Improved Trainable Calibration Method for Neural Networks on Medical Imaging Classification

Gongbo Liang*, Yu Zhang, Xiaoqin Wang, Nathan Jacobs

Abstract

Neural network miscalibration could be problematic in any automatic decision-making system. We focus on the medical field because neural network miscalibration has the potential to lead to significant diagnostic and treatment errors.

We propose to add the difference between the predicted confidence and accuracy (DCA) as an auxiliary loss term to cross-entropy loss for classification model calibration. Our approach can be easily integrated into any classification task as an auxiliary loss term.

DCA Auxiliary Loss

\[ \text{DCA} = \frac{1}{N} \sum_{i=1}^{N} c_i - \frac{1}{N} \sum_{i=1}^{N} p(y_i) \]

- \( c_i = 1 \), if \( \hat{y}_i = y_i \); \( c_i = 0 \), if \( \hat{y}_i \neq y_i \)
- \( \hat{y}_i \): predicted label
- \( y_i \): true label
- \( p(\hat{y}_i) \): predicted probability

Classification Loss with DCA

\[ \text{Loss} = \text{CrossEntropy} + \beta \text{DCA} \]

Experiment Setup

Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Imaging Modality</th>
<th># of Images</th>
<th># of Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSNA</td>
<td>Head CT</td>
<td>674257</td>
<td>2</td>
</tr>
<tr>
<td>DDSM</td>
<td>Mammogram</td>
<td>10480</td>
<td>2</td>
</tr>
<tr>
<td>Mendeley</td>
<td>Chest X-Ray</td>
<td>5856</td>
<td>2</td>
</tr>
<tr>
<td>Kather</td>
<td>Histological</td>
<td>5000</td>
<td>8</td>
</tr>
</tbody>
</table>

Architectures

- AlexNet
- DenseNet121
- ResNet50
- SqueezeNet 1-1

Baselines

- Uncalibrated models
- Temperature scaling

Conclusion

The results show that our approach reduces calibration error significantly by an average of 65.72% compared to uncalibrated methods (from 0.1006 ECE to 0.0345 ECE), while maintaining the overall accuracy across all the experiments—83.08% and 83.58% for the uncalibrated method and our method, respectively. The proposed method is also approximately 20% better on calibration than temperature scaling, on average.

Fig. 1: Overall performance of uncalibrated models, temperature scaling, and the proposed method. A dot shows the ECE, Accuracy, or ECE vs. Accuracy of one model.

Fig. 2: The training/testing loss and accuracy for both uncalibrated model (left) and the proposed method (right) on Mendeley dataset.

Fig. 3: Four t-SNE plots visualize the learned features for temperature scaling and the proposed method on Mendeley dataset (Left Two) and Kather dataset (Right Two).

* gb.liang@uky.edu (www.gb-liang.com)