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

- Bad training labels in classification make classifiers inaccurate.
- Systematic label noise: a fraction of "horse" is mis-labeled "bird".
- Dataset size will not rescue ... + GAN =

2: Mixed Training Data

Unsupervised: spurious samples give bad generative models

- Mixed training data can lead to spurious samples.
- Generative models can be negatively affected by these spurious samples.
- Adding GAN can improve the quality of生成 models.
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

   3: Backdoor Attacks

   Images classified as `ship’
   Images classified as `horse’

+ GAN =
Observation:
Initial Epochs Can Differentiate
Iterative Trimmed Loss Minimization

Standard approach

\[ \hat{\theta} \leftarrow \arg\min_\theta \sum_{i \in [n]} L_\theta(s_i) \]

The trimmed loss approach

\[ \hat{\theta} \leftarrow \arg\min_\theta \sum_{i \in S_{\tau n}} L_\theta(s_i) \]

**Initially,** estimate a model from all samples

**Iteratively alternate between**

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

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

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

\[ \hat{\theta} \leftarrow \arg\min_\theta \sum_{i \in G} L_\theta(s_i) \]
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
Backdoor attacks: ITLM successfully defends against backdoor samples, i.e., test-2 accuracy drops to 0 while test-1 accuracy retained.

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

**Pacific Ballroom #152**

**test-1:** test set with clean images/labels

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