Learning New Tricks From Old Dogs: Multi-Source Transfer Learning From Pre-Trained Networks

Joshua Lea, Prasanna Sattigeri, and Gregory Wornell
1 Dept. EECS, MIT, 2 MIT-IBM Watson AI Lab, IBM Research

Problem Setup
Consider an ensemble of features $f_1(x), \ldots, f_k(x)$ extracted from networks pre-trained on unknown objectives, and a target classification task with data $X \in X$ and labels $Y \in Y$. We wish to train a classifier on very few samples using the pre-trained features without altering the existing networks (black-box feature access).

Examples:
- Distributed transfer learning (e.g. learning from multiple mobile devices, each with their own network)
- Rapid adaptation to new environments with multiple candidate source models to transfer from
- Learning from old networks for which the original training data is lost

Table: Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>1-Shot Acc</th>
<th>5-Shot Acc</th>
<th>10-Shot Acc</th>
<th>20-Shot Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Single Source SVM</td>
<td>98.9 ± 0.3</td>
<td>96.9 ± 0.5</td>
<td>97.7 ± 1.0</td>
<td>79.0 ± 1.2</td>
</tr>
<tr>
<td>Best Single Source MCW</td>
<td>96.2 ± 1.1</td>
<td>69.0 ± 0.2</td>
<td>97.4 ± 2.4</td>
<td>76.4 ± 1.2</td>
</tr>
<tr>
<td>Multi-Source SVM</td>
<td>64.7 ± 3.0</td>
<td>72.8 ± 2.7</td>
<td>76.4 ± 2.7</td>
<td>81.0 ± 0.6</td>
</tr>
<tr>
<td>Multi-Source MCW</td>
<td>69.9 ± 3.0</td>
<td>78.1 ± 0.8</td>
<td>80.1 ± 0.8</td>
<td>81.7 ± 0.2</td>
</tr>
</tbody>
</table>

Experimental Results - Source Selection

Average values of $\sum \sigma_i$ for each source task for the 5-shot transfer learning task on the CIFAR-100 dataset, with the target task of “apple vs. fish.”

References

Experimental Setup

- We ran experiments on three image datasets: (a) CIFAR-100, (b) Stanford Dogs, and (c) Tiny ImageNet.
- For each dataset, we divide the images into a set of smaller, mutually-exclusive classification tasks.
- We select one task as the target and the remainder as the sources.
- For each source task, we train a neural net with the LeNet architecture [3] for classification.
- We use the penultimate layer of these nets as feature functions, and compute MCW parameters with respect to the target task.
- We compare the classification accuracy with that of an SVM trained on the same features.


- Given multiple pre-trained feature functions $f_1, \ldots, f_k$, the Maximal Correlation Objective is given by:

$$ L = \mathbb{E}_{P_{X,Y}} \left[ (F^T(X)G(Y)) \right] $$

(1)

This objective separates out as:

$$ L = \sum_{i=1}^k \mathbb{E}_{P_{X,Y}} \left[ f_i(x)g_i(y) \right] $$

(2)

- We can solve each term separately to find the associated $g_1, \ldots, g_k$ and $\sigma_1, \ldots, \sigma_k$ by taking conditional expectations over the empirical distribution of target samples:

$$ g_i(y) = \mathbb{E}_{P_{X,Y}} \left[ f_i(x) \mid y \right] $$

$$ \sigma_i = \mathbb{E}_{P_{X,Y}} \left[ f_i(x)g_i(y) \right] $$

- We then construct the approximate distribution:

$$ P_{X,Y}^\ast(y|x) = P_{Y}^\ast(y) \left[ 1 + \sum_{i=1}^k \sigma_i f_i(x)g_i(y) \right] $$

And apply an ML estimator to predict $y$ given $x$. 

Maximal Correlation Functions

- Hirschfeld-Gebelein-Rényi Maximal Correlation:

$$ \sigma = \sup_{f_1,\ldots,f_k} \mathbb{E}[f(X)y(Y)] $$

(3)

The optimal $f$ and $g$ are maximal correlation functions, and have been shown to be universally optimal in an information-preserving sense.

- For a fixed $f_i$, the optimal $g$ is given by: [1]

$$ g_i(y) = \mathbb{E}_{P_{X,Y}} \left[ f_i(x) \mid y \right] $$

- We ran experiments on three image datasets: (a) CIFAR-100, (b) Stanford Dogs, and (c) Tiny ImageNet.

- For each dataset, we divide the images into a set of smaller, mutually-exclusive classification tasks.

- We select one task as the target and the remainder as the sources.

- For each source task, we train a neural net with the LeNet architecture [3] for classification.

- We use the penultimate layer of these nets as feature functions, and compute MCW parameters with respect to the target task.

- We compare the classification accuracy with that of an SVM trained on the same features.

- For each source task for the 5-shot transfer learning task on the CIFAR-100 dataset, with the target task of “apple vs. fish.”

- Experimental results on target task for the CIFAR-100 dataset using the MCW method with different subsets of source networks.

- Experimental results on target task for the Stanford Dogs dataset (10 source tasks, 5-way classification).

- Experimental results on target task for the Tiny ImageNet dataset (10 source tasks, 5-way classification).