Learning to Self-Train for Semi-Supervised Few-Shot Classification

Xinzhe Li, Qianru Sun, Yaoyao Liu, Shibao Zheng, Qin Zhou, Tat-Seng Chua, Bernt Schiele

NeurIPS 2019
Motivation

Leverage unlabeled data for FSL

Meta-learn to cherry pick the data

Learning to minimize the noise from unlabeled data
Pseudo-labeling:

- **Initialization**
  - Support set $S$
  - Unlabeled set $R$

- **Few-shot training**
  - predicting

- **Cherry-picking**
  - Pseudo-labeled (noisy)
  - Hard selection
  - Soft weighting
  - Pseudo-labeled set $R^p$ (selected $\rightarrow$ weighted)

Re-training:

- **Initialization**
  - Pseudo-labeled set $R^p$
  - Support set $S$

- **Re-training**
  - re-training

Val./Test:

- **Fine-tuning**
  - fine-tuned model

- **Validation or Test**
  - Query set $Q$
  - Loss or accuracy

Detailed steps of Cherry-picking:

- **Select**
  - Top-Z per class
  - Pseudo-labeled (selected)

- **Concatenate**
  - With every prototype
  - Prototypes of classes

- **Copy**
  - Pseudo-labeled (selected)

- **SWN**
  - Soft weights

- **Soft weights**
  - Pseudo-labeled set $R^p$ (weighted)
Pseudo-Labeling

Train a fast learner using support set

\[ \theta_t \leftarrow \theta_{t-1} - \alpha \nabla_{\theta_{t-1}} L(S; \Phi_{ss}, \theta_{t-1}) \]

Label the unsupervised dataset

\[ Y^R = f_{[\Phi_{ss}, \theta_T]}(R), \]
Cherry picking

Soft-weight the semi-supervised samples

\[ w_{i,c} = f_{\Phi_{swn}} \left( \left[ f_{\Phi_{ss}}(x_i); \frac{\sum_k f_{\Phi_{ss}}(x_{c,k})}{K} \right] \right), \]
Self-train

Merge support set and weakly-labeled data and retrain

\[ \theta_t \leftarrow \theta_{t-1} - \alpha \nabla_{\theta_{t-1}} L(S \cup R^p; [\Phi_{swn}, \Phi_{ss}, \theta_{t-1}]), \]

\[ L(S \cup R^p; [\Phi_{swn}, \Phi_{ss}, \theta_t]) = \begin{cases} 
L_{ce}(f_{[\Phi_{swn}, \Phi_{ss}, \theta_t]}(x_i), y_i), & \text{if } (x_i, y_i) \in S, \\
L_{ce}(w_i \odot f_{[\Phi_{swn}, \Phi_{ss}, \theta_t]}(x_i), y_i), & \text{if } (x_i, y_i) \in R^p,
\end{cases} \]

Update params at different time step

\[ \Phi_{swn} =: \Phi_{swn} - \beta_1 \nabla_{\Phi_{swn}} L(Q; [\Phi_{swn}, \Phi_{ss}, \theta_m]), \]

\[ [\Phi_{ss}, \theta'] =: [\Phi_{ss}, \theta'] - \beta_2 \nabla_{[\Phi_{ss}, \theta']} L(Q; [\Phi_{swn}, \Phi_{ss}, \theta_T]), \]
Inner loop of self-train

Outer loop

Input an episode $S$, $Q$ and $R$

Inner loop

- support set $S$
- pseudo labeled set $R^p$
- unlabeled set $R$

Re-train step 1

$\theta_1 \rightarrow \theta_{m-1} \rightarrow \theta_m \rightarrow \theta_{m+1} \rightarrow \theta_{T-1}$

Re-train step $m$

$\theta_m(\Phi_{swn})$

Fine-tune step $m+1$

Query set $Q$

$f[\Phi_{ss};\theta_m(\Phi_{swn})]$

loss of $\Phi_{swn}$

Fine-tune step $T$

Query set $Q$

$f[\Phi_{ss};\theta_T([\Phi_{ss},\theta'])]$

loss of $[\Phi_{ss},\theta']$

meta update $[\Phi_{swn}, \Phi_{ss}, \theta']$

deploy
<table>
<thead>
<tr>
<th>Few-shot Learning Method</th>
<th>Backbone</th>
<th>miniImageNet (test)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1-shot</td>
<td>5-shot</td>
<td></td>
</tr>
<tr>
<td><strong>Data augmentation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delta-encoder, [29]</td>
<td>VGG-16 (pre)</td>
<td>58.7</td>
<td>73.6</td>
<td></td>
</tr>
<tr>
<td><strong>Gradient descent</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAML, [3]</td>
<td>4 CONV</td>
<td>48.70 ± 1.75</td>
<td>63.11 ± 0.92</td>
<td></td>
</tr>
<tr>
<td>Bilevel Programming, [5]</td>
<td>ResNet-12⁺</td>
<td>50.54 ± 0.85</td>
<td>64.53 ± 0.68</td>
<td></td>
</tr>
<tr>
<td>MetaGAN, [42]</td>
<td>ResNet-12</td>
<td>52.71 ± 0.64</td>
<td>68.63 ± 0.67</td>
<td></td>
</tr>
<tr>
<td>adaResNet, [19]</td>
<td>ResNet-12⁺</td>
<td>56.88 ± 0.62</td>
<td>71.94 ± 0.57</td>
<td></td>
</tr>
<tr>
<td>LEO, [27]</td>
<td>WRN-28-10 (pre)</td>
<td>61.76 ± 0.08</td>
<td>77.59 ± 0.12</td>
<td></td>
</tr>
<tr>
<td>MTL, [32]</td>
<td>ResNet-12 (pre)</td>
<td>61.2 ± 1.8</td>
<td>75.5 ± 0.9</td>
<td></td>
</tr>
<tr>
<td>MetaOpt-SVM, [10]⁺</td>
<td>ResNet-12</td>
<td>62.64 ± 0.61</td>
<td>78.63 ± 0.46</td>
<td></td>
</tr>
<tr>
<td><strong>LST (Ours)</strong></td>
<td>recursive, hard, soft</td>
<td>ResNet-12 (pre)</td>
<td><strong>70.1 ± 1.9</strong></td>
<td><strong>78.7 ± 0.8</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Few-shot Learning Method</th>
<th>Backbone</th>
<th>tieredImageNet (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1-shot</td>
</tr>
<tr>
<td><strong>Gradient descent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAML, [3] (by [13])</td>
<td>ResNet-12</td>
<td>51.67 ± 1.81</td>
</tr>
<tr>
<td>LEO, [27]</td>
<td>WRN-28-10 (pre)</td>
<td>66.33 ± 0.05</td>
</tr>
<tr>
<td>MTL, [32] (by us)</td>
<td>ResNet-12 (pre)</td>
<td>65.6 ± 1.8</td>
</tr>
<tr>
<td>MetaOpt-SVM, [10]⁺</td>
<td>ResNet-12</td>
<td>65.99 ± 0.72</td>
</tr>
<tr>
<td><strong>LST (Ours)</strong></td>
<td>recursive, hard, soft</td>
<td>ResNet-12 (pre)</td>
</tr>
</tbody>
</table>