Semi-supervised Learning Models, Group of Methods
“MixUp”
Outline

- MixUp (Zhang et al., ICLR 2018)
  - data augmentation method: train models on convex combination two inputs and two labels
- MixUp strategies for semi-supervised learning
  - ICT (Verma et al., IJCAI 2019)
  - MixMatch (Berthelot et al., NeurIPS 2019)
- MixUp strategies in NLP (Guo et al., 2019, Thulasidasan et al., NeurIPS 2019)
  - Word embeddings
  - Final vector representation
MixUp: Intuition

- ML models rely and are overconfident on existing examples
- We need to fill in the gap between the points
- The paper bases on vicinal risk minimization (VRM) theory instead of empirical risk minimization (ERM)
  - train on augmented examples neighboring original examples
MixUp: Data augmentation

- Convex combination of two examples

  \[ \lambda \sim \text{Beta}(\alpha, \alpha) \]
  \[ (x_i, y_i), (x_j, y_j) \sim D_l \]
  \[ x_i' = \lambda x_i + (1 - \lambda)x_j \]
  \[ y_i' = \lambda y_i + (1 - \lambda)y_j \]

- Vicinal objective function

  \[ \mathcal{L}(\theta) = \ell(p(y|x_i'; \theta), y_i') \]
MixUp: Benefits

- Agnostic data augmentation: work on image, speech, text (with modification), tabular data, stabilizing GANs
- Generalize better to out-of-distribution examples
- Robustness to adversarial examples
- Smooth decision boundaries, reduce memorization and overfitting
MixUp strategies for semi-supervised learning: ICT (Verma et al., IJCAI 2019)

- Mixing unlabeled examples
MixUp strategies for semi-supervised learning: MixMatch (Berthelot et al., NeurIPS 2019)

- Mixing both labeled and unlabeled examples as follows
  - Label predictions for unlabeled examples, averaging output distribution for $K$ augmented unlabeled examples
    \[
    y_{ul} = \frac{1}{K} \sum_{k=1}^{K} p(y|\hat{u}_u; \theta)
    \]
  - Sharpening (temperature annealing) label predictions, forcing output distribution towards one-hot vector by adjusting temperature $T \to 0$
    \[
    y_{ul,c} = \frac{\frac{1}{T} y_{ul,c}}{\sum_{c=1}^{C} \frac{1}{T} y_{ul,c}}
    \]
  - MixUp all labeled and unlabeled examples
- Warning: during training, we pass $K$ duplicated unlabeled images at the same time, so we need to adjust batchnorm updates to only take one set unlabeled images
NLP MixUp Strategies

- Word embeddings
  - Mixing embeddings of aligned words between two sentences (padding sentences if one is shorter than the other)
- Final vector representation
## NLP MixUp Results

<table>
<thead>
<tr>
<th>Random Tune</th>
<th>Trec</th>
<th>SST-1</th>
<th>SST-2</th>
<th>Subj</th>
<th>MR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN- KIM Impl. (Kim, 2014b)</td>
<td>91.2</td>
<td>45.0</td>
<td>82.7</td>
<td>89.6</td>
<td>76.1</td>
</tr>
<tr>
<td>CNN- HarvardNLP Impl. ¹</td>
<td>88.2</td>
<td>42.2</td>
<td>83.5</td>
<td>89.2</td>
<td>75.9</td>
</tr>
<tr>
<td>CNN - Our Impl.</td>
<td>90.2±0.20</td>
<td>43.6±0.19</td>
<td>82.3±0.47</td>
<td>90.6±0.45</td>
<td>75.5±0.36</td>
</tr>
<tr>
<td>CNN+wordMixup</td>
<td>90.9±0.42</td>
<td>45.2±0.90</td>
<td>82.8±0.45</td>
<td>92.9±0.41</td>
<td>78.0±0.39</td>
</tr>
<tr>
<td>CNN+senMixup</td>
<td>92.1±0.31</td>
<td>45.2±0.22</td>
<td>83.0±0.35</td>
<td>92.7±0.38</td>
<td>77.9±0.76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Tune</th>
<th>Trec</th>
<th>SST-1</th>
<th>SST-2</th>
<th>Subj</th>
<th>MR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-StanfordNLP Impl. (Tai et al., 2015)</td>
<td>N/A</td>
<td>46.4</td>
<td>84.9</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>LSTM-AgrLearn Impl. (Guo et al., 2018a)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>90.2</td>
<td>76.2</td>
</tr>
<tr>
<td>LSTM - Our Impl.</td>
<td>86.5±0.61</td>
<td>45.9±0.58</td>
<td>84.4±0.35</td>
<td>90.9±0.42</td>
<td>77.2±0.75</td>
</tr>
<tr>
<td>LSTM + wordMixup</td>
<td>90.5±0.50</td>
<td>48.2±0.18</td>
<td>86.3±0.35</td>
<td>93.1±0.49</td>
<td>78.0±0.33</td>
</tr>
<tr>
<td>LSTM + senMixup</td>
<td>89.4±0.40</td>
<td>48.3±0.77</td>
<td>86.7±0.33</td>
<td>91.9±0.34</td>
<td>77.9±0.33</td>
</tr>
</tbody>
</table>
Thank you!