PCL: Proxy-based Contrastive Learning for Domain Generalization

Xufeng Yao\textsuperscript{1}, Yang Bai\textsuperscript{1}, Xinyun Zhang\textsuperscript{1}, Yuechen Zhang\textsuperscript{1}, Qi Sun\textsuperscript{1}, Ran Chen\textsuperscript{1}, RuiYu Li\textsuperscript{2}, Bei Yu\textsuperscript{1}

\textsuperscript{1}The Chinese University of Hong Kong
\textsuperscript{2}SmartMore
{xfyao,byu}@cse.cuhk.edu.hk

May. 27, 2022
Background and Motivation
• DG aims to train a model from multiple source domains that can generalize well on target domain.

• Contrastive learning offers a potential solution, but is not effective in DG.

• We aims to use proxy-based contrastive learning to address the problem.
(a) Contrastive-based Loss

Contrastive loss: sample-to-sample pairs

(b) Proxy-based Loss

Proxy loss: proxy-to-sample pairs
### Complexity comparison

<table>
<thead>
<tr>
<th>Loss function</th>
<th>positive pair</th>
<th>negative pair</th>
<th>relations</th>
<th>category</th>
<th>training complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>softmax CE loss</td>
<td>((w_y, x_i))</td>
<td>((w_1, x_i), (w_2, x_i), \ldots, (w_j, x_i))</td>
<td>proxy-to-sample</td>
<td>proxy-based</td>
<td>(O(CN))</td>
</tr>
<tr>
<td>Contrastive loss</td>
<td>((x_i, x_i^*))</td>
<td>((x_i, x_1), (x_i, x_2), \ldots, (x_1, x_n))</td>
<td>sample-to-sample</td>
<td>pair-based</td>
<td>(O(N^2))</td>
</tr>
<tr>
<td>MS Loss</td>
<td>((x_i, x_j) \ldots (x_i, x_m))</td>
<td>((x_i, x_1), (x_i, x_2) \ldots (x_1, x_n))</td>
<td>sample-to-sample</td>
<td>pair-based</td>
<td>(O(N^2))</td>
</tr>
<tr>
<td>Triplet Loss</td>
<td>((x_i, x_j) \ldots (x_i, x_m))</td>
<td>((x_i, x_1), (x_i, x_2) \ldots (x_1, x_n))</td>
<td>sample-to-sample</td>
<td>pair-based</td>
<td>(O(N^3))</td>
</tr>
</tbody>
</table>

- **Pair-based loss**: rich sample-to-sample pairs, high complexity
- **Proxy-based loss**: low complexity, high generalization
Proxy-based Contrastive Learning
Review of softmax CE loss

- **Pros**: Learn a proxy for each class efficiently.
- **Pros**: Low complexity, safe convergence.
- **Cons**: Miss rich sample-to-sample pairs.

\[
\mathcal{L}_{\text{CE}} = -\log \frac{\exp(w_c^T z_i)}{\exp(w_c^T z_i) + \sum_{j=1}^{c-1} \exp(w_j^T z_i)},
\]

(1)
Review of Contrastive Loss

- **Pros**: Leverage dense sample-to-sample pairs.
- **Pros**: Implicit hard pair mining.
- **Cons**: High complexity, unstable convergence.

\[
\mathcal{L}_{CL} = - \log \frac{\exp(\mathbf{z}_i^T \mathbf{z}_+ \cdot \alpha)}{\exp(\mathbf{z}_i^T \mathbf{z}_+ \cdot \alpha) + \sum \exp(\mathbf{z}_i^T \mathbf{z}_- \cdot \alpha)},
\]  
(2)
Implicit hard pair mining in contrastive loss

- By controlling $\alpha$, contrastive loss implicitly conduct hard pair mining.
- Sufficient pairs guarantee the performance.

\[
\mathcal{L}_{CL} = \lim_{\alpha \to \infty} \frac{1}{\alpha} - \log\left( \frac{\exp(\alpha \cdot s_p)}{\exp(\alpha \cdot s_p) + \sum_{j=1}^{N-1} \exp(\alpha \cdot s^j_n)} \right) \\
= \lim_{\alpha \to \infty} \frac{1}{\alpha} \log(1 + \sum_{j=1}^{N-1} \exp(\alpha(s^j_n - s_p))) \\
= \max[s^j_n - s_p]_+, \tag{3}
\]
High complexity may impede the performance

Gradients of the positive similarity score

$\sum_{j=1}^{N-1} \frac{\partial L}{s_n^j} = \left| \frac{\partial L}{s_p} \right|$
Combine Softmax CE and Contrastive Loss

- **Softmax**: Low complexity, overlook sample-to-sample pairs
- **Contrastive**: High complexity, rich pairs, unstable convergence.

\[
\mathcal{L}_{\text{PCL}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathbf{w}_c^\top \mathbf{z}_i \cdot \alpha)}{Z},
\]

(4)

where \( Z \) is given by:

\[
Z = \exp(\mathbf{w}_c^\top \mathbf{z}_i \cdot \alpha) + \sum_{k=1}^{C-1} \exp(\mathbf{w}_k^\top \mathbf{z}_j \cdot \alpha) + \sum_{j=1, j \neq i}^{K} \exp(\mathbf{z}_i^\top \mathbf{z}_j \cdot \alpha).
\]

(5)
Experimental Results
Visualizaiton of learned features

Baseline

PCL

Ours
Table: Comparison with state-of-the-art methods on OfficeHome benchmark with ResNet-50 imagenet-pretrained model

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>A</th>
<th>C</th>
<th>P</th>
<th>R</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixstyle</td>
<td>51.1</td>
<td>53.2</td>
<td>68.2</td>
<td>69.2</td>
<td>60.4</td>
</tr>
<tr>
<td>SagNet</td>
<td>63.4</td>
<td>54.8</td>
<td>75.8</td>
<td>78.3</td>
<td>68.1</td>
</tr>
<tr>
<td>CORAL</td>
<td>65.3</td>
<td>54.4</td>
<td>76.5</td>
<td>78.4</td>
<td>68.7</td>
</tr>
<tr>
<td>SWAD</td>
<td>66.1</td>
<td>57.7</td>
<td>78.4</td>
<td>80.2</td>
<td>70.6</td>
</tr>
<tr>
<td>Ours</td>
<td>67.3</td>
<td>59.9</td>
<td>78.7</td>
<td>80.7</td>
<td>71.6</td>
</tr>
</tbody>
</table>

1 Zhou; Kaiyang; et al. Domain generalization with mixstyle. ICLR 2021.
2 Nam; Hyeonseob; et al. Reducing domain gap by reducing style bias. CVPR. 2021.
3 Sun; Baochen; and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. ECCV 2016.
4 Cha; Junbum; et al. Swad: Domain generalization by seeking flat minima. NeurIPS; 2021.
Table: Comparison with state-of-the-art methods on PACS benchmark with ResNet-50 imagenet-pretrained model

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>A</th>
<th>C</th>
<th>P</th>
<th>S</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixstyle</td>
<td>86.8</td>
<td>79.0</td>
<td>96.6</td>
<td>78.5</td>
<td>85.2</td>
</tr>
<tr>
<td>CORAL</td>
<td>88.3</td>
<td>80.0</td>
<td>97.5</td>
<td>78.8</td>
<td>86.2</td>
</tr>
<tr>
<td>SagNet</td>
<td>87.4</td>
<td>80.7</td>
<td>97.1</td>
<td>80.0</td>
<td>86.3</td>
</tr>
<tr>
<td>SWAD</td>
<td>89.3</td>
<td>83.4</td>
<td>97.3</td>
<td>82.5</td>
<td>88.1</td>
</tr>
<tr>
<td>Ours</td>
<td>90.2</td>
<td>83.9</td>
<td>98.1</td>
<td>82.6</td>
<td>88.7</td>
</tr>
</tbody>
</table>

5 Zhou; Kaiyang; et al. Domain generalization with mixstyle. ICLR 2021.
6 Sun; Baochen; and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. ECCV; 2016.
7 Nam; Hyeonseob; et al. Reducing domain gap by reducing style bias. CVPR. 2021.
Table: Comparison with state-of-the-art methods on TerraIncognita benchmark with ResNet-50 imagenet-pretrained model

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Location100</th>
<th>Location38</th>
<th>Location43</th>
<th>Location46</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixstyle⁹</td>
<td>54.3</td>
<td>34.1</td>
<td>55.9</td>
<td>31.7</td>
<td>44.0</td>
</tr>
<tr>
<td>CORAL¹⁰</td>
<td>51.6</td>
<td>42.2</td>
<td>57.0</td>
<td>39.8</td>
<td>47.7</td>
</tr>
<tr>
<td>SagNet¹¹</td>
<td>53.0</td>
<td>43.0</td>
<td>57.9</td>
<td>40.4</td>
<td>48.6</td>
</tr>
<tr>
<td>SWAD¹²</td>
<td>55.4</td>
<td>44.9</td>
<td>59.7</td>
<td>39.9</td>
<td>50.0</td>
</tr>
<tr>
<td>Ours</td>
<td>58.7</td>
<td>46.3</td>
<td>60.0</td>
<td>43.6</td>
<td>52.1</td>
</tr>
</tbody>
</table>

---

¹⁰ Sun; Baochen; and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. ECCV; 2016.
¹¹ Nam; Hyeonseob; et al. Reducing domain gap by reducing style bias. CVPR. 2021.
### Experimental Results

Table: Comparison with state-of-the-art methods on DomainNet benchmark with ResNet-50 ImageNet pre-trained model

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>clip</th>
<th>info</th>
<th>paint</th>
<th>quick</th>
<th>real</th>
<th>sketch</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixstyle(^{13})</td>
<td>51.9</td>
<td>13.3</td>
<td>37.0</td>
<td>12.3</td>
<td>46.1</td>
<td>43.4</td>
<td>34.0</td>
</tr>
<tr>
<td>SagNet(^{14})</td>
<td>57.7</td>
<td>19.0</td>
<td>45.3</td>
<td>12.7</td>
<td>58.1</td>
<td>48.8</td>
<td>40.3</td>
</tr>
<tr>
<td>CORAL(^{15})</td>
<td>59.2</td>
<td>19.7</td>
<td>46.6</td>
<td>13.4</td>
<td>59.8</td>
<td>50.1</td>
<td>41.5</td>
</tr>
<tr>
<td>SWAD(^{16})</td>
<td>66.0</td>
<td>22.4</td>
<td>53.5</td>
<td>16.1</td>
<td>65.8</td>
<td>55.5</td>
<td>46.5</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>67.9</strong></td>
<td><strong>24.3</strong></td>
<td><strong>55.3</strong></td>
<td><strong>15.7</strong></td>
<td><strong>66.6</strong></td>
<td><strong>56.4</strong></td>
<td><strong>47.7</strong></td>
</tr>
</tbody>
</table>

\(^{13}\) Zhou; Kaiyang; et al. Domain generalization with mixstyle. ICLR 2021.  
\(^{14}\) Nam; Hyeonseob; et al. Reducing domain gap by reducing style bias. CVPR. 2021.  
\(^{15}\) Sun; Baochen; and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. ECCV; 2016.  
\(^{16}\) Cha; Junbum; et al. Swad: Domain generalization by seeking flat minima. NeurIPS; 2021.
THANK YOU!