

Learning with Bad Training Data via Iterative Trimmed Loss Minimization

Yanyao Shen and Sujay Sanghavi

Department of ECE, The University of Texas at Austin

*In Proceedings of the 36th International Conference on
Machine Learning, ICML 2019*

Abstract

- 訓練データの一部が破損している場合の学習に対するフレームワークの提案
- 訓練データに綺麗なデータ (**clean samples**) と悪いデータ (**bad samples**) が含まれている場合に訓練データの精度を向上させるのは難しい

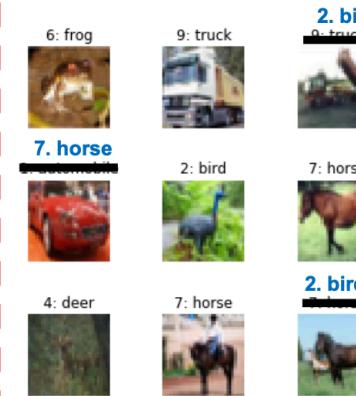


- (a)一度全サンプルで学習して現在のlossが最も低いサンプルを選択
 - (b)それらサンプルのみで再学習
-
- この手順に従うことで一般線形モデルにおいて、最適解に達成することを証明する
 - 実験的にも提案手法の有効性を証明する
 - (a)ラベルのみ誤っている深層学習分類機
 - (b)悪いデータが含まれたGAN
 - (c)敵対的に組みを持つ深層画像分類機

Introduction

- 機械学習モデルにおいてSOTAな性能を達成するには以上に大きな訓練データを用いた大規模な学習を必要とする.
- 機械学習モデルは、使用される訓練データの質に非常に敏感である。
データの質が悪い場合、優れた結果を得ることができないことが多い。
<例>CIFAR 10で自動車の30%を飛行機としてラベリングするとWideResNet-16で学習したとき、精度が90%→70%に低下

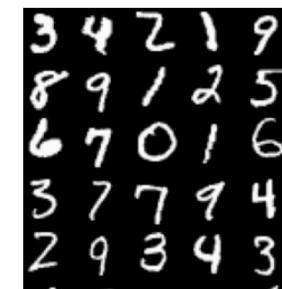
1: Bad Training Labels in Classification
Supervised: noise in training labels makes classifiers inaccurate



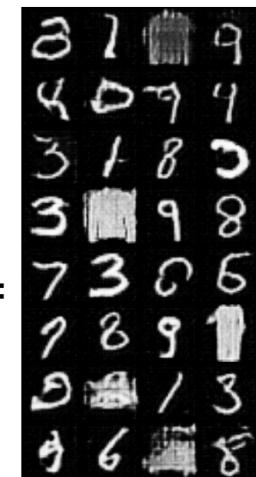
Systematic label noise:
a fraction of "horse" is mis-labeled "bird"

Dataset size will not rescue ...

2: Mixed Training Data
Unsupervised: spurious samples give bad generative models



+ GAN =



3: Backdoor Attacks



Images classified as 'ship' Images classified as 'horse'

Introduction

本論文では、教師あり学習での場合では、ラベリングの誤り、教師なし学習での場合では不適切なデータが含まれるような十分に精査されていないデータセットに興味がある

汚染された訓練データでも対処できる以下のような簡単な手法を提案する

モデルの学習の際に

1. 全てのサンプルで学習する
2. 大きな損失を持つサンプルを除去
3. 残ったサンプルで再度学習する

右の図は, clean sampleとbad sampleを含んだ訓練データをWideResNet-16で学習した結果を示している(Figure 1).

訓練データに破損がある状況で,
学習が進むとbad sampleの方が精度が高くなることもあるが,
初期のエポックでは, clean sample
の精度がbad sampleの精度よりも
高いことがわかる.

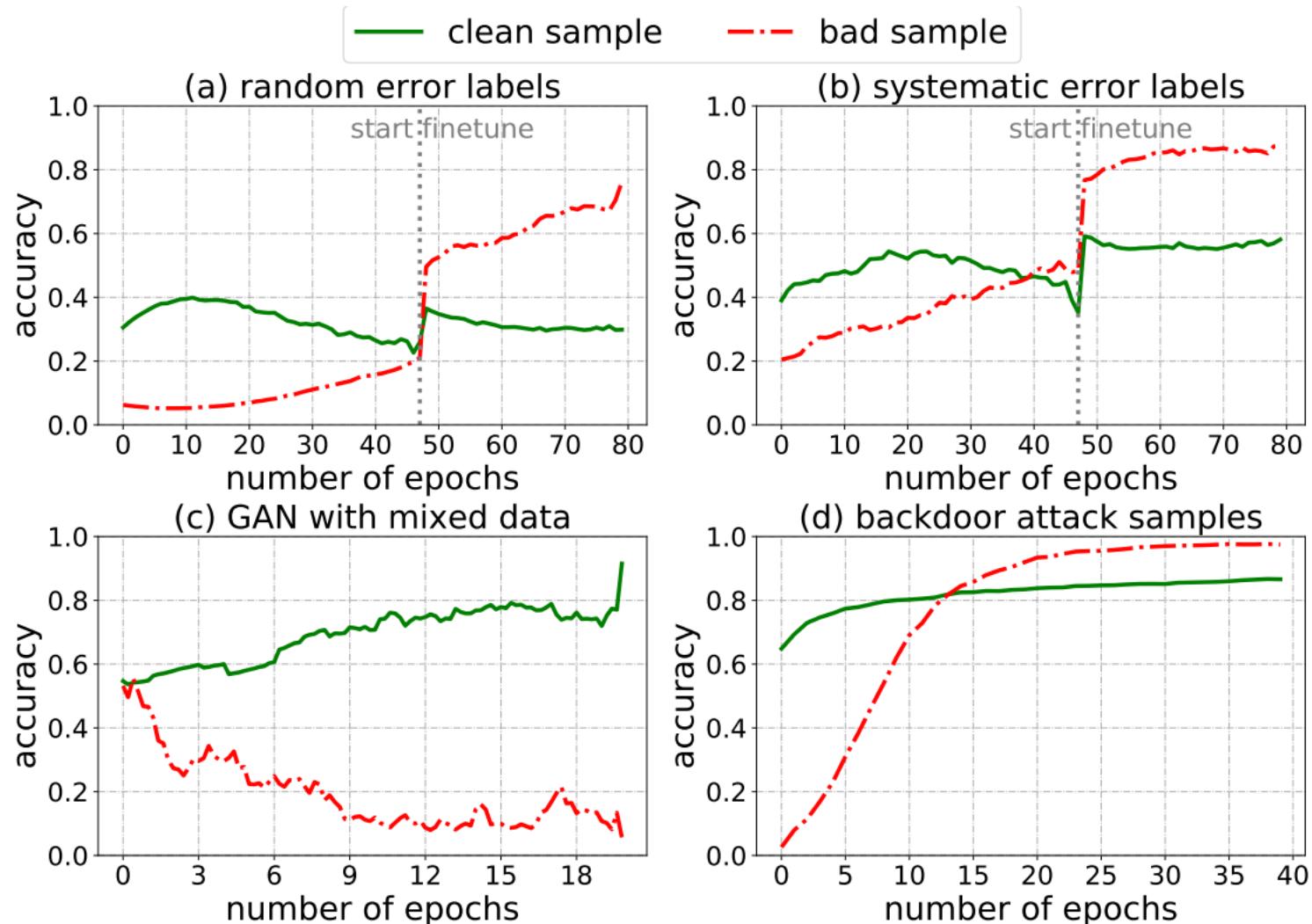


Figure 1: **Observation:** Evolution of model accuracy for clean and bad samples, as a function of training epochs, for four different tainted data settings: (a) classification for CIFAR-10 with 40% random errors in labels, (b) classification for CIFAR-10 with 40% systematic errors in labels, (c) DC-GAN trained on unlabeled mixture of 70% MNIST images with 30% Fashion-MNIST images, (d) backdoor attack on classification for CIFAR-10 with 250 watermarked backdoor attack samples as described in [TLM18]. The CIFAR-10 classifications are done using the WideResNet-16 of [ZK16]. In all instances models are trained on the respective tainted data. **Early on, models are more accurate on the good samples.**

Trimmed loss

- **trimmed loss**を最小化することが目的

n 個のサンプルがあるとき, その中から αn 個のサンプルを選択して,
選ばれた部分集合の損失を最小化する



- trimmed lossを最小化させるために
反復的な方法**Iterative Trimmed Loss Minimization** (ITLM)を提案

Contribution

汚染されたデータを用いた学習に対する一般的なアプローチとしてITLMを提案し、その性能を理論的にも実験的にも調査したこと

- (a) clean samplesが一般線形モデルからくる場合、ITLMは、少なくとも線形で真のモデルに収束することを示した。
- (b) ITLMはbad labelsが含まれる分類問題で適用できることを実験的に示した。
random labelを含むCIFAR-10の分類問題において、ITLMは既存のSOTAな結果を上回った。
- (c) bad imagesを含む画像生成タスクにもITLMを適用することに成功した。

Related Work

- Robust regression

線形回帰設定におけるtrimmed loss最小化を行う研究は, [Hos95、RVD06、SSvdHT13] で行われている。しかしながら, 他の研究と比較して, 損失関数の選択に影響を受けにくい。また, Robust regressionに関する最近の別の研究では, 低次元[DKK + 18、PSBR18、KKM18]と高次元[CCM13、BDLS17、LSC18]の両方の設定で, 敵対者が入力と出力の両方を書き換える強力なロバストネスを検討している。これらのアルゴリズムは通常, 例えば本論文で考察するアルゴリズムと比較してはるかに多くの計算を必要とする。

- Noisy label problems

ノイズの多いラベルを含んだ分類も非常に興味深いものである。[RZYU18]は, DNN学習中にcleanな検証データを大いに参照することにより, 重みの再学習を行うアルゴリズムを提案した。

Setup and (Exact) Trimmed Loss Estimator

Trimmed loss estimator

$$\hat{\theta}^{(\text{TL})} = \arg \min_{\theta \in \mathcal{B}} \min_{S: |S|=\lfloor \alpha n \rfloor} \sum_{i \in S} f_\theta(s_i).$$

sample : s_1, \dots, s_n

model parameters : θ

loss function : $f_\theta(\cdot)$

$\hat{\theta}^{(TL)}$ を見つけるためにサイズ $\lfloor \alpha n \rfloor$ の部分集合 S と θ を調整する必要がある

Assumption

- 以下を仮定する

θ^* はtrimmed lossの期待誤差のglobal minimumである

Assumption 1 (Identification condition for θ^*). *For every $\epsilon > 0$ there exists a $\delta > 0$ such that if $\theta \in \mathcal{B} \setminus \mathcal{U}(\theta^*, \epsilon)$, we have that $F(\theta) - F(\theta^*) > \delta$.*

一般的な仮定

Assumption 2 (Regularity conditions). *D_θ is absolutely continuous for any $\theta \in \mathcal{B}$. d_θ is bounded uniformly in $\theta \in \mathcal{B}$, and is locally positive in a neighborhood of its α -quantile. $f_\theta(s)$ is differentiable in θ for $\theta \in \mathcal{U}(\theta^*, \epsilon)$, for some $\epsilon > 0$.*

\mathcal{B} : コンパクトなパラメータ空間

全てのサンプルはi.i.d

D_θ : f_θ の分布関数

d_θ : f_θ の密度関数

$S(\theta) = \mathbb{E}_s[f_\theta(s)]$: 期待誤差

$S_n(\theta) = \frac{1}{n} \sum_{i=1}^n f_\theta(s_i)$: 経験誤差

$F(\theta) := \mathbb{E}[f_\theta(s)I(f_\theta(s) \leq D_\theta^{-1}(\alpha))]$: trimmed lossの期待誤差

$\mathcal{U}(\theta, \epsilon) := \{\tilde{\theta} \mid |S(\tilde{\theta}) - S(\theta)| < \epsilon, \tilde{\theta} \in \mathcal{B}\}$: 期待誤差の差が ϵ 以内であるパラメータ集合

Lemma 3

Assumption 1 (Identification condition for θ^*). *For every $\epsilon > 0$ there exists a $\delta > 0$ such that if $\theta \in \mathcal{B} \setminus \mathcal{U}(\theta^*, \epsilon)$, we have that $F(\theta) - F(\theta^*) > \delta$.*

Assumption 2 (Regularity conditions). *D_θ is absolutely continuous for any $\theta \in \mathcal{B}$. d_θ is bounded uniformly in $\theta \in \mathcal{B}$, and is locally positive in a neighborhood of its α -quantile. $f_\theta(s)$ is differentiable in θ for $\theta \in \mathcal{U}(\theta^*, \epsilon)$, for some $\epsilon > 0$.*

$$\hat{\theta}^{(\text{TL})} = \arg \min_{\theta \in \mathcal{B}} \min_{S: |S| = \lfloor \alpha n \rfloor} \sum_{i \in S} f_\theta(s_i).$$



$\hat{\theta}^{(TL)}$ の経験誤差は $n \rightarrow \infty$ で θ^* の経験誤差と一致する

Lemma 3. *Under Assumptions 1 and 2, the estimator $\hat{\theta}^{(\text{TL})}$ satisfies: $|S_n(\hat{\theta}^{(\text{TL})}) - S_n(\theta^*)| \rightarrow 0$ with probability 1, as $n \rightarrow \infty$.*

Iterative Trimmed Loss Minimization

Algorithm 1 Iterative Trimmed Loss Minimization (ITLM)

- 1: **Input:** Samples $\{s_i\}_{i=1}^n$, number of rounds T , fraction of samples α
- 2: **(Optional) Initialize:** $\theta_0 \leftarrow \arg \min_{\theta} \sum_{i \in [n]} f_{\theta}(s_i)$
- 3: **For** $t = 0, \dots, T - 1$ **do**
- 4: Choose samples with smallest current loss f_{θ_t} :

$$S_t \leftarrow \arg \min_{S: |S|=\lfloor \alpha n \rfloor} \sum_{i \in S} f_{\theta_t}(s_i)$$

- 5: $\theta_{t+1} = \text{ModelUpdate}(\theta_t, S_t, t)$

- 6: **Return:** θ_T



通常のミニバッチSGDを M 回繰り返す

Algorithm 2 BatchSGD_ModelUpdate(θ, S, t)

- 1: **Input:** Initial parameter θ , set S , round t
- 2: **Choose:** Step size η , number M of gradient steps, batch size N
- 3: **(Optional)** Re-initialize θ^0 randomly
- 4: **For** $j = 1, \dots, M$ **do**
- 5: $B_j \leftarrow \text{random_subset}(S, N)$
- 6: $\theta^j \leftarrow \theta^{j-1} - \eta \left(\frac{1}{N} \sum_{i \in B_j} \nabla_{\theta} f_{\theta^{j-1}}(s_i) \right)$
- 7: **Return:** θ^M

Theoretical Guarantees for Generalized Linear Models

- 誤差を含む一般線形モデルについてITLMを分析する
サンプル (x, y) は以下のように与えられる

$$\begin{aligned} y &= \omega(\phi(x)^\top \cdot \theta^*) + e, && (\text{clean samples}) \\ y &= r + e, && (\text{bad samples}) \end{aligned} \tag{1}$$

- 二乗誤差を考える

$$f_\theta(x, y) = (y - \omega(\phi(x)^\top \cdot \theta))^2$$

x : input
 y : output
 ϕ : embedding function
 w : link function
 r : arbitrary, random corruptions
 e : random subgaussian noise with parameter σ^2
 θ^* : ground truth
 α^* : the fraction of clean samples

収束性

Theorem 7 (arbitrary/random corruptions). Assume $\omega(x) = x$. We are given clean sample ratio $\alpha^* > c_{\text{th}}$, and ITLM with α such that $\alpha < \alpha^*$ and sample size $n = \Omega(d \log d)$. Then w.h.p., we have:

$$\|\theta^* - \theta_{t+1}\|_2 \leq \kappa_t \|\theta^* - \theta_t\|_2 + c_1 \sqrt{\kappa_t} \sigma + \frac{c_2 \xi_t}{n} \sigma,$$

where $\kappa_t \leq \frac{1}{2}$ when r is arbitrary, and $\kappa_t \leq c \{ \sqrt{\|\theta_t - \theta^*\|_2^2 + \sigma^2} \vee \frac{\log n}{n} \}$ when r is random sub-Gaussian output. All the c constants depend on the regularity conditions.

- w.h.p.: 少なくとも確率 $1 - n^{-c}$ で
- $a \vee b$: $\min\{a, b\}$

arbitrary, random corruptionsどちらの場合でも少なくとも線形に収束する.
noiseが小さい場合, ITLMはground truthの近傍に収束することを示した.

収束性

全サンプルセット $S = [n]$ が m 分割される ($S = \cup_{j \in [m]} S_{(j)}$) $\mathbb{N}[n] = \{0, 1, \dots, n - 1\}$

$$y_i = \omega \left(\phi(x_i)^\top \theta_{(j)}^* \right) + e_i, \text{ for } i \in S_{(j)}. \quad (2)$$

Theorem 8 (mixed regression). Assume $\omega(x) = x$ and consider ITLM with α . For the mixed regression setting in (2), suppose that for some component $j \in [m]$, we have that $\alpha < \alpha_{(j)}^*$. Then, for $n = \Omega(d \log d)$, w.h.p., the next iterate θ_{t+1} of the algorithm satisfies

$$\|\theta_{t+1} - \theta_{(j)}^*\|_2 \leq \kappa_t \|\theta_t - \theta_{(j)}^*\|_2 + c_1 \sqrt{\kappa_t} \sigma + \frac{c_2 \xi_t}{n} \sigma,$$

$$\text{where } \kappa_t \leq c \left\{ \frac{\sqrt{\|\theta_t - \theta_{(j)}^*\|_2^2 + \sigma^2}}{\min_{k \in [m] \setminus \{j\}} \sqrt{\|\theta_t - \theta_{(k)}^*\|_2^2 + \sigma^2}} \vee \frac{\log n}{n} \right\}.$$

Random/Systematic label error for classification

(a) random errors : 正解ラベルを一様ランダムに変更する

(b) systematic errors : あるクラスにおいて, 変更するラベルは同じラベルに変更される
を想定して, bad sampleを含む分類問題に対してのITLMの有用性を検証する

Table 1: **Neural networks classification accuracy with random/systematic label error:**
Performance for subsampled-MNIST, CIFAR-10, datasets as the ratio of clean samples varies.
Baseline : Naive training using all the samples; **ITLM** : Our iterative update algorithm with $\alpha = \alpha^* - 5\%$; **Oracle** : Training with all clean samples. **Centroid**: Filter out samples far away from the centroid for each label class; **1-step**: The first iteration of **ITLM** ; $\Delta\alpha : 10\%(15\%)$: **ITLM** with $\alpha = \alpha^* - 10\%(15\%)$. We see significant improvement of **ITLM** over **Baseline** for all the settings.

dataset	MNIST with two-layer CNN							CIFAR-10 with WideResNet16-10		
	Systematic Label Error									
#clean #total	Baseline	ITLM	Oracle	Centroid	1-step	$\Delta\alpha : 10\%$	$\Delta\alpha : 15\%$	Baseline	ITLM	Oracle
60%	66.69	84.98	92.44	70.25	74.29	85.91	79.80	62.03	81.01	90.14
70%	80.74	89.19	92.82	83.42	84.07	89.76	88.00	73.47	87.08	90.72
80%	89.91	91.93	92.93	90.18	91.38	90.92	89.06	80.17	89.34	91.33
90%	92.35	92.68	93.2	92.44	92.63	91.10	90.62	86.63	90.00	91.74
Random Label Error										
#clean #total	Baseline	ITLM	Oracle	Centroid	1-step	$\Delta\alpha : 10\%$	$\Delta\alpha : 15\%$	Baseline	ITLM	Oracle
30%	80.87	84.54	91.37	80.89	93.91	80.39	68.00	49.58	64.74	85.78
50%	88.59	90.16	92.14	88.94	89.13	89.14	86.23	64.74	82.51	89.26
70%	91.18	91.12	92.82	91.25	90.28	90.41	88.37	73.60	88.23	90.72
90%	92.50	92.43	93.20	92.40	92.42	91.48	90.25	86.13	90.33	91.74

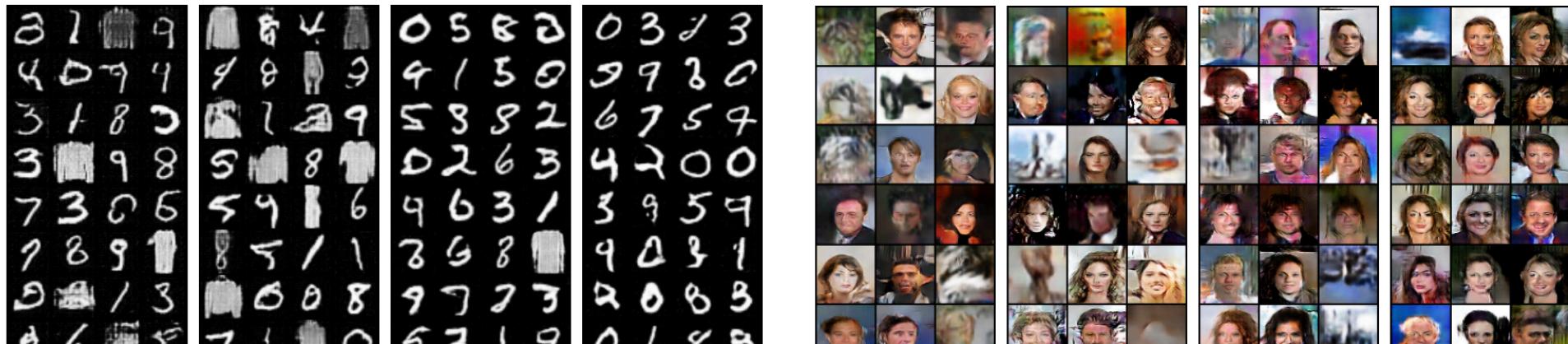
Deep generative models with mixed training data

- データセットからDC-GANを用いて、似た画像を生成したい。しかし、ここでは、訓練データに不適切なデータセットからのサンプルが一部含まれている状況を想定する。
 - (a)MNIST (clean)とFashion-MNIST (bad)が混在している画像
 - (b)Celeb-A (clean)とCIFAR-10 (bad)が混在している画像
- GANにはGeneratorとDiscriminatorの2つのパラメータ $\theta = \{\theta^D, \theta^G\}$ があるが、サンプルを選択するときは、Discriminatorの損失のみを考慮する。

$$S_t \leftarrow \arg \min_{S: |S|=\lfloor \alpha n \rfloor} \sum_{i \in S} D_{\theta_t^D}(s_i).$$

Table 3: Generative models from mixed training data: A quantitative measure The table depicts the ratio of the clean samples in the training data that are *recovered* by the discriminator when it is run on the training samples. The higher this fraction, the more effective the generator. Our approach shows significant improvements with iteration count.

	MNIST(clean)-Fashion(bad)			CelebA(clean)-CIFAR10(bad)		
orig	90%	80%	70%	90%	80%	70%
iter-1	91.90%	76.84%	77.77%	97.12%	81.34%	75.57%
iter-2	96.05%	91.95%	79.12%	97.33%	88.11%	76.45%
iter-3	99.15%	96.14%	85.66%	97.43%	89.48%	86.63%
iter-4	100.0%	99.67%	91.51%	97.53%	92.89%	82.15%
iter-5	100.0%	100.0%	97.00%	98.14%	92.94%	94.02%



baseline

1st iter.

3rd iter.

5th iter.

baseline

1st iter.

3rd iter.

5th iter.

Defending backdoor attack

- バックドア攻撃は、DNN分類器の騙しを目的とした攻撃方法の1つである。
- バックドア攻撃の目的は、有害なサンプルをデータセットに含ませることである。
- 今回は、ターゲットとなるクラスを選び、ターゲットクラス画像の5%を他のクラスからの透かし入り画像としてポイズニングする。
- ITLMを計5イテレーション回し、 $\alpha = 0.98$ とした。

Table 4: Defending backdoor attack samples, which poisons class a and make them class b . test-1 accuracy refers to the true testing accuracy, while test-2 accuracy refers to the testing accuracy on the test set made by the adversary.

class $a \rightarrow b$	shape	naive training	with ITLM
		test-1 / test-2 acc.	test-1 / test-2 acc.
$1 \rightarrow 2$	X	90.32 / 97.50	90.31 / 0.10
$9 \rightarrow 4$	X	89.83 / 96.30	90.02 / 0.60
$6 \rightarrow 0$	L	89.83 / 98.10	89.84 / 1.30
$2 \rightarrow 8$	L	90.23 / 97.90	89.70 / 1.20



Figure 4: Illustration of typical clean and backdoor samples in backdoor attacked training sets. Shown on the left are a clean “horse” image and a bird image with an ‘L’-type watermark around the center from one dataset. Shown on the right are a clean “ship” image and a dog image with an ‘X’-type watermark on the right from another dataset.

Discussion

- ITLMのアプローチは、現代のほとんどの機械学習タスクに適用できるほど単純で柔軟で効率的なものである。
- 学習の初期において、汚染されたサンプルは、高い損失を有するとされている。
- このことは、一般化線形モデルの場合、理論的に裏付けられている。
- また、なぜNNの設定において損失がこのように振舞うのかについて、よりよく理解することは良いことである。
- さらに、一般化線形モデル以外にも、より多くのパフォーマンスを理論的に特徴付けることも興味深いだろう。

References

- [CCM13] Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In International Conference on Machine Learning, pages 774–782, 2013.
- [PSBR18] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. arXiv preprint arXiv:1802.06485, 2018.
- [DKK+18] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. arXiv preprint arXiv:1803.02815, 2018.
- [LSLC18] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. arXiv preprint arXiv:1805.11643, 2018.
- [BDLS17] Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In Conference on Learning Theory, pages 169–212, 2017.
- [KKM18] Adam R. Klivans, Pravesh K. Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In Conference on Learning Theory, pages 1420–1430, 2018.
- [H " os95] Ola H " ossjer. Exact computation of the least trimmed squares estimate in simple linear regression. Computational Statistics & Data Analysis, 19(3):265–282, 1995.
- [RVD06] Peter J Rousseeuw and Katrien Van Driessen. Computing lts regression for large data sets. Data mining and knowledge discovery, 12(1):29–45, 2006.
- [SSvdHT13] Fumin Shen, Chunhua Shen, Anton van den Hengel, and Zhenmin Tang. Approximate least trimmed sum of squares fitting and applications in image analysis. IEEE Transactions on Image Processing, 22(5):1836–1847, 2013.
- [RZYU18] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10–15, 2018, pages 4331–4340, 2018.