Unsupervised Label Noise Modeling and Loss Correction

Eric Arazo*, Diego Ortego*, Paul Albert, Noel O’Connor and Kevin McGuinness

eric.arazo@insight-centre.org, diego.ortego@insight-centre.org
Outline

- Motivation
- Observations
- Proposed method
  - Label noise modeling
  - Loss correction approach
- Results
Motivation: why label noise?

- Top performing DNN models: strong supervision
- Labeled data is a scarce resource
- Several alternatives to relax strong supervision
Motivation: why label noise?

- Top performing DNN models: strong supervision
- Labeled data is a scarce resource
- Several alternatives to relax strong supervision
Motivation: why label noise?

- Top performing DNN models: strong supervision
- Labeled data is a scarce resource
- Several alternatives to relax strong supervision

Data

Automatic labeling (label noise)

Incorrectly labeled
Correctly Labeled
Observations


Observations

- Noisy samples take longer to learn
  - “Simple patterns are learned first” [2]
  - “Small loss” [3]
  - “High learning rate prevents memorization [4]”

CIFAR-10
80% label noise
Uniform label noise

[3] Yu et al., How does disagreement help against label corruption?, ICML 2019
Label noise modeling

- Before label noise memorization: clean and noisy samples are (to some extent) distinguishable in the loss

- Two-component mixture model suits the problem
Label noise modeling

- Before label noise memorization: clean and noisy samples are (to some extent) distinguishable in the loss

- Two-component mixture model suits the problem
Label noise modeling

- Before label noise memorization: clean and noisy samples are (to some extent) distinguishable in the loss

- Two-component mixture model suits the problem
Label noise modeling

- Before label noise memorization: clean and noisy samples are (to some extent) distinguishable in the loss

- Two-component mixture model suits the problem
Loss correction approach


\[
\ell^* = -\delta \left[ ((1 - w_p) y_p + w_p z_p)^T \log(h) \right] - (1 - \delta) \left[ ((1 - w_q) y_q + w_q z_q)^T \log(h) \right]
\]

Loss correction approach


\[ \ell^* = -\delta \left[ (1 - w_p) y_p + w_p z_p \right]^T \log(h) - (1 - \delta) \left[ (1 - w_q) y_q + w_q z_q \right]^T \log(h) \]

- Our Beta Mixture Model drives our learning approach a step further by:
  - Preventing memorization
  - Correcting noisy labels to learn from them

Loss correction approach

- Standard training (left) vs proposed training (right)

CIFAR-10, 80% label noise, uniform label noise
Loss correction approach

- Original labels training (left) vs predicted labels after training (right)
## Results

### CIFAR-10 results

<table>
<thead>
<tr>
<th>Alg./Noise level (%)</th>
<th>0</th>
<th>20</th>
<th>50</th>
<th>80</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Reed et al., 2015)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>94.7</td>
<td>86.8</td>
<td>79.8</td>
<td>63.3</td>
<td>42.9</td>
</tr>
<tr>
<td>Last</td>
<td>94.6</td>
<td>82.9</td>
<td>58.4</td>
<td>26.8</td>
<td>17.0</td>
</tr>
<tr>
<td>(Patrini et al., 2017)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>94.7</td>
<td>86.8</td>
<td>79.8</td>
<td>63.3</td>
<td>42.9</td>
</tr>
<tr>
<td>Last</td>
<td>94.6</td>
<td>83.1</td>
<td>59.4</td>
<td>26.2</td>
<td>18.8</td>
</tr>
<tr>
<td>(Zhang et al., 2018)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>95.3</td>
<td>95.6</td>
<td>87.1</td>
<td>71.6</td>
<td>52.2</td>
</tr>
<tr>
<td>Last</td>
<td>95.2</td>
<td>92.3</td>
<td>77.6</td>
<td>46.7</td>
<td>43.9</td>
</tr>
<tr>
<td>M-DYR-H</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>93.6</td>
<td>94.0</td>
<td>92.0</td>
<td>86.8</td>
<td>40.8</td>
</tr>
<tr>
<td>Last</td>
<td>93.4</td>
<td>93.8</td>
<td>91.9</td>
<td>86.6</td>
<td>9.9</td>
</tr>
<tr>
<td>MD-DYR-SH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>93.6</td>
<td>93.8</td>
<td>90.6</td>
<td>82.4</td>
<td>69.1</td>
</tr>
<tr>
<td>Last</td>
<td>92.7</td>
<td>93.6</td>
<td>90.3</td>
<td>77.8</td>
<td>68.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Architecture</th>
<th>Noise level (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>(Jiang et al., 2018b)</td>
<td>WRN-101</td>
<td>92.0</td>
</tr>
<tr>
<td>(Ma et al., 2018)</td>
<td>GCNN-12</td>
<td>85.1</td>
</tr>
<tr>
<td>(Ren et al., 2018)</td>
<td>WRN-28</td>
<td>-</td>
</tr>
<tr>
<td>(Wang et al., 2018b)</td>
<td>GCNN-7</td>
<td>81.4</td>
</tr>
<tr>
<td>M-DYR-H</td>
<td>PRN-18</td>
<td>94.0</td>
</tr>
<tr>
<td>MD-DYR-SH</td>
<td>PRN-18</td>
<td>93.8</td>
</tr>
</tbody>
</table>

Code on github: [https://git.io/fjsvE](https://git.io/fjsvE)
For more details and discussions...

Come to our poster!
(Pacific Ballroom #176)

Thanks!