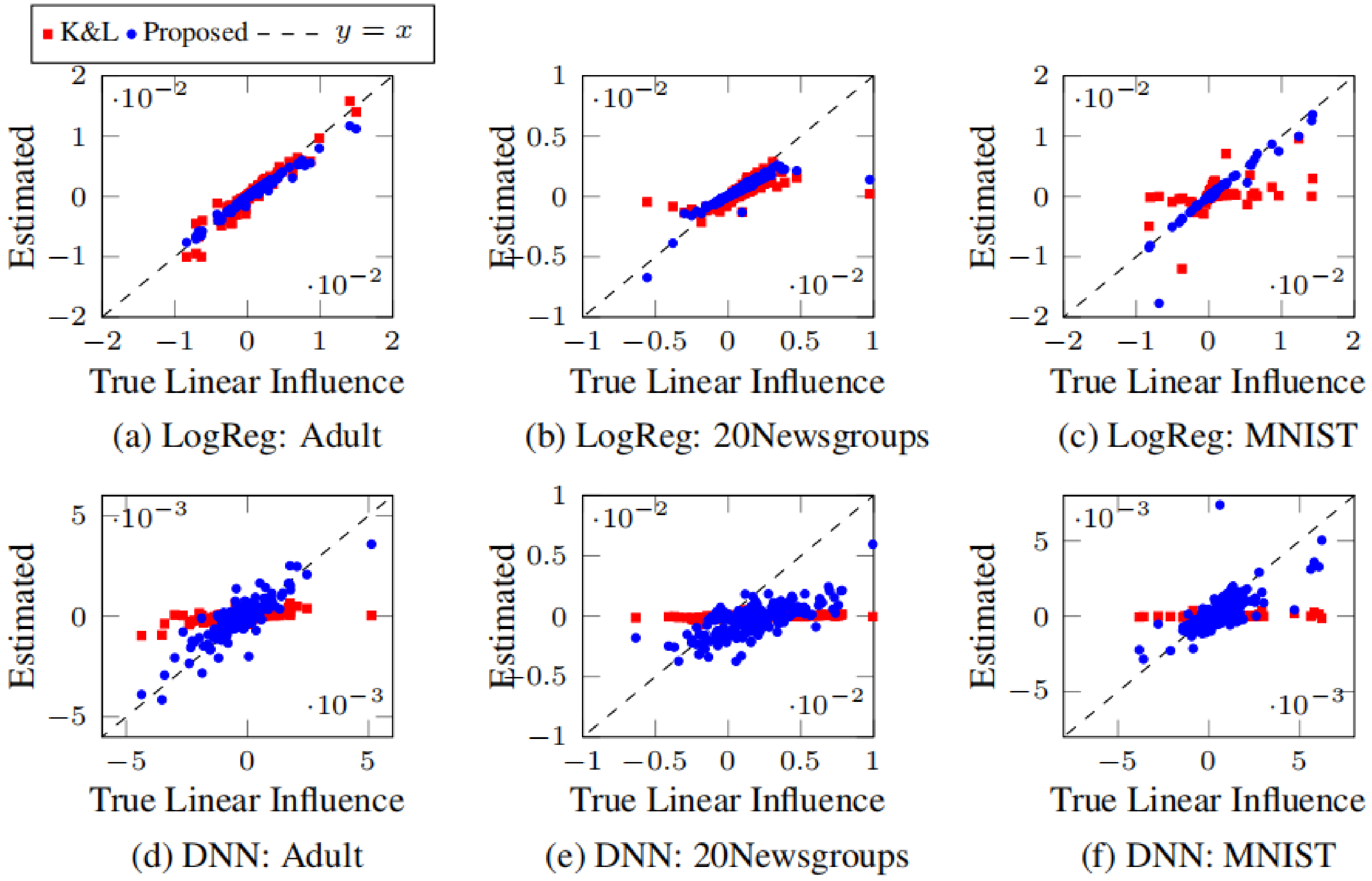
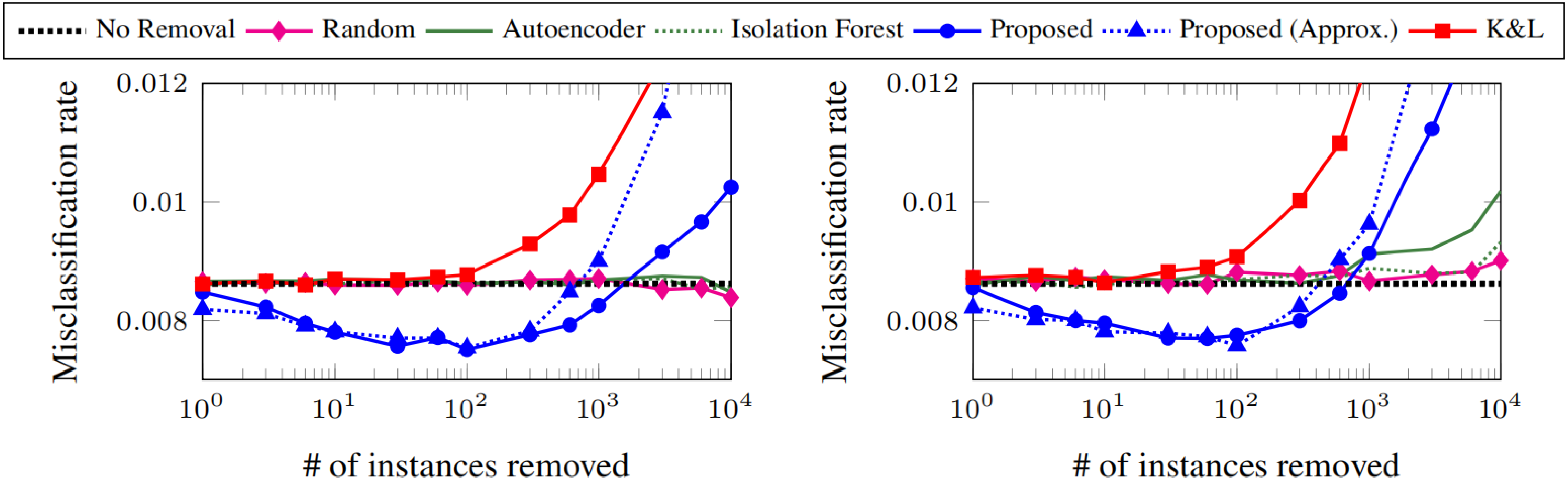


Figure 1: Estimated linear influences for linear logistic regression (LogReg) and deep neural networks (DNN) for all the 200 training instances. K&L denotes the method of Koh and Liang [2017].

Experiments



(a) MNIST: Retrain Last (b) MNIST: Retrain All

Figure 6: MNIST: Average misclassification rates on the test set after data cleansing over 30 experiments.

Experiments

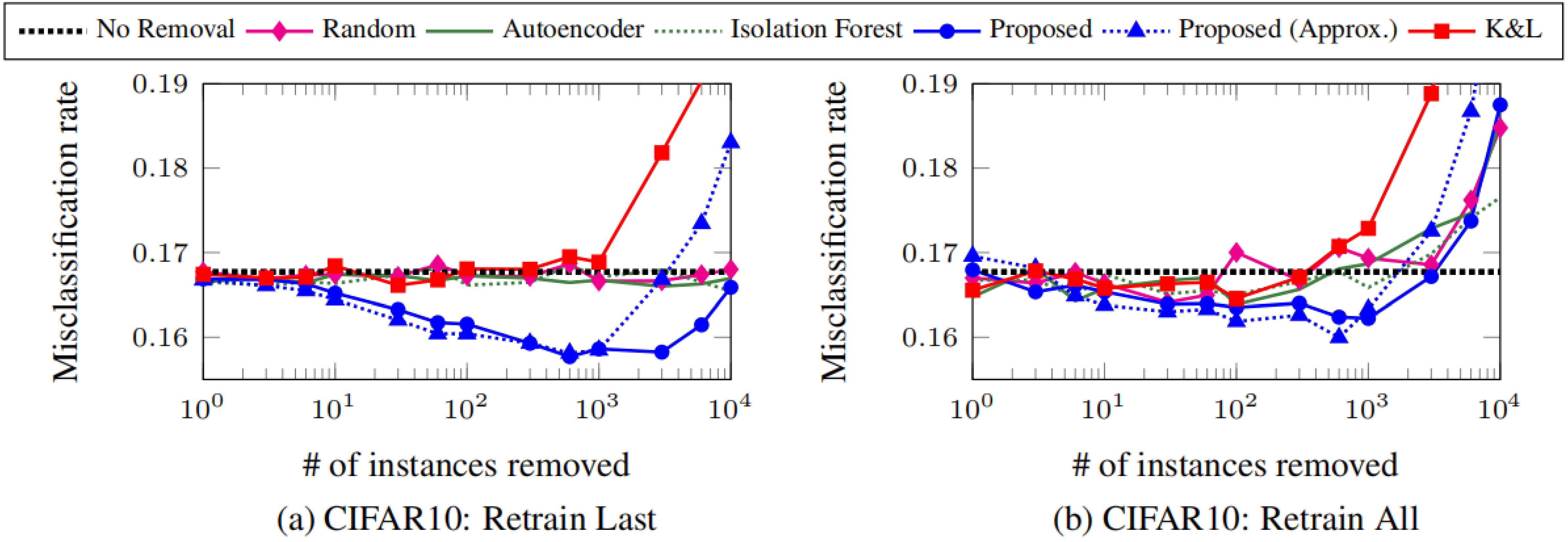
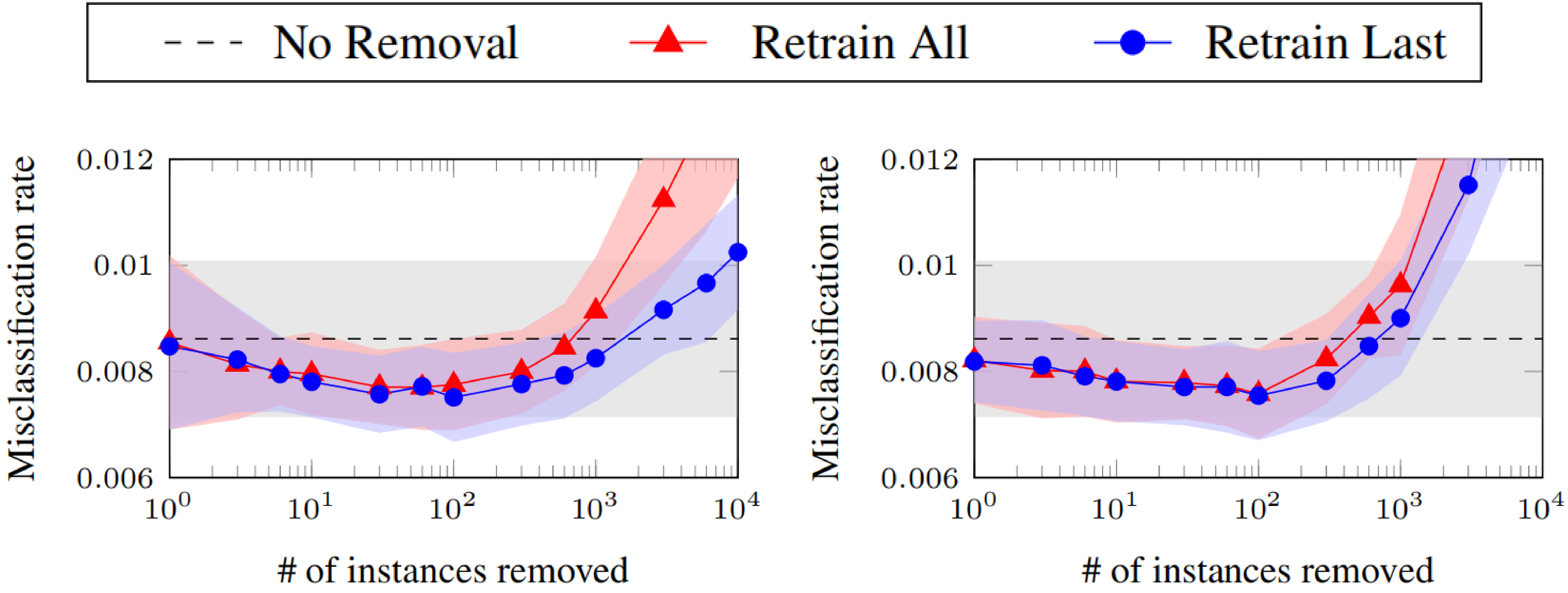


Figure 8: CIFAR10: Average misclassification rates on the test set after data cleansing over 30 experiments.

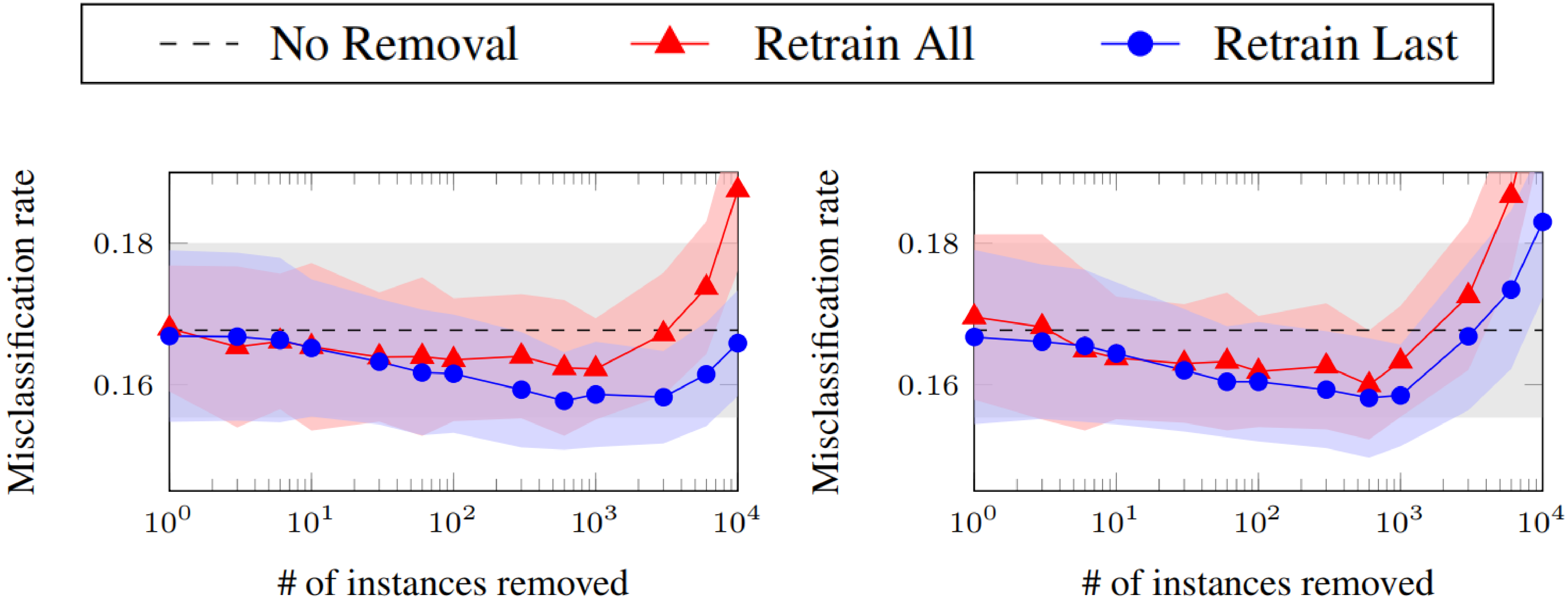
Experiments



(e) Proposed

(f) Proposed (Approx.)

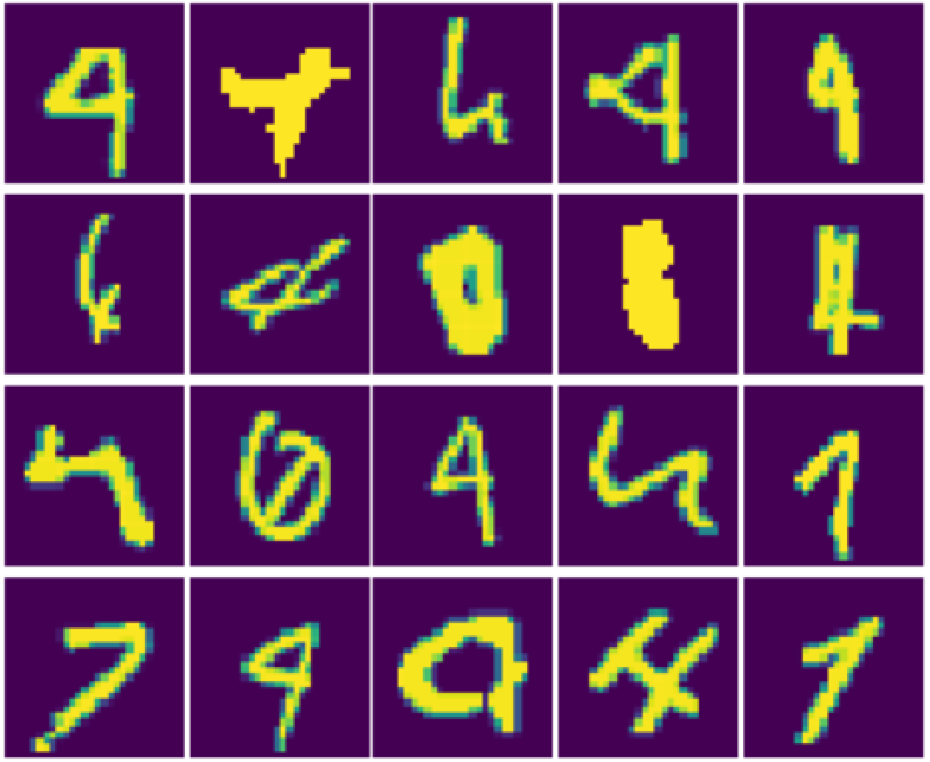
Experiments



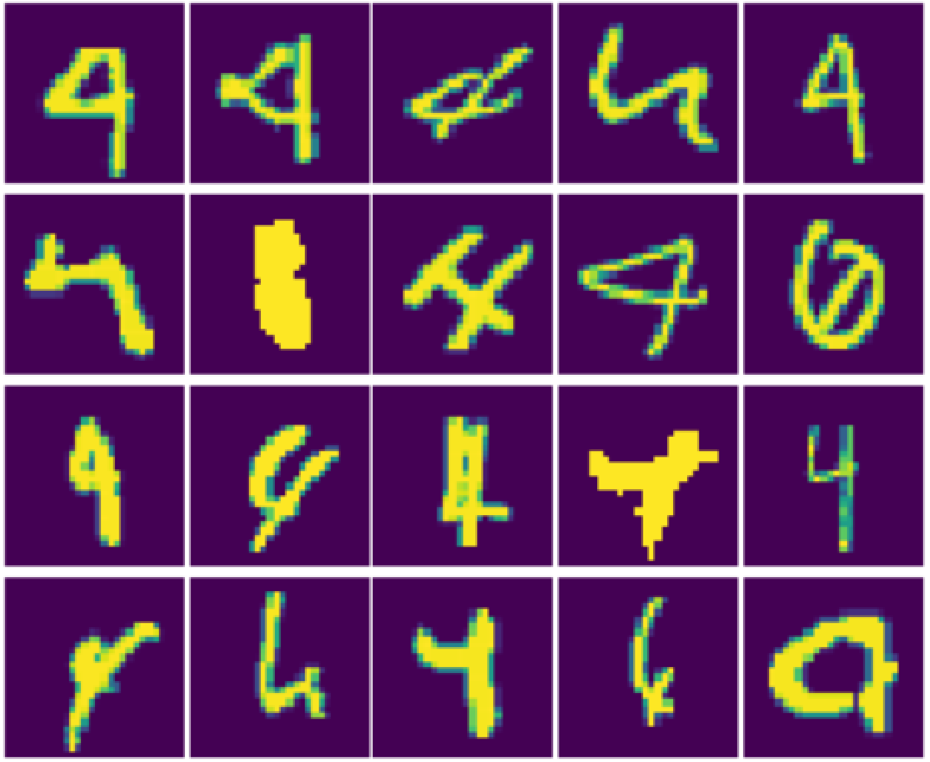
(e) Proposed

(f) Proposed (Approx.)

Experiments

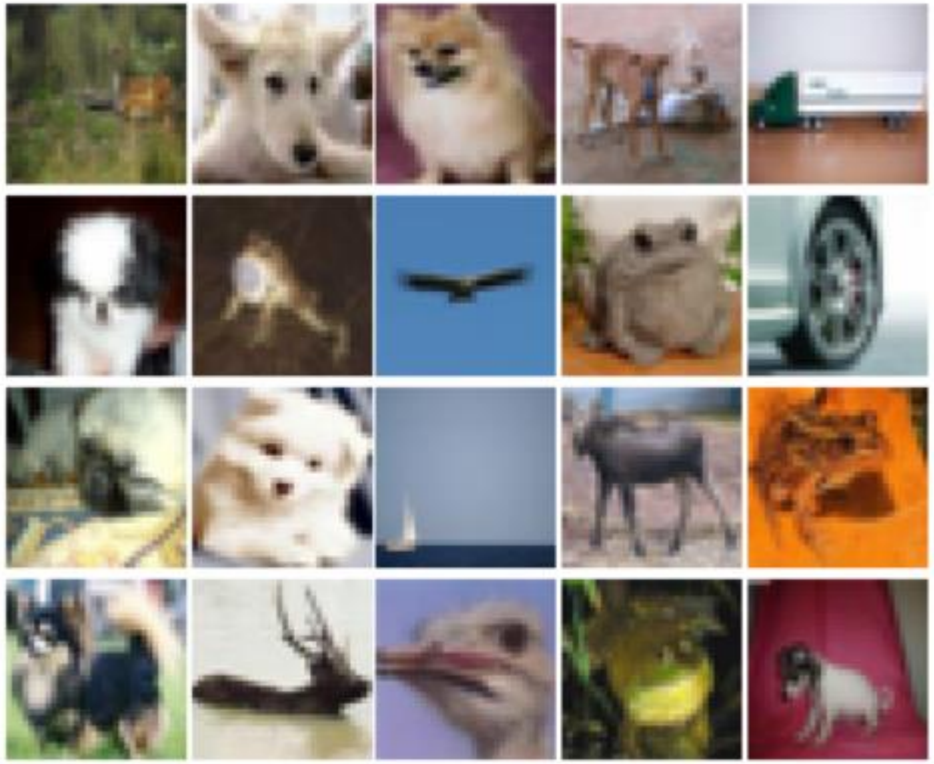


(c) Proposed



(d) Proposed (Approx.)

Experiments



(c) Proposed

(d) Proposed (Approx.)

Thanks