Local Clustering with Mean Teacher for Semi-supervised learning

Authors: Zexi Chen, Benjamin Button, Bharathkumar Ramachandra, Tianfu Wu, Ranga Raju Vatsavai
Introduction to Consistency-based approaches

Consistency-based approaches for semi-supervised learning (SSL)
• Encourage consistent probability predictions between a teacher-student pair for the same data under perturbations

Loss term: \( \mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{cons} \)

\[ \mathcal{L}_{cons} = \mathbb{E}_{\{x_i\}_{i=1}^{n_l} + \{x_i\}_{i=1}^{n_u}} \left[ D\left[ f(x_i; \xi'; \theta'), f(x_i; \xi''; \theta'') \right] \right] \]

Figure 1. General framework for consistency-based approaches
Research Motivation

Issue with consistency-based methods: Confirmation bias

- Caused by inaccurate learning targets generated by the teacher model
- It would trap some unlabeled data samples in low-density regions or enforce them into high-density regions of incorrect class in feature space.

Figure 2. General framework for consistency-based approaches
Method-Idea illustration

**Clustering assumption:** Samples are likely to have the same class label if there is a path connecting them passing through regions of high density only.

**Local consistency assumption:** Nearby samples are likely to have the same label. Samples on the same structure (typically, a manifold) are likely to have the same label.

Propose a local clustering method that clusters data points locally by minimizing the pairwise distance between neighboring points in feature space.

![Figure 3](image)

Figure 3. An illustration of the intuition behind Local clustering in feature space. Each point represents the intermediate learned representation of one data sample.
Related Work

Several types of consistency-based methods:

\[ \mathcal{L}_{cons} = \mathbb{E}_{\{x_i\}^{n_i+n_u}_{i=1}} D[f(x_i, \xi'; \theta'), f(x_i, \xi''; \theta'')] \]

\( \Pi \) model\(^2\)

\[ \theta' = \theta'' \]

Temporal Ensembling\(^2\)
- The EMA of probability predictions of the student model as the teacher model’s predictions

Virtual Adversarial Training (VAT)\(^3\)
- Impose adversarial perturbations to either the inputs or intermediate feature vectors that would maximize the difference in predictions between the student model and teacher model.

Mean Teacher\(^1\)

\[ \theta''_t = \alpha \theta'_{t-1} + (1 - \alpha) \theta' \]

Note: the citation-style references are presented in the end of the slides
Method

Local clustering:
• Build on top of Mean Teacher
• Idea is to pull those misclassified unlabeled data to the high-density regions of their correct class in feature space
• Can be treated as a new regularizer to Mean Teacher
Method

Local clustering:
• At each training iteration, build a weighted graph in feature space from a sub-batch of labeled data and a sub-batch of unlabeled data

The local clustering loss:

\[ \mathcal{L}_{lc} = \mathbb{E}_{\{x_i\}_{i=1}^{n_l}, \{x_j\}_{j=1}^{n_u}} \left[ w_{ij} \| g(x_i; \theta'_g) - g(x_j; \theta'_g) \|^2 \right] + \mathbb{E}_{\{x_m\}_{m=1}^{n_u}, \{x_n\}_{n=1}^{n_u}} \left[ w_{mn} \| g(x_m; \theta'_g) - g(x_n; \theta'_g) \|^2 \right] \]

The edge weight function

\[ w_{ij} = \begin{cases} \exp \left( -\frac{\| z_i - z_j \|^2}{\epsilon} \right) & \text{if } \| z_i - z_j \|^2 \leq \epsilon \\ 0 & \text{otherwise} \end{cases} \]

Total loss:

\[ \mathcal{L} = \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{cons} + \lambda_2 \mathcal{L}_{lc} \]
Results – Datasets

We conduct experiments on two widely used semi-supervised image classification benchmark datasets: SVHN and CIFAR-10

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of training images</th>
<th>Number of test images</th>
<th>Number of action classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVHN</td>
<td>73,257</td>
<td>26,032</td>
<td>10</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>50,000</td>
<td>10,000</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 8: Video action recognition datasets used for this task

Figure 4. Sample images in SVHN (left) and CIFAR-10 (right)
Results

Compares with state-of-the-art methods on SVHN and CIFAR-10

- Train the models on SVHN training images with 500 and 1,000 randomly labeled samples
- Train the models on CIFAR-10 training images with 2,000 and 4,000 randomly labeled samples
- Test error rate percentage is reported as evaluation metric

<table>
<thead>
<tr>
<th>Method</th>
<th>SVHN</th>
<th></th>
<th>CIFAR-10</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n_l = 500$</td>
<td>$n_l = 1000$</td>
<td>$n_l = 2,000$</td>
<td>$n_l = 4,000$</td>
</tr>
<tr>
<td>FM-GAN (Salimans et al. 2016)</td>
<td>18.44 ± 4.80</td>
<td>8.11 ± 1.30</td>
<td>19.61 ± 2.09</td>
<td>18.63 ± 2.32</td>
</tr>
<tr>
<td>Bad GAN (Dai et al. 2017)</td>
<td>-</td>
<td>7.42 ± 0.65</td>
<td>-</td>
<td>14.41 ± 0.30</td>
</tr>
<tr>
<td>Local GAN (Qi et al. 2018)</td>
<td>5.48 ± 0.29</td>
<td>4.73 ± 0.29</td>
<td>-</td>
<td>14.23 ± 0.27</td>
</tr>
<tr>
<td>II model (Laine and Aila 2016)</td>
<td>6.65 ± 0.53</td>
<td>4.82 ± 0.17</td>
<td>-</td>
<td>12.36 ± 0.31</td>
</tr>
<tr>
<td>TempEns (Laine and Aila 2016)</td>
<td>5.12 ± 0.13</td>
<td>4.42 ± 0.16</td>
<td>-</td>
<td>12.16 ± 0.31</td>
</tr>
<tr>
<td>Mean Teacher (Tarvainen and Valpola 2017)</td>
<td>4.18 ± 0.27</td>
<td>3.95 ± 0.19</td>
<td>15.73 ± 0.31</td>
<td>12.31 ± 0.28</td>
</tr>
<tr>
<td>VAdD (Park et al. 2018)</td>
<td>-</td>
<td>4.16 ± 0.08</td>
<td>-</td>
<td>11.32 ± 0.11</td>
</tr>
<tr>
<td>VAT + EntMin (Miyato et al. 2018)</td>
<td>-</td>
<td>3.86 ± 0.11</td>
<td>-</td>
<td>10.55 ± 0.05</td>
</tr>
<tr>
<td>TempEns + SNTG (Luo et al. 2018)</td>
<td>4.46 ± 0.26</td>
<td>3.98 ± 0.21</td>
<td>13.64 ± 0.32</td>
<td>10.93 ± 0.14</td>
</tr>
<tr>
<td>MT + SNTG (Luo et al. 2018)</td>
<td>3.99 ± 0.24</td>
<td>3.86 ± 0.27</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MT*</td>
<td>3.91 ± 0.11</td>
<td>3.80 ± 0.09</td>
<td>12.37 ± 0.29</td>
<td>9.93 ± 0.16</td>
</tr>
<tr>
<td><strong>MT + LC (ours)</strong></td>
<td><strong>3.54 ± 0.17</strong></td>
<td><strong>3.35 ± 0.09</strong></td>
<td><strong>11.56 ± 0.31</strong></td>
<td><strong>9.26 ± 0.16</strong></td>
</tr>
</tbody>
</table>

Table 1: Error rate percentage comparison with the sota methods on SVHN and CIFAR-10
Results

Visualize the test error as a function of training epoch on SVHN and CIFAR-10

Figure 5. Smoothed test error curves of MT and MT+LC on SVHN (left) and CIFAR-10 (right)
Results

Ablation Study:
• Study the effect of **cut-off threshold** $\epsilon$
• Study the effect of **LC loss weight** $\lambda_2$

Figure 6. The test errors of MT + LC with different cut-off thresholds (left) and loss weights (right)
Visualization

Visual comparison of MT and MT + LC on intermediate feature representations on test data

Figure 7. t-SNE visualization of CIFAR-10 test data features obtained by MT (left) and MT + LC (right)
Summary & future work

• We developed a novel local clustering method to address the confirmation bias issues existing in Mean Teacher method

• Future work:
  • Design adaptive cut-off threshold
  • Use a pair-wise metric learning method to determine the pairwise similarity
References for slide 5


Thank You!