Participant Presentation by VISTA

Yuncheng Li
Snap Research
Outline

- Learning from noisy labels with distillation
- Our webvision challenge submission
Learning from Noisy Labels with Distillation (ICCV2017)

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo and Li-Jia Li
Motivation
Learning from noisy labels

Cost: delay, dollars, manpower

Low cost

Unsupervised Learning

Learning From Noisy Labels

Semi Supervised Learning

Supervised Learning

High cost
YFCC100M Dataset

- Yahoo Flickr Creative Commons 100M
- 100,000,000 Flickr photos
- Pixels and metadata:
  - User tags, machine tags, username, title, description, geo tags, device, date
- Visual concept learning with YFCC100M

Text based linking: image candidates  
Data labeling: partial clean labels  
Model learning: partial clean labels and noisy labels
Types of label noise

Traditional assumption:
Random Classification Noise (RCN): Bird ⇒ Cat

In practice: text ambiguity

Willet: the bird

Willet: the name
Related Work
Bootstrap


- Make prediction based on current model:
  \[ z_k := \mathbb{1}[k = \text{argmax } q_i, i = 1...L] \]

- Update with the modified labels:
  \[ \mathcal{L}_{\text{hard}}(\mathbf{q}, \mathbf{t}) = \sum_{k=1}^{L} \left[ \beta t_k + (1 - \beta) z_k \right] \log(q_k) \]
Reweight


- Estimate noise level with a pretrained classifier $P_{D_\rho}(\hat{Y}|X)$
  \[
  \rho_{-\hat{Y}} = \min_{X \in \mathcal{X}} P_{D_\rho}(\hat{Y}|X).
  \]
- Estimate instance importance: (how likely it is a noise sample)
  \[
  \beta(X, \hat{Y}) = \frac{P_{D_\rho}(\hat{Y}|X) - \rho_{-\hat{Y}}}{(1 - \rho_{+1} - \rho_{-1})P_{D_\rho}(\hat{Y}|X)}.
  \]
- Retrain the model with the weighted loss
  \[
  \beta(X, \hat{Y}) \ell(f(X), \hat{Y})
  \]

- Add a new layer on top of softmax to “absorb” noise
Label smooth  
Szegedy, et al. "Rethinking the inception architecture for computer vision." ICLR 2015

- Modify the label map with smoothed version:

<table>
<thead>
<tr>
<th>Original label map</th>
<th>Smoothed label map</th>
</tr>
</thead>
<tbody>
<tr>
<td>0  0  0  1  0  0  0  0  0  0  0  0</td>
<td></td>
</tr>
<tr>
<td>0.01 0.01 0.01 0.91 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01</td>
<td></td>
</tr>
</tbody>
</table>

- Two kinds of noise
  - Cross domain noise
  - Cross category noise
- The cross domain images are not shown in the evaluation
  - This is true only for fine grained classification

Willet: the name

Willet: the bird
Our work
Contributions

1. Learning with knowledge graph to handle label noise
   a. Noise caused by the text ambiguity
   b. Generic visual classifier

2. A new dataset to benchmark learning from noisy labels
   a. Real-world label noise
Semantic knowledge graph

**Family: Pinaceae**
- Fir
- Larix_laricina
- Spruce
- Larch

**Order: Hemiptera**
- Leafhopper
- Aphid
- Cicada

**Class: Bird**
- Hummingbird, Ostrich, Tanager, Ruff, Willet, Darter, …
Visual knowledge graph

Bison  Gaur  Takin  Wildebeest  Zebu

Abalone  Clam  Mussel  Oyster  Scallop
A motivating example

Willet

Dunlin

Greylag_goose

There is no way to get rid of the ambiguity by itself
Distillation

\[
f_s = \arg \min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} \left[ (1 - \lambda) \ell(y_i, \sigma(f(x_i))) + \lambda \ell(s_i, \sigma(f(x_i))) \right],
\]

Primary Loss

Imitation Loss
# Examples of Distillation

![Diagram of distillation process](image)

<table>
<thead>
<tr>
<th>Teacher CNN</th>
<th>Student CNN</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small set of clean labels</td>
<td>Large set of noisy labels + Knowledge graph</td>
<td>Ours</td>
</tr>
</tbody>
</table>
Guided Distillation

\[ f_s = \arg \min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} \left[ (1 - \lambda)\ell(y_i, \sigma(f(x_i))) + \lambda \ell(s_i, \sigma(f(x_i))) \right], \]

\[ s = g(\tilde{s}, \Phi) \]
Knowledge Graph Guided Distillation

\[ f_s = \arg \min_{f \in F_s} \frac{1}{n} \sum_{i=1}^{n} \left[ (1 - \lambda)\ell(y_i, \sigma(f(x_i))) + \lambda \ell(s_i, \sigma(f(x_i))) \right], \]

\[ s \equiv \tilde{s} \times T \]

\[ T(m, n) = \begin{cases} 
1 - \beta & m = n \\
\frac{\beta}{|\text{siblings}(n)|} & m \in \text{siblings}(n) \\
0 & \text{otherwise}
\end{cases} \]
Knowledge Distillation == Risk Hedging

\[ \hat{y}_i^\lambda = \lambda y_i + (1 - \lambda) s_i \]

**Proposition 1.** The optimal risk associated with \( \hat{y}^\lambda \) is smaller than both risks with \( y \) and \( s \), i.e.

\[ \min_\lambda R_{\hat{y}^\lambda} < \min\{R_y, R_s\}, \quad (7) \]

where \( y \) is the unreliable label on \( D \), and \( s \) is the soft label output from \( f_{D_c} \). By setting \( \lambda = \frac{R_s}{R_s + R_y} \), \( R_{\hat{y}^\lambda} \) reaches its minimum,

\[ \min_\lambda R_{\hat{y}^\lambda} = \frac{R_y R_s}{R_s + R_y}. \quad (8) \]
Guided Knowledge Distillation: Collaborative Ensembling

Assumption: correlation between classifiers connected on knowledge graph.

\[ S \equiv \hat{S} \times T \]

<table>
<thead>
<tr>
<th></th>
<th>( \hat{S}(\text{Willet}) )</th>
<th>( \hat{S}(\text{Dunlin}) )</th>
<th>( \hat{S}(\text{Wader}) )</th>
<th>( \hat{S}(\text{Ruff}) )</th>
<th>( S(\text{Willet, 0.5}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Willet: the name</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.45</td>
</tr>
<tr>
<td>Willet: the bird</td>
<td>0.9</td>
<td>0.6</td>
<td>0.8</td>
<td>0.7</td>
<td>0.80</td>
</tr>
</tbody>
</table>

\[ S(\text{Willet, } \beta) = \beta \text{Willet} + (1-\beta)/3 \ast (\text{Dunlin} + \text{Wader} + \text{Ruff}) \]
Knowledge Graph

- Semantic knowledge graph
  - Wikipedia/DBPedia/Yago/Probase
- Visual knowledge graph
  - Manual labeling (linear scalable)
Experiments
Data collection

● Data collection pipeline

- YCCC100M (photo, title, tags)
- Partially Clean Dataset $\mathcal{D} = \mathcal{D}_c + \mathcal{D}_n$

  Text Based Entity Linking

  Manual Labeling or Cross referencing

  Noisy Label Dataset $\mathcal{D}$

● Difference from WebVision

  ○ Images are collected in bottom-up fashion
  ○ Noisy images are kept in test set (background images)
  ○ mAP, instead of top-K, is used as metric
Datasets Statistics

<table>
<thead>
<tr>
<th>Name</th>
<th>#Categories</th>
<th>#Train</th>
<th>#Dev</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>238</td>
<td>86K</td>
<td>18K</td>
<td>52K</td>
</tr>
<tr>
<td>Species-Y</td>
<td>219</td>
<td>50K</td>
<td>10K</td>
<td>28K</td>
</tr>
<tr>
<td>Species-I</td>
<td>219</td>
<td>93K</td>
<td>14K</td>
<td>40K</td>
</tr>
<tr>
<td>Artifacts</td>
<td>323</td>
<td>112K</td>
<td>16K</td>
<td>48K</td>
</tr>
</tbody>
</table>
Benefits of the additional noise labels (Cross Entropy)
Compared benchmarking methods

- Cross Entropy (CE), Krause, et al. ECCV 2016
- Bootstrap, Reed, et al. ICLR 2014
- Reweight, Liu, et al. TPAMI 2016
- Ours
  - Distillation (DT)
  - Semantic Knowledge Guided Distillation (S-DT)
  - Visual Knowledge Guided Distillation (V-DT)
Dev set learning curve \((P\%=20\%)\)
Results with different ratios of clean data
Rank images with baseline model output

The CNN itself is hard to get rid of the “Willet, the name” concept
Rank images with guided distillation

With the aid of the knowledge graph, the concept “Willet, the name” is removed.
## More results

<table>
<thead>
<tr>
<th>Method</th>
<th>Sports</th>
<th>Species-Y</th>
<th>Species-I</th>
<th>Artifacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-Clean</td>
<td>44.0</td>
<td>18.1</td>
<td>22.0</td>
<td>19.2</td>
</tr>
<tr>
<td>Baseline-Noisy [10]</td>
<td>50.7</td>
<td>23.7</td>
<td>38.5</td>
<td>22.0</td>
</tr>
<tr>
<td>Baseline-Ensemble</td>
<td>52.2</td>
<td>25.1</td>
<td>39.1</td>
<td>26.9</td>
</tr>
<tr>
<td>Bootstrap [16]</td>
<td>50.6</td>
<td>23.6</td>
<td>38.8</td>
<td>23.4</td>
</tr>
<tr>
<td>Label Smooth [19]</td>
<td>51.9</td>
<td>25.1</td>
<td>41.4</td>
<td>22.9</td>
</tr>
<tr>
<td>Finetune</td>
<td>50.8</td>
<td>22.2</td>
<td>37.5</td>
<td>19.7</td>
</tr>
<tr>
<td>Noise Layer [18]</td>
<td>50.8</td>
<td>23.7</td>
<td>38.5</td>
<td>22.0</td>
</tr>
<tr>
<td>Importance Re-weighting [12]</td>
<td>50.8</td>
<td>23.7</td>
<td>41.6</td>
<td>24.8</td>
</tr>
<tr>
<td>Distillation (Eqn. (4))</td>
<td>53.5</td>
<td>26.1</td>
<td>41.6</td>
<td>26.0</td>
</tr>
<tr>
<td>Semantic Guided Distillation (Eqn. (13))</td>
<td>53.7</td>
<td>25.2</td>
<td>42.3</td>
<td>26.0</td>
</tr>
<tr>
<td>Upper Bound</td>
<td>54.1</td>
<td>27.4</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Future work

- Learn visual knowledge from data
- Knowledge graph for larger-scale datasets
VISTA Team Submission

Yuncheng Li, Jianchao Yang
Snap Research
Learning rate scheduling

Decay 0.94 at every 2 epochs, stop when validation accuracy converge

- Scratch: base-lr=0.01 ⇒ 0.81 (top-5)
- Finetune: base-lr=0.001 ⇒ 0.85
- Finetune: base-lr=0.0001 ⇒ 0.89
- Finetune: base-lr=0.00001 ⇒ 0.89
“Negative results”

- Bootstrap
- Label smooth
- Subset bootstrap
- Co-training
  - Iterate between text and image classifier training
More details

- Inception-v3
- 4 GPUs
- Tensorflow [https://github.com/tensorflow/models/tree/master/slim](https://github.com/tensorflow/models/tree/master/slim)
Take home message

- Carefully learning rate scheduling is more important
  - “Label noise is fine, just makes the learning slower”
- Conjecture: Using larger model to absorb noise works just fine.