IOMatch: Simplifying Open-Set Semi-Supervised Learning with Joint Inliers and Outliers Utilization

Zekun Li\textsuperscript{1}, Lei Qi\textsuperscript{2}, Yinghuan Shi\textsuperscript{1*}, Yang Gao\textsuperscript{1}

\textsuperscript{1}Nanjing University  \textsuperscript{2}Southeast University
Background

• Standard Semi-SL assumptions can be hard to satisfy.
  • In practice, unlabeled data may contain unseen classes (outliers).
  • Existing Semi-SL methods suffer from open-set unlabeled data.
    • It is impossible to generate correct close-set pseudo-labels for outliers.

Illustration of Open-Set SSL

[NeurIPS'18] Oliver et al.
Motivation

• Intuition: Outliers are harmful? Remove them first!
  • It is a common strategy employed in previous works:
    • Detect the outliers first and then filter them out of pseudo-labeling.
    • Detection based on predictions or with additional network modules:

Motivation

• The intuitive detect-and-filter strategy can easily fail.
  • We can hardly obtain a reliable outlier detector at the beginning.
    • Especially when labels are extremely scarce.
  • An unreliable detector harms more than outliers themselves.

Motivation

- The intuitive detect-and-filter strategy can easily fail.
  - We can hardly obtain a reliable outlier detector at the beginning.
    - Especially when labels are extremely scarce.
  - An unreliable detector harms more than outliers themselves.
    - Numerous inliers may be wrongly removed.
    - Such errors are difficult to rectify.

Can we utilize open-set unlabeled data without exactly distinguishing between inliers and outliers?
Approach

• **Key idea:** exploit unified open-set targets.
  • A standard closed-set classifier to predict an unlabeled sample
    • Most likely to belong to which seen class \((c_1/c_2/c_3)\)
      \[ p = [p_{c_1}, p_{c_2}, p_{c_3}] = [0.7, 0.2, 0.1] \]
  • An extra multi-binary classifier to predict
    • Probability of truly belonging to each seen class or not
      \[ o_{c_1} = [o_{c_1}, \overline{o_{c_1}}] = [0.4, 0.6] \]
      \[ o_{c_2} = [o_{c_2}, \overline{o_{c_2}}] = [0.1, 0.9] \]
      \[ o_{c_3} = [o_{c_3}, \overline{o_{c_3}}] = [0.2, 0.8] \]
Key idea: exploit unified open-set targets.

- Fuse these two predictions to estimate the likelihood of a sample
  - Being an inlier of $c_1$: $p_{c_1} \times o_{c_1} = 0.7 \times 0.4 = 0.28$
  - Being an outlier similar to $c_1$: $p_{c_1} \times \overline{o_{c_1}} = 0.7 \times 0.6 = 0.42$
  - Same for other seen classes...
    - Being an inlier of $c_1/c_2/c_3$: $[0.28, 0.02, 0.02]$
    - Being an outlier: $0.42 + 0.18 + 0.08 = 0.68$

- Then we obtain the open-set target:
  - Probability of [Bird, Cat, Dog, Outlier] = $[0.28, 0.02, 0.02, 0.68]$
• **Key idea:** exploit unified open-set targets.
  • Unified open-set targets are produced for both inliers and outliers.
  • Optimize an open-set classifier via pseudo-labeling.

![Diagram](image)
Approach

• **IOMatch demonstrates remarkable simplicity.**
  - All the classifiers in IOMatch are concurrently optimized.
    - No more need for a pre-training (warm-up) stage for an outlier detector.
  - All the learning objectives are cross-entropy losses.
    - Easy for implementation.
    - Easy to tune hyper-parameters.
IOMatch achieves impressive performance.

- Compared with the SOTA standard and open-set Semi-SL methods.
- For both closed-set and open-set evaluation.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class split (Seen / Unseen)</td>
<td>6 / 4</td>
<td>20 / 80</td>
</tr>
<tr>
<td>Number of labels per class</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>MixMatch [3]</td>
<td>NeurIPS’19</td>
<td>43.08 ± 1.79</td>
</tr>
<tr>
<td>ReMixMatch [2]</td>
<td>ICLR’20</td>
<td>72.82 ± 1.81</td>
</tr>
<tr>
<td>FixMatch [30]</td>
<td>NeurIPS’20</td>
<td>81.58 ± 6.63</td>
</tr>
<tr>
<td>CoMatch [20]</td>
<td>ICCV’21</td>
<td>86.08 ± 1.08</td>
</tr>
<tr>
<td>FlexMatch [14]</td>
<td>NeurIPS’21</td>
<td>73.34 ± 4.42</td>
</tr>
<tr>
<td>SimMatch [47]</td>
<td>CVPR’22</td>
<td>79.84 ± 4.76</td>
</tr>
<tr>
<td>FreeMatch [37]</td>
<td>ICLR’23</td>
<td>79.26 ± 4.11</td>
</tr>
<tr>
<td>UASD [7]</td>
<td>AAAI’20</td>
<td>35.25 ± 1.07</td>
</tr>
<tr>
<td>DS3L [10]</td>
<td>ICML’20</td>
<td>39.09 ± 1.24</td>
</tr>
<tr>
<td>MTCF [42]</td>
<td>ECCV’20</td>
<td>49.15 ± 6.12</td>
</tr>
<tr>
<td>T2T [16]</td>
<td>ICCV’21</td>
<td>73.89 ± 1.55</td>
</tr>
<tr>
<td>OpenMatch [27]</td>
<td>NeurIPS’21</td>
<td>43.63 ± 3.26</td>
</tr>
<tr>
<td>SAFE-STUDENT [14]</td>
<td>CVPR’22</td>
<td>59.28 ± 1.18</td>
</tr>
</tbody>
</table>

IOMatch | Ours | 89.68 ± 2.04 | 93.87 ± 0.16 | 53.73 ± 2.12 | 67.28 ± 1.10 | 56.31 ± 2.29 | 69.77 ± 0.58 | 50.83 ± 0.99 | 64.75 ± 0.52 |
Conclusions

• In open-set Semi-SL, it is really challenging, but not mandatory, to exactly identify outliers before pseudo-labeling.

• What truly matters is the idea of joint inliers and outliers utilization.
  • Producing unified open-set targets is just one approach for this.

• We are working towards more realistic Semi-SL!
  • Tackling more practical challenges: imbalanced class distribution, domain shifts, and fine-grained categories…
  • With stronger techniques: self-supervised learning, LLMs, and VLMs…
Looking forward to further discussion!

10:30 am – 12:30 pm

Poster #152 @ Room Nord

Code: https://github.com/nukezil/IOMatch


Contact me: lizekun@smail.nju.edu.cn