Transfer Learning Algorithms for Image Classification

Ariadna Quattoni
MIT, CSAIL

Advisors:
Michael Collins
Trevor Darrell
Motivation

Goal:

- We want to be able to build classifiers for thousands of visual categories.
- We want to exploit rich and complex feature representations.

Problem:

- We might only have a few labeled samples per category.

Solution:

- Transfer Learning, leverage labeled data from multiple related tasks.
We study efficient transfer algorithms for image classification which can exploit supervised training data from a set of related tasks:

- Learn an image representation using supervised data from auxiliary tasks automatically derived from unlabeled images + meta-data.

- A transfer learning model based on joint regularization and an efficient optimization algorithm for training jointly sparse classifiers in high dimensional feature spaces.
A method for learning image representations from unlabeled images + meta-data

Create auxiliary problems

Structure Learning

[Ando & Zhang, JMLR 2005]

Visual Representation

\( F : I \rightarrow R^h \)

Structure Learning:

\[ f_k(x) = (\nu_k^t \theta^t x) \]

Task specific parameters

Shared Parameters
A method for learning image representations from unlabeled images + meta-data
Outline

- An overview of transfer learning methods.
- A joint sparse approximation model for transfer learning.
- Asymmetric transfer experiments.
- An efficient training algorithm.
- Symmetric transfer image annotation experiments.
Transfer Learning: A brief overview

The goal of transfer learning is to use labeled data from related tasks to make learning easier. Two settings:

- **Asymmetric transfer:**
  Resource: Large amounts of supervised data for a set of related tasks.
  Goal: Improve performance on a target task for which training data is scarce.

- **Symmetric transfer:**
  Resource: Small amount of training data for a large number of related tasks.
  Goal: Improve average performance over all classifiers.
Transfer Learning: A brief overview

- Three main approaches:
  - Learning priors over parameters: [Raina 2006, Lawrence et al. 2004]
  - Learning relevant shared features [Torralba 2004, Obozinsky 2006]
Feature Sharing Framework:

- Work with a rich representation:
  - Complex features, high dimensional space
  - Some of them will be very discriminative (hopefully)
  - Most will be irrelevant

- Related problems may share relevant features.

- If we knew the relevant features we could:
  - Learn from fewer examples
  - Build more efficient classifiers

- We can train classifiers from related problems together using a regularization penalty designed to promote joint sparsity.
<table>
<thead>
<tr>
<th>Church</th>
<th>Airport</th>
<th>Grocery Store</th>
<th>Flower-Shop</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Church" /></td>
<td><img src="image2" alt="Airport" /></td>
<td><img src="image3" alt="Grocery Store" /></td>
<td><img src="image4" alt="Flower-Shop" /></td>
</tr>
<tr>
<td>$w_{1,1}$</td>
<td>$w_{1,2}$</td>
<td>$w_{1,3}$</td>
<td>$w_{1,4}$</td>
</tr>
<tr>
<td>$w_{2,1}$</td>
<td>$w_{2,2}$</td>
<td>$w_{2,3}$</td>
<td>$w_{2,4}$</td>
</tr>
<tr>
<td>$w_{3,1}$</td>
<td>$w_{3,2}$</td>
<td>$w_{3,3}$</td>
<td>$w_{3,4}$</td>
</tr>
<tr>
<td>$w_{4,1}$</td>
<td>$w_{4,2}$</td>
<td>$w_{4,3}$</td>
<td>$w_{4,4}$</td>
</tr>
<tr>
<td>$w_{5,1}$</td>
<td>$w_{5,2}$</td>
<td>$w_{5,3}$</td>
<td>$w_{5,4}$</td>
</tr>
<tr>
<td>Church</td>
<td>Airport</td>
<td>Grocery Store</td>
<td>Flower-Shop</td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>---------------</td>
<td>-------------</td>
</tr>
<tr>
<td><img src="Church.png" alt="Image" /></td>
<td><img src="Airport.png" alt="Image" /></td>
<td><img src="Grocery.png" alt="Image" /></td>
<td><img src="Flower.png" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$w_{1,1}$</th>
<th>$w_{1,2}$</th>
<th>$w_{1,3}$</th>
<th>$w_{1,4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="W1_1.png" alt="Image" /></td>
<td><img src="W1_2.png" alt="Image" /></td>
<td><img src="W1_3.png" alt="Image" /></td>
<td><img src="W1_4.png" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$w_{2,1}$</th>
<th>$w_{2,2}$</th>
<th>$w_{2,3}$</th>
<th>$w_{2,4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="W2_1.png" alt="Image" /></td>
<td><img src="W2_2.png" alt="Image" /></td>
<td><img src="W2_3.png" alt="Image" /></td>
<td><img src="W2_4.png" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$w_{3,1}$</th>
<th>$w_{3,2}$</th>
<th>$w_{3,3}$</th>
<th>$w_{3,4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="W3_1.png" alt="Image" /></td>
<td><img src="W3_2.png" alt="Image" /></td>
<td><img src="W3_3.png" alt="Image" /></td>
<td><img src="W3_4.png" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$w_{4,1}$</th>
<th>$w_{4,2}$</th>
<th>$w_{4,3}$</th>
<th>$w_{4,4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="W4_1.png" alt="Image" /></td>
<td><img src="W4_2.png" alt="Image" /></td>
<td><img src="W4_3.png" alt="Image" /></td>
<td><img src="W4_4.png" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$w_{5,1}$</th>
<th>$w_{5,2}$</th>
<th>$w_{5,3}$</th>
<th>$w_{5,4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="W5_1.png" alt="Image" /></td>
<td><img src="W5_2.png" alt="Image" /></td>
<td><img src="W5_3.png" alt="Image" /></td>
<td><img src="W5_4.png" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td>Church</td>
<td>Airport</td>
<td>Grocery Store</td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
<td>---------</td>
<td>---------------</td>
</tr>
<tr>
<td></td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
</tbody>
</table>

\[ w_{2,1} + w_{2,2} - w_{2,3} - w_{2,4} \]

\[ w_{3,2} - w_{3,2} + w_{4,3} - w_{4,4} \]

\[ w_{4,1} + w_{4,2} - w_{4,3} + w_{4,4} \]

\[ + - w_{5,3} + \]
Torralba et al. [2004] developed a joint boosting algorithm based on the idea of learning additive models for each class that share weak learners.

Obozinski et al. [2006] proposed $L_{1,2}$ joint penalty and developed a blockwise boosting scheme based on Boosted-Lasso.
Our Contribution

A new model and optimization algorithm for training jointly sparse classifiers:

- Previous approaches to joint sparse approximation have relied on greedy coordinate descent methods.

- We propose a simple an efficient global optimization algorithm with guaranteed convergence rates.

- Our algorithm can scale to large problems involving hundreds of problems and thousands of examples and features.

- We test our model on real image classification tasks where we observe improvements in both asymmetric and symmetric transfer settings.
Outline

- An overview of transfer learning methods.
- A joint sparse approximation model for transfer learning.
- Asymmetric transfer experiments.
- An efficient training algorithm.
- Symmetric transfer image annotation experiments.
**Notation**

Collection of Tasks

\[ \mathbf{D} = \{D_1, D_2, \ldots, D_m\} \]

\[ D_k = \{(x_1^k, y_1^k), \ldots, (x_{n_k}^k, y_{n_k}^k)\} \]

\[ \mathbf{x} \in \mathbb{R}^d \quad y \in \{+1, -1\} \]

**Joint Sparse Approximation**

\[
\begin{bmatrix}
W_{1,1} & W_{1,2} & \cdots & W_{1,m} \\
W_{2,1} & W_{2,2} & \cdots & W_{2,m} \\
& & \ddots & \\
& & & \ddots \\
W_{d,1} & W_{d,2} & \cdots & W_{d,m}
\end{bmatrix}
\]
Consider learning a single sparse linear classifier of the form:

$$f(x) = w \cdot x$$

We want a few features with non-zero coefficients.

Recent work suggests to use $L_1$ regularization:

$$\arg\min_w \sum_{(x,y) \in D} l(f(x), y) + Q \sum_{j=1}^{d} |w_j|$$

- Classification error
- $L_1$ penalizes non-sparse solutions

Donoho [2004] proved (in a regression setting) that the solution with smallest $L_1$ norm is also the sparsest solution.
Joint Sparse Approximation

Setting: Joint Sparse Approximation

\[ f_k(x) = w_k \cdot x \]

\[ \arg \min_{w_1, w_2, \ldots, w_m} \sum_{k=1}^{m} \frac{1}{|D_k|} \sum_{(x, y) \in D_k} l(f_k(x), y) + Q R(w_1, w_2, \ldots, w_m) \]

- Average Loss on Collection D
- Penalizes solutions that utilize too many features
Joint Regularization Penalty

- How do we penalize solutions that use too many features?

\[
\begin{bmatrix}
W_{1,1} & W_{1,2} & \cdots & W_{1,m} \\
W_{2,1} & W_{2,2} & \cdots & W_{2,m} \\
\vdots & \vdots & \ddots & \vdots \\
W_{d,1} & W_{d,2} & \cdots & W_{d,m}
\end{bmatrix}
\]

Coefficients for classifier 2

\[R(W) = \# \text{non-zero rows}\]

- Would lead to a hard combinatorial problem.
Joint Regularization Penalty

- We will use a $L_{1-\infty}$ norm [Tropp 2006]

$$R(W) = \sum_{i=1}^{d} \max_{k}(|W_{ik}|)$$

- This norm combines:

  - The $L_{\infty}$ norm on each row promotes non-sparsity on each row.

    → Share features

  - An $L_{1}$ norm on the maximum absolute values of the coefficients across tasks promotes sparsity.

    → Use few features

- The combination of the two norms results in a solution where only a few features are used but the features used will contribute in solving many classification problems.
Joint Sparse Approximation

- Using the $L_{1,\infty}$ norm we can rewrite our objective function as:

$$
\min_{w} \sum_{k=1}^{m} \frac{1}{|D_k|} \sum_{(x,y) \in D_k} l(f_k(x), y) + Q \sum_{i=1}^{d} \max(|W_{ik}|)
$$

- For any convex loss this is a convex objective.

- For the hinge loss: $l(f(x), y) = \max(0, 1 - yf(x))$
  the optimization problem can be expressed as a linear program.
Joint Sparse Approximation

- Linear program formulation (hinge loss):

- Objective:

  $$\min_{[w, \varepsilon, t]} \sum_{k=1}^{m} \frac{1}{|D_k|} \sum_{j=1}^{|D_k|} \varepsilon_j^k + Q \sum_{i=1}^{d} t_i$$

- Max value constraints:

  for \(k = 1:m\) and for \(i = 1:d\)

  $$-t_i \leq w_{ik} \leq t_i$$

- Slack variables constraints:

  for \(k = 1:m\) and for \(j = 1:|D_k|\)

  $$y_j^k f_k(x_j^k) \geq 1 - \varepsilon_j^k$$

  $$\varepsilon_j^k \geq 0$$
Outline

- An overview of transfer learning methods.
- A joint sparse approximation model for transfer learning.
- **Asymmetric transfer experiments.**
- An efficient training algorithm.
- Symmetric transfer image annotation experiments.
Setting: Asymmetric Transfer

- SuperBowl
- Sharon
- Danish Cartoons
- Australian Open
- Trapped Miners
- Golden globes
- Grammys
- Figure Skating
- Iraq
- Academy Awards

- Train a classifier for the 10th held out topic using the relevant features $\mathbf{R}$ only.

- Learn a representation using labeled data from 9 topics.

- Learn the matrix $\mathbf{W}$ using our transfer algorithm.

- Define the set of relevant features to be: $R = \{r : \max_k (|w_{rk}|) > 0\}$
Results

Asymmetric Transfer
Outline

- An overview of transfer learning methods.
- A joint sparse approximation model for transfer learning.
- Asymmetric transfer experiments.
- An efficient training algorithm.
- Symmetric transfer image annotation experiments.
Limitations of the LP formulation

- The LP formulation can be optimized using standard LP solvers.

- The LP formulation is feasible for small problems but becomes intractable for larger data-sets with thousands of examples and dimensions.

- We might want a more general optimization algorithm that can handle arbitrary convex losses.
**L_{1-\infty}** Regularization: Constrained Convex Optimization Formulation

\[
\arg\min_w \sum_{k=1}^{m} \frac{1}{|D_k|} \sum_{(x,y) \in D_k} l(f_k(x), y) \quad \text{A convex function}
\]

\[
s.t. \sum_{i=1}^{d} \max_k(|W_{ik}|) \leq C \quad \text{Convex constraints}
\]

- We will use a Projected SubGradient method.
  - Main advantages: simple, scalable, guaranteed convergence rates.

- Projected SubGradient methods have been recently proposed:
  - L_2 regularization, i.e. SVM [Shalev-Shwartz et al. 2007]
  - L_1 regularization [Duchi et al. 2008]
Euclidean Projection into the $L_{1,\infty}$ ball

$$\mathbf{P}_{1,\infty} : \min_{B,\mu} \quad \frac{1}{2} \sum_{i,j} (B_{i,j} - A_{i,j})^2$$

subject to:

$$\forall i, j \quad B_{i,j} \leq \mu_i$$

$$\sum_i \mu_i = C$$

$$\forall i, j \quad B_{i,j} \geq 0$$

$$\forall i \quad \mu_i \geq 0$$
Characterization of the solution

Let $\mu$ be the optimal maximums of problem $P_{1,\infty}$. The optimal matrix $B$ of $P_{1,\infty}$ satisfies that:

\[
A_{i,j} \geq \mu_i \implies B_{i,j} = \mu_i \\
A_{i,j} \leq \mu_i \implies B_{i,j} = A_{i,j} \\
\mu_i = 0 \implies B_{i,j} = 0
\]
Characterization of the solution

At the optimal solution of \( P_{1,\infty} \) there exists a constant \( \theta \geq 0 \) such that for every \( i \) either:

- \( \mu_i > 0 \) and \( \sum_j (A_{i,j} - B_{i,j}) = \theta \)
- \( \mu_i = 0 \) and \( \sum_j A_{i,j} \leq \theta \)
Mapping to a simpler problem

- We can map the projection problem to the following problem which finds the optimal maximums $\mu$:

$$
M_{1,\infty} : \text{find } \mu, \theta \\
\text{s.t. } \sum_{i} \mu_i = C \\
\sum_{j : A_{i,j} \geq \mu_i} (A_{i,j} - \mu_i) = \theta , \forall i \text{ s.t. } \mu_i > 0 \\
\sum_{j} A_{i,j} \leq \theta , \forall i \text{ s.t. } \mu_i = 0 \\
\forall i \text{ } \mu_i \geq 0 \ ; \ \theta \geq 0
$$

For any matrix $A$ and a constant $C$ such that $C < ||A||_{1,\infty}$, there is a unique solution $\mu^*, \theta^*$ to the problem $M_{1,\infty}$. 
Efficient Algorithm for: $M_{1,\infty}$, in pictures

4 Features, 6 problems, $C=14 \sum_{i=1}^{d} \max(|A_{ik}|) = 29$
Complexity

- The total cost of the algorithm is dominated by a sort of the entries of \( A \).

- The total cost is in the order of: \( O(dm \log(dm)) \)
Outline

- An overview of transfer learning methods.
- A joint sparse approximation model for transfer learning.
- Asymmetric transfer experiments.
- An efficient training algorithm.
- Symmetric transfer image annotation experiments.
Synthetic Experiments

- Generate a jointly sparse parameter matrix $W$:

- For every task we generate pairs: $(x_i^k, y_i^k)$
  
  where $y_i^k = \text{sign}(w_k^t x_i^k)$

- We compared three different types of regularization (i.e. projections):
  
  - $L_{1-\infty}$ projection
  - L2 projection
  - L1 projection
Synthetic Experiments

Test Error

Performance on predicting relevant features
Dataset: Image Annotation

- 40 top content words
- Raw image representation: Vocabulary Tree (Grauman and Darrell 2005, Nister and Stewenius 2006)
- 11000 dimensions
The differences are statistically significant
Results

Joint-Sparsity Accuracy Tradeoff L1

Average AUC

% non-zero features

Joint-Sparsity Accuracy Tradeoff L1\inf

Average AUC

% non-zero features
Dataset: Indoor Scene Recognition

- 67 indoor scenes.
- Raw image representation: similarities to a set of unlabeled images.
- 2000 dimensions.
Results:
Summary of Thesis Contributions

- A method that learns image representations using unlabeled images + meta-data.

- A transfer model based on performing a joint loss minimization over the training sets of related tasks with a shared regularization.

- Previous approaches used greedy coordinate descent methods. We propose an efficient global optimization algorithm for learning jointly sparse models.

- A tool that makes implementing an $L_{1-\infty}$ penalty as easy and almost as efficient as implementing the standard L1 and L2 penalties.

- We presented experiments on real image classification tasks for both an asymmetric and symmetric transfer setting.
Future Work

- Online Optimization.
- Task Clustering.
- Combining feature representations.
- Generalization properties of $L_{1-\infty}$ regularized models.
Thanks!