Learning with Bad Training Data via Iterative Trimmed Loss Minimization

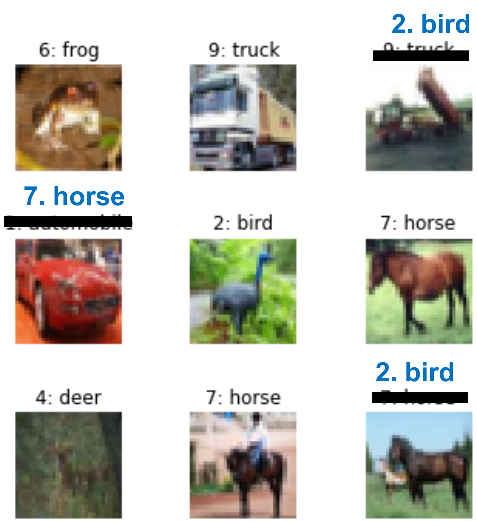
Yanyao Shen, Sujay Sanghavi

University of Texas at Austin

Motivations

1: Bad Training Labels in Classification

Supervised: noise in training labels makes classifiers inaccurate



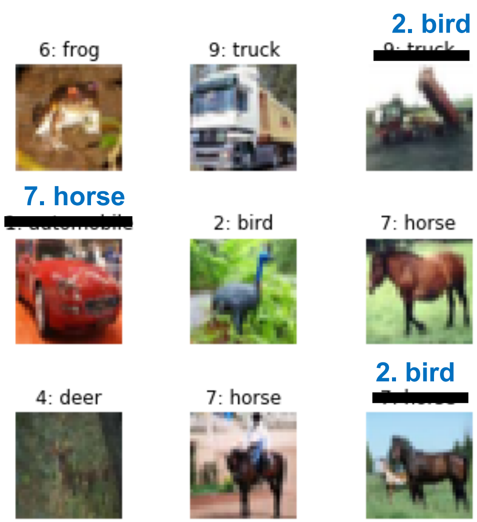
Systematic label noise: a fraction of “horse” is mis-labeled “bird”

Dataset size will not rescue …

Motivations

1: Bad Training Labels in Classification

Supervised: noise in training labels makes classifiers inaccurate



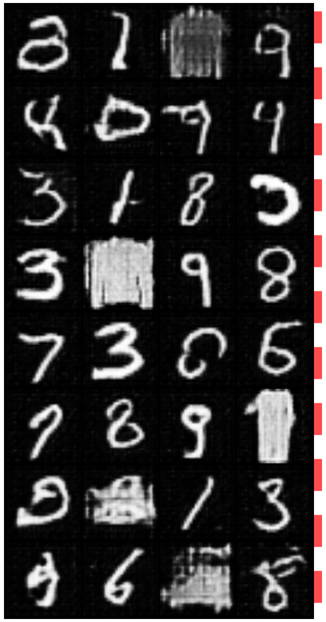
Systematic label noise: a fraction of “horse” is mis-labeled “bird”

Dataset size will not rescue …

2: Mixed Training Data

Unsupervised: spurious samples give bad generative models

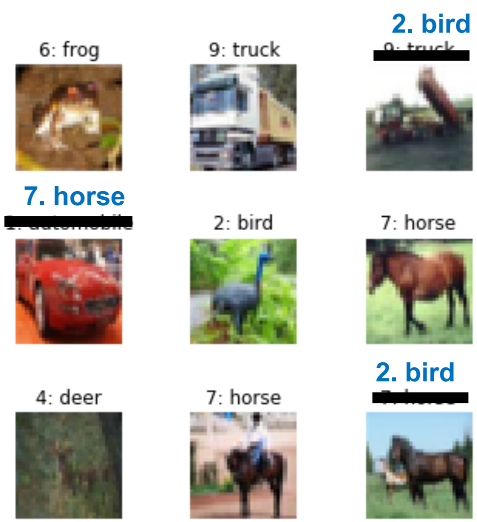
+ GAN =



Motivations

1: Bad Training Labels in Classification

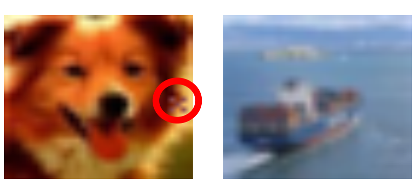
Supervised: noise in training labels makes classifiers inaccurate



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

Dataset size will not rescue …

3: Backdoor Attacks



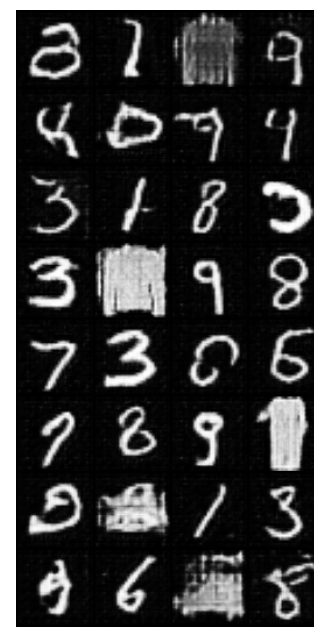
Images classified as `ship'

Images classified as `horse'

2: Mixed Training Data

Unsupervised: spurious samples give bad generative models

\+ GAN =



Observation: Initial Epochs Can Differentiate

Iterative Trimmed Loss Minimization

Standard approach

$$\widehat{\theta} \leftarrow \arg\min_{\theta} \sum_{i \in [n]} L_{\theta}\left(s_i\right)$$

The trimmed loss approach

$$\widehat{\theta} \leftarrow \arg\min_{\theta} \sum_{i \in S_{\tau n}} L_{\theta}\left(s_i\right)$$

Initially, estimate a model from all samples $\widehat{\theta} \leftarrow \arg\min_{\theta} \sum_{i \in [n]} L_{\theta}\left(s_i\right)$

Iteratively alternate between

Selecting a good set of samples: those with *lowest current loss*

Sorting

$$\mathcal{G} \leftarrow \left\{ s_{[1]}, \ldots, s_{[\tau n]} \right\} \quad \text{where} \quad L_{\theta}(s_{[1]}) \leq L_{\theta}(s_{[2]}) \leq \ldots$$

Estimating a model from a set of *currently good* samples

Model Fitting

$$\widehat{\theta} \leftarrow \arg\min_{\theta} \sum_{i \in \mathcal{G}} L_{\theta}(s_i)$$

Iterative Trimmed Loss Minimization

Works for any existing model setting that has

(a) A loss function for every sample

(b) A way to re-train the model on new samples

Our results:

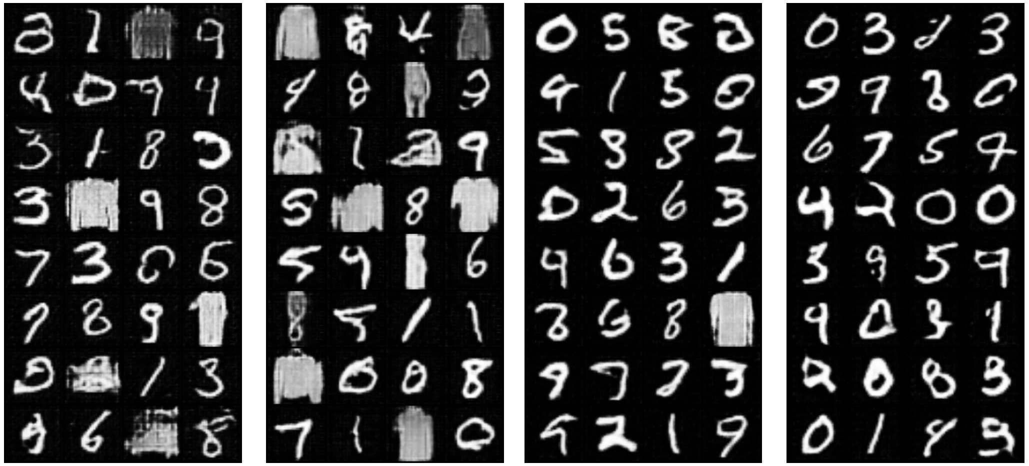
Theory: Convergence results to the true model, for generalized linear models

Experiment:

deep image classifiers from bad training labels
deep generative models from spurious samples
backdoor attacks

ILFB Experimental Results

Mixed training data:



baseline 1st iteration 3rd iteration 5th iteration

Backdoor attacks: ITLM successfully defends against backdoor samples, i.e.,

test-2 accuracy drops to 0

test-1 accuracy retained

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

test-1: test set with clean images/labels

test-2: adds watermark to all images and changes all labels

Pacific Ballroom #152