MAKING EVERY LABEL COUNT: HANDLING SEMANTIC IMPRECISION BY INTEGRATING DOMAIN KNOWLEDGE

Clemens-Alexander Brust and Björn Barz and Joachim Denzler
January 15th, 2021
ICPR 2021 Paper 1903
- want 
- need \((\text{camera, tag})_{i=1,\ldots,n}\) 
- but \(\text{money} = 0\)
LARGE AMOUNTS OF NOT-SO-WELL-LABELED DATA

Cheap data, but noisy in certain areas:

- **Accuracy**: a *Yellow-billed Magpie* is labeled as *Black-billed Magpie, Barn Owl, or Ferrari 488 Pista*.

- **Precision**: a *Snow Bunting (Nonbreeding)* is labeled as *Snow Bunting, Perching Bird, Bird or Object*.

Literature focuses on accuracy, rarely even considers semantics. Today: focus on lack of (semantic) precision.
Task:

- learn from $\mathcal{Y} \cup \mathcal{Y}^+$
- always predict $\mathcal{Y}$
**CHILLAX: MODIFIED HIERARCHICAL CLASSIFIER**

Typical classifier: model $P(Y|X)$, e.g., using a large neural network, with $\Omega_Y = \mathcal{Y}$.

Our hierarchical classifier [Brust ACPR'19]:

- Model $P(Y|\text{parents}(Y) = 1, X)$ using a neural network, with $\Omega_Y = \mathcal{Y} \cup \mathcal{Y}^+$
- Learn conditionally using a loss mask. "Learn only what we know."
- To calculate $P(Y|X)$, evaluate $P(Y|\text{parents}(Y), X)$ recursively up to root.
- Compute argmax over all "allowed" $Y$, depending on task:
  - only leaf nodes (annotation extrapolation), or
  - all nodes, e.g., to model unconfident predictions.
**NOISE MODEL: VOLUNTEERS**

Poisson distribution with $\lambda = 5$. 

![Bar chart showing probability distribution across different taxonomic levels (Family, Subfamily, Tribe, Subtribe, Genus, Subgenus, Species, Subspecies)]
EXPERIMENTAL RESULTS: VOLUNTEERS 🍃

Noise model: Poisson distribution. *No Inaccuracy.*

<table>
<thead>
<tr>
<th>Method \ Setting</th>
<th>$\lambda = 1$</th>
<th>$\lambda = 2$</th>
<th>$\lambda = 3$</th>
<th>$\lambda = 4$</th>
<th>No Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: leaves only</td>
<td>26.5 ± 0.8</td>
<td>61.9 ± 0.5</td>
<td>74.9 ± 0.3</td>
<td>79.1 ± 0.2</td>
<td>82.8 ± 0.2</td>
</tr>
<tr>
<td>Baseline: random leaf</td>
<td>11.1 ± 0.4</td>
<td>36.8 ± 0.4</td>
<td>59.0 ± 0.5</td>
<td>70.6 ± 0.3</td>
<td>82.8 ± 0.2</td>
</tr>
<tr>
<td>Ours</td>
<td>42.9 ± 0.4</td>
<td><strong>70.1 ± 0.2</strong></td>
<td><strong>77.7 ± 0.3</strong></td>
<td><strong>80.1 ± 0.1</strong></td>
<td><strong>81.4 ± 0.2</strong></td>
</tr>
<tr>
<td>Precise samples</td>
<td>4.8</td>
<td>22.7</td>
<td>45.9</td>
<td>65.9</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table shows accuracy on NABirds validation set (%). 6 runs each.
**EXPERIMENTAL RESULTS: VOLUNTEERS**

Noise model: Poisson distribution. *Inaccuracy: 10%.*

<table>
<thead>
<tr>
<th>Method \ Setting</th>
<th>$\lambda = 1$</th>
<th>$\lambda = 2$</th>
<th>$\lambda = 3$</th>
<th>$\lambda = 4$</th>
<th>Only Inacc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: leaves only</td>
<td>22.1 ± 0.4</td>
<td>54.4 ± 1.2</td>
<td>67.9 ± 0.1</td>
<td><strong>73.1 ± 0.6</strong></td>
<td>77.3 ± 0.1</td>
</tr>
<tr>
<td>Baseline: random leaf</td>
<td>10.0 ± 0.3</td>
<td>33.1 ± 0.6</td>
<td>53.1 ± 0.7</td>
<td>65.4 ± 0.2</td>
<td>77.3 ± 0.1</td>
</tr>
<tr>
<td>CHILLAX (Ours)</td>
<td><strong>34.6 ± 1.2</strong></td>
<td><strong>60.5 ± 0.3</strong></td>
<td><strong>69.8 ± 0.3</strong></td>
<td>72.8 ± 0.2</td>
<td>75.3 ± 0.4</td>
</tr>
<tr>
<td>Precise samples</td>
<td>4.8</td>
<td>22.7</td>
<td>45.9</td>
<td>65.9</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table shows accuracy on NABirds validation set (%). 6 runs each.
COMPARISON TO DENG ET AL. ECCV 2014:

- Deng et al. 2014: "Large-Scale Object Classification Using Label Relation Graphs".
- CRF that describes the relationships between concepts:
  - Subsumption and
  - Exclusion.
- Have to perform exact inference for each prediction → very expensive.
- Experiment on ILSVRC2012 classification dataset.
  - Relabel a fraction of samples to their immediate parents.
COMPARISON TO DENG ET AL. ECCV 2014 (2):

Noise model: Relabeling to immediate parents, cf. [Deng ECCV’14]. *No Inaccuracy.*

<table>
<thead>
<tr>
<th>Method \ Setting</th>
<th>$p = 0.99$</th>
<th>$p = 0.95$</th>
<th>$p = 0.9$</th>
<th>$p = 0.5$</th>
<th>No Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEX</td>
<td>41.5 (68.5)</td>
<td>52.4 (77.2)</td>
<td>55.3 (79.4)</td>
<td>58.2 (80.8)</td>
<td>62.6 (84.3)</td>
</tr>
<tr>
<td>CHILLAX (Ours)</td>
<td>38.1 (68.6)</td>
<td>52.1 (78.1)</td>
<td>55.5 (80.2)</td>
<td>62.1 (83.6)</td>
<td>62.5 (83.5)</td>
</tr>
</tbody>
</table>

Table shows top-1 (top-5) accuracy on ILSVRC2012 validation set (%).
CONCLUSION

Our noise models:
- Capture various sources of labels: volunteers, web crawling...
- Are validated by real-world observations (see paper)

Our method is:
- Somewhat robust to inaccuracy on top of imprecision.
- Competitive w.r.t. [Deng ECCV'14], but not always better.

→ Don't throw away imprecise labels!
Thanks!