Meta-Learning of Structured Representation by Proximal Mapping

Mao Li, Yingyi Ma, Xinhua Zhang
University of Illinois at Chicago
Motivation

Goal of meta-learning: Extract **prior structures** from a set of **tasks** that allows efficient learning of **new tasks**.

Examples of structural regularities:

- **Instance level**
  - **Input layers**: transformation beyond group-based diffeomorphism
  - **Within layers**: sparsity, disentanglement, spatial invariance, structured gradient accounting for data covariance, manifold smoothness
  - **Between layers**: equivariance, contractivity, robustness under dropout and adversarial perturbations of preceding nodes

- **Batch/Dataset level**
  - multi-view, multi-modality, multi-domain
  - diversity, fairness, privacy, causal structure
Existing Approaches

- Data Augmentation

✓ boost prediction performance

✗ unclear the improvement is due to the learned representation or due to a better classifier.
Existing Approaches

- **Auto-encoder**

  ![Diagram of auto-encoder](image)

  \[ \min \ell(x, \hat{x}) \]

  - √ learned the most salient features
  - × usually used as an initialization for subsequent supervised task
  - × not amendable to end-to-end learning

  **Our goal**: learn representations that explicitly encode structural priors in an end-to-end fashion.
Existing Approaches

- Regularization

\[ \min_f \ \text{Empirical Risk}(f) + R(f) \]

✓ simple and efficient

✗ contention of weights between regularizer and supervised performance
Proposed Method

Morph a representation $z$ towards a structured one by proximal mapping:

$$z \mapsto \arg\min_{x \in C} \frac{\lambda}{2} \| x - z \|^2 + L(x)$$

- $z$: mini-batch or single-example
- a mini-batch $\leftrightarrow$ a task in meta-learning
- proximal mapping $\leftrightarrow$ task-specific base learner

Embed the proximal mapping as a layer into deep networks

Advantages

- Decoupling the regularization and supervised learning
- Extend meta-learning to unsupervised base learners
Proposed Method

Morph a representation $z$ towards a structured one by proximal mapping:

\[
    z \mapsto \arg\min_{x \in C} \frac{\lambda}{2} \| x - z \|^2 + L(x)
\]

$\mathbf{L}$: graph-Laplacian (for smoothness on manifold)
MetaProx for Multi-view Learning

In multiview learning, observations are available as pairs of views: \( \{x_i, y_i\} \).

Figure 1: training framework of MetaProx
MetaProx for Multi-view Learning

feature extraction:

$X \rightarrow f(X)$

$Y \rightarrow g(Y)$
MetaProx for Multi-view Learning

② proximal mapping: promote high correlation between two views

\[ \arg \min_{P,Q} \frac{1}{2} \| P - f(X) \|^2 + \frac{1}{2} \| Q - g(Y) \|^2 + \text{CCA}(P, Q) \]

\[ \text{CCA}(P, Q) := \min_{U,V} -\text{tr}(U^T P Q^T V), \]
\[ \text{s.t.} \quad U^T P P^T U = I \]
\[ V^T Q Q^T V = I \]
\[ u_i^T P Q^T v_j = 0, \forall i \neq j \text{ from 1 to } k. \]
MetaProx for Multi-view Learning

Supervised task

\[
\begin{align*}
\min_{f,g,h} \text{loss} \quad & \left( \begin{array}{c}
\arg \min_{P,Q} \frac{1}{2} \| P - f(X) \|^2 \\
+ \frac{1}{2} \| Q - g(Y) \|^2 \\
+ \text{CCA}(P, Q)
\end{array} \right), \text{ground true label}
\end{align*}
\]

\[h: \text{supervised predictor}\]
MetaProx for Multi-view Learning

\[ \text{supervised task} \]

\[
\min_{f,g,h} \text{loss} \left( h \left[ \begin{array}{c} \arg \min_{P,Q} \frac{1}{2} \| P - f(X) \|^2 \\ + \frac{1}{2} \| Q - g(Y) \|^2 \\ + \text{CCA}(P,Q) \end{array} \right], \text{ground true label} \right), \text{optimize over red variables} \]
Experiment Results

Multi-view image classification

- **Dataset**: a subset of Sketchy (20 classes)

\[
\{(\text{butterfly}), \text{‘butterfly’}; \ldots \ldots ; (\text{cat}), \text{‘cat’}\}
\]

![Test accuracy for image classification](image)

**Test accuracy for image classification**
Experiment Results

Crosslingual word embedding

- **Dataset:** WS353, SimLex999
- **Metric:** Spearman’s correlation between the rankings by model and human

Table 1: Spearman’s correlation for word similarities

<table>
<thead>
<tr>
<th></th>
<th>WS-353</th>
<th>WS-SIM</th>
<th>WS-REL</th>
<th>SimLex999</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EN</td>
<td>DE</td>
<td>EN</td>
<td>DE</td>
</tr>
<tr>
<td>Baseline</td>
<td>73.35</td>
<td>52.68</td>
<td>77.84</td>
<td>63.34</td>
</tr>
<tr>
<td>linearCCA</td>
<td>73.79</td>
<td>68.45</td>
<td>76.06</td>
<td>73.02</td>
</tr>
<tr>
<td>DCCA</td>
<td>73.86</td>
<td>69.09</td>
<td><strong>78.69</strong></td>
<td>74.13</td>
</tr>
<tr>
<td>DCCAE</td>
<td>72.39</td>
<td><strong>69.67</strong></td>
<td>75.74</td>
<td>74.65</td>
</tr>
<tr>
<td>MetaProx</td>
<td><strong>75.38</strong></td>
<td>69.19</td>
<td>78.28</td>
<td><strong>75.40</strong></td>
</tr>
<tr>
<td>DEPEMB</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(English, German)

word 1

word 2

word n
At the poster:
More details and discussions

Thanks!

MetaProx ≠

“Efficient Meta Learning via Minibatch Proximal Update” (NeurIPS 2019)

“Meta-Learning with Implicit Gradients” (NeurIPS 2019)

modeling ≠ optimization