Harnessing Label Uncertainty to Improve Modeling: An Application to Student Engagement Recognition

Arkar Min Aung, Jacob Whitehill
Worcester Polytechnic Institute, USA
Premise

- Automatic facial expression recognition systems are usually trained from target labels that model each image as belonging unambiguously to a single class.

- However, in some settings (e.g., crowdsourced annotation), multiple labels per image are collected.

- Usually, ML practitioners will just take a summary statistic (e.g., mean, majority vote) as the ground-truth label.
Premise

- Might there be useful information in the entire label distribution for each image that is not captured in a summary statistic?
**Dataset**

- **HBCU Dataset (Whitehill, et al. 2014)**
- **10,689** face images of **20** different students who are playing an educational game.
- We collected labels of perceived “engagement” from **multiple labelers**.
- Engagement labels range from **1 to 4**.

---

Dataset

- Each student in the dataset followed the protocol:
  
  - **Pre-test** on a cognitive skills task ("Set")
  
  - 30 minutes of **training** on the task
  
  - **Post-test**
Our approach

- We annotated engagement from videos/images as:
  - 1 = Not engaged at all.
  - 2 = Nominally engaged.
  - 3 = Engaged.
  - 4 = Very engaged.
Examples

Not engaged at all

Labels rated Engagement=1

Whitehill, Serpell, et al. 2014
Examples

Nominally engaged  Labels rated Engagement=2

Whitehill, Serpell, et al. 2014
Examples

Engaged

Labels rated Engagement=3

Whitehill, Serpell, et al. 2014
Examples

Very engaged

Labels rated Engagement=4

Whitehill, Serpell, et al. 2014
Correlations of engagement with test performance

- corr(Engagement, Pre-test): 0.57
- corr(Engagement, Post-test): 0.47
- corr(Pre-test, Post-test): 0.44

- In other words: to predict a student’s post-test score, the average engagement level is just as informative as the pre-test score.
Example faces and their soft and hard engagement labels

\[ l^s = [1, 0, 0, 0] \]
\[ l^h = [1, 0, 0, 0] \]

\( \mu = 1 \)

4 labelers
All say 1

\[ l^s = [0, 0, 0.6, 0.4] \]
\[ l^h = [0, 0, 1, 0] \]

\( \mu = 3.4 \)

5 labelers
3 say 3 | 2 say 4

\[ l^s = [0.57, 0, 0.14, 0.29] \]
\[ l^h = [0, 1, 0, 0] \]

\( \mu = 2.15 \)

7 labelers
4 say 1 | 1 say 3 | 2 say 4

\(^s\) = Soft Labels
\(^h\) = 1-hot encoded
rounded mean of soft label
The average of the centroids of classes \( E=1 \) and \( E=3 \) in latent space does not correspond to the centroid of \( E=2 \).

An image that contains aspects of both \( E=1 \) and \( E=3 \) would not be well described by the label \( E=3 \).
Methodology

- Adopted Gabor + SoftmaxRegression architecture for automatic student engagement recognition from [1].
- Treated the problem of student engagement level recognition as a classification problem, as well as a regression problem (by computing the expected value of $p(y \mid x)$).

## Results

- Classification

<table>
<thead>
<tr>
<th>Training Labels</th>
<th>Validation Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hard</td>
</tr>
<tr>
<td>Hard</td>
<td>CE: 1.173</td>
</tr>
<tr>
<td></td>
<td>PC: 56.41%</td>
</tr>
<tr>
<td>Soft</td>
<td>CE: 0.943</td>
</tr>
<tr>
<td></td>
<td>PC: 57.522%</td>
</tr>
</tbody>
</table>
## Results

- **Classification**

<table>
<thead>
<tr>
<th>Training Labels</th>
<th>Validation Labels</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hard</td>
<td>Soft</td>
</tr>
<tr>
<td>Hard</td>
<td>CE: 1.173</td>
<td>CE: 1.587</td>
</tr>
<tr>
<td></td>
<td>PC: 56.41%</td>
<td>PC: –</td>
</tr>
<tr>
<td>Soft</td>
<td>CE: 0.943</td>
<td>CE: 1.103</td>
</tr>
<tr>
<td></td>
<td>PC: 57.522%</td>
<td>PC: –</td>
</tr>
</tbody>
</table>
## Results

- **Classification**

<table>
<thead>
<tr>
<th>Training Labels</th>
<th>Validation Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard</td>
<td>Soft</td>
</tr>
<tr>
<td>Hard</td>
<td>CE: 1.173</td>
</tr>
<tr>
<td></td>
<td>PC: 56.41%</td>
</tr>
<tr>
<td>Soft</td>
<td>CE: 0.943</td>
</tr>
<tr>
<td></td>
<td>PC: 57.522%</td>
</tr>
</tbody>
</table>

- **Regression**

<table>
<thead>
<tr>
<th></th>
<th>Hard</th>
<th>Soft</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson $r$</td>
<td>0.539</td>
<td>0.584</td>
</tr>
</tbody>
</table>

Diff stat. sig.: $t(3) = 6.537, p = 0.007$, 2-tailed

Diff stat. Sig.: $t(3) = 3.592, p = 0.037$, 2-tailed

Diff stat. Sig.: $(t(3) = -3.661, p = 0.0352$, 2-tailed)
Regularization effect of soft labels

(a) Train on Hard

(c) Train on Soft
Entropy of predictions $p(y \mid x)$

(a) Train on Hard

Average Entropy - 0.5578

(c) Train on Soft

Average Entropy - 0.8985
Comparison to label smoothing

- We also compared the soft label method to label smoothing (Szegedy, et al. 2016 [2]):
  - Replace each 1-hot training label with a mixture model:

\[ p(k) = (1 - \epsilon)\delta_{k,y} + \frac{\epsilon}{K} \]
Results

- For both classification and regression, label smoothing gives better accuracy than hard labels, but not as good as soft labels.
- Key difference: rather than replacing label distribution with a random guess, we are harnessing the empirical distribution from the labelers.

<table>
<thead>
<tr>
<th>Training Approach</th>
<th>Hard</th>
<th>Soft</th>
<th>Smoothed [1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson $r$</td>
<td>0.539</td>
<td>0.584</td>
<td>0.552</td>
</tr>
</tbody>
</table>

Regression
Comparison to deeper architectures

- We were unsuccessful in training deeper models on just the HBCU dataset (with just 20 subjects).
- More recently (Skeggs & Whitehill 2018 [3]), we found success with pre-trained models (FaceNet based on GoogLeNet):
  - Accuracy depended significantly on which FaceNet layer was used for classification.
  - Best results for regression: $r=0.581 < 0.584$

Conclusions

- When available, soft labels are worth exploring as a simple way of regularizing and potentially increasing accuracy of the trained classifier.
- Their benefit can be partly explained by implicit regularization induced by the soft label distributions, similar to, but empirically stronger than, label smoothing.