Deep Asymmetric Multi-task Feature Learning

Reference:

Reproduced by: Aobo Yang, Yujia Mu, David Yao, Qi Liu

2019/12/05
Motivation

- **Multi-task Learning (original)**
  - Learn deep representation
  - Problem of negative transfer

- **Asymmetric Multi-task Feature Learning (advanced)**
  - Learn deep representation
  - Prevent negative transfer
  - Unscalable and inefficient to deep learning

- **Deep Asymmetric Multi-task Feature Learning (more advanced)**
  - Learn deep representation
  - Prevent negative transfer
  - Less noisy representations
  - Scalable and efficient
Background

- **Multi-task learning:**
  - **Definition:**
    - Jointly train multiple task predictors
    - Allow knowledge transferring
  - **Drawbacks:**
    - Existence of negative transfer

- **Asymmetric Multi-task Feature Learning**
  - **Definition:**
    - Allow asymmetric knowledge transfer through inter-task regulation
    - Proposed to solve the above negative transfer
  - **Drawbacks:**
    - Fails to reconstructed from the combination of parameters for tasks
    - Poorly scalable
Related Work

● **Multitask Learning**
  ○ **Definition**: Jointly train a set of task predictors
  ○ Learning process allows knowledge transfer between predictors
  ○ Main limitation: cannot prevent negative transfer

● **Asymmetric Multitask Learning**
  ○ **Definition**: Break the symmetry in the knowledge transfer direction
  ○ Proposed in order to solve the problem of negative transfer
  ○ Main limitation: not scalable and hard to transfer to deep learning

● **Autoencoders**
  ○ **Definition**: transform input features and decode back to the original
  ○ Use a sparse nonlinear autoencoder term
  ○ Purpose: denoise of the latent features
Target Task

- **Asymmetric Multitask Feature Learning**
  - Learns latent features
  - Weighting up reliable task predictors; Weighting down the unpredictable ones (To prevent negative transfer)
  - Extending multitask learning to DNN with top layer feedback connections

- **Benchmarking**
  - Image classification using both the shallow and deep neural network on synthetic datasets

- **Expected Effects**
  - Better performance
  - More useful features learnt
An Intuitive Figure Showing WHY Claim
Proposed Solution

- Asymmetric multi-task feature learning (AMTFL): a completely new type of regularization to prevent the negative transfer from unreliable tasks to the shared latent features
  - Reconstruct latent features with task predictors’ parameters
  - Enforce reconstruction to be done by reliable tasks only
  - Since task parameters are constructed by features, the reconstruction is like autoencoder

Multiple task parameters \((w)\) are constructed by a set of latent features \((l)\). Unreliable task \((w_3)\) pollutes the latent features.

Encourage asymmetric transfer by using reliable task parameters \((w_1, w_2)\) to reconstruct the latent features \((l)\).
The AMTFL framework is defined as

$$\min_{L,S,A} \sum_{t=1}^{T} \left\{ (1 + \alpha \|a_t^o\|_1) \mathcal{L}(L, s_t; X_t, y_t) + \mu \|s_t\|_1 \right\}$$
$$+ \gamma \|Z - \sigma(ZSA)\|^2_F + \lambda \|L\|^2_F.$$  (6)

Where

$$W = LS$$

$L \in \mathbb{R}^{d \times k}$

$L$ is a collection of $k$ latent base

$S \in \mathbb{R}^{k \times T}$

$S$ is the coefficient matrix for linearly combining the bases

$Z = \sigma(XL)$

Nonnegative feature matrix with ReLU nonlinear transformation

$A \in \mathbb{R}^{T \times k}$

Task-to-feature transfer matrix

The model parameters $W$ can be decomposed to $L$ and $S$
Implementation

\[ \min_{L,S,A} \sum_{t=1}^{T} \left\{ (1 + \alpha \|a_t^0\|_1) \mathcal{L}(L, s_t; X_t, y_t) + \mu \|s_t\|_1 \right\} \]

\[ + \gamma \|Z - \sigma(ZSA)\|_F^2 + \lambda \|L\|_F^2. \]  \hspace{1cm} (6)

L1 regularization to make S sparse. The assumption is that each task sparsely rely on the shared latent vectors.

L2 regularization
Implementation

Sparsity regularization. Multiplied by the amount of training loss, making the ongoing transfer from hard task more sparse than the easy ones

\[
\min_{L,S,A} \sum_{t=1}^{T} \left\{ (1 + \alpha \| a_t^0 \|_1) \mathcal{L}(L, s_t; X_t, y_t) + \mu \| s_t \|_1 \right\} \\
+ \gamma \| Z - \sigma(ZSA) \|_F^2 + \lambda \| L \|_F^2.
\]

Reconstruction regularization. The goal of the autoencoder-like term is to reconstruct feature Z from model output ZS

Any generic loss
Since the framework considers asymmetric transfer in the feature space, it can be generalized to deep network with multiple layers.

- autoencoding regularization term $Z$ is formulated at the second-last layer.

$$
\min_{A, \{W^{(l)}\}_{l=1}^L} \sum_{t=1}^T \left\{ (1 + \alpha \| a_t^o \|_1) L_t + \mu \| w_t^{(L)} \|_1 \right\} \\
+ \gamma \left\| \sigma(Z W^{(L)} A) - Z \right\|_F^2 + \lambda \sum_{l=1}^{L-1} \left\| W^{(l)} \right\|_F^2,
$$

Where

$$
Z = \sigma(W^{(L-1)}) \sigma(W^{(L-2)}) \ldots \sigma(XW^{(1)}))
$$
Data Summary

- **For shallow models:**
  - AWA-A
  - MNIST
  - School
  - Room

- **For deep models:**
  - MNIST-Imbalanced
  - CUB-200
  - AWA-C
  - ImageNet-Small
Experimental Results

- For shallow models:

*Table 1.* Performance of the linear and shallow baselines and our asymmetric multi-task feature learning model. We report the RMSE for regression and mean classification error(%) for classification, along with the standard error for 95% confidence interval.

<table>
<thead>
<tr>
<th></th>
<th>AWA-A</th>
<th>MNIST</th>
<th>School</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>STL</td>
<td>37.6±0.5</td>
<td>14.8±0.6</td>
<td>10.16±0.08</td>
<td>45.9±1.4</td>
</tr>
<tr>
<td>GO-MTL</td>
<td>35.6±0.2</td>
<td>14.4±1.3</td>
<td>9.87±0.06</td>
<td>47.1±1.4</td>
</tr>
<tr>
<td>AMTL</td>
<td>33.4±0.3</td>
<td>12.9±1.4</td>
<td>10.13±0.08</td>
<td>40.8±1.5</td>
</tr>
<tr>
<td>NN</td>
<td>26.3±0.3</td>
<td>8.96±0.9</td>
<td>9.89±0.03</td>
<td>44.5±2.0</td>
</tr>
<tr>
<td>MT-NN</td>
<td>26.2±0.3</td>
<td>8.76±1.0</td>
<td>9.91±0.04</td>
<td>41.7±1.7</td>
</tr>
<tr>
<td>AMTFL</td>
<td><strong>25.2±0.3</strong></td>
<td><strong>8.68±0.9</strong></td>
<td>9.89±0.09</td>
<td><strong>40.4±2.4</strong></td>
</tr>
</tbody>
</table>
Experimental Results

- For deep models:

Table 2. Classification performance of the deep learning baselines and Deep-AMTFL. The reported numbers for MNIST-Imbalanced and CUB datasets are averages over 5 runs.

<table>
<thead>
<tr>
<th>Model</th>
<th>MNIST-Imbal.</th>
<th>CUB</th>
<th>AWA-C</th>
<th>Small</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>8.13</td>
<td>46.18</td>
<td>33.36</td>
<td>66.54</td>
</tr>
<tr>
<td>MT-CNN</td>
<td>8.72</td>
<td>43.92</td>
<td>32.80</td>
<td>65.69</td>
</tr>
<tr>
<td>Deep-AMTL</td>
<td>8.52</td>
<td>45.26</td>
<td>32.32</td>
<td>65.61</td>
</tr>
<tr>
<td>Deep-AMTFL</td>
<td><strong>5.82</strong></td>
<td><strong>43.75</strong></td>
<td><strong>31.88</strong></td>
<td><strong>64.49</strong></td>
</tr>
</tbody>
</table>
Experimental Analysis

● For shallow models:
  ○ AMTFL outperforms the baselines on most datasets.
  ○ The only exception is the School dataset, on which GO-MTL obtains the best performance, but is due to the strong homogeneity among the tasks in this particular dataset.

● For deep models:
  ○ Deep-AMTFL outperforms all baselines, including MT-CNN and Deep-AMTL.
  ○ It shows the effectiveness of our asymmetric knowledge transfer from tasks to features, and back to tasks in deep learning frameworks.
Reproduction

- In our implementation, we tried to reproduce the results for the MNIST dataset.
- We use the CNN (Lenet-Conv) mentioned in the paper.
- Since the paper does not include all the hyperparameters, we cannot exactly reproduce the numbers, but the gap is trivial (~1%).
- Following is the comparison between with AMTFL and without it.

<table>
<thead>
<tr>
<th>Model</th>
<th>MT-CNN</th>
<th>Deep-AMTFL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.9026</td>
<td>0.9301</td>
</tr>
</tbody>
</table>

Test

Test and compare the accuracies

```
[9]  # test normal
    print('Normal model\'s accuracy:', test_acc)

    # test AMTFL
    print('AMTFL model\'s accuracy:', test_acc)
```

Normal model's accuracy: 0.9026
AMTFL model's accuracy: 0.9301
Conclusion and Future Work

- Propose a novel deep asymmetric multi-task feature learning framework, effectively prevent negative transfer resulting from symmetric influences of each task in feature learning.

- The predictors can asymmetrically affect the learning of shared representations by introducing an asymmetric feedback connections.

- Experimental results show that our model significantly outperforms asymmetric multi-task learning for both shallow and deep frameworks.
## Division of Work

<table>
<thead>
<tr>
<th></th>
<th>Slide</th>
<th>Coding</th>
<th>Presentation</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Qi Liu</strong></td>
<td>motivation, background, related work</td>
<td>data preprocessing, PCA</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td><strong>David Yao</strong></td>
<td>target task, intuitive figure of why claim</td>
<td>AWA dataset, data preprocessing</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td><strong>Aobo Yang</strong></td>
<td>Solution, Implementation</td>
<td>Cross entropy loss, AMTFL regularization, Test and analysis</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td><strong>Yujia Mu</strong></td>
<td>Data Summary; Experimental results; Experimental analysis; Conclusion and future work.</td>
<td>MNIST-imbalanced; CNN Lenet-Conv</td>
<td>√</td>
<td></td>
</tr>
</tbody>
</table>