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.

Thesis Contributions

- 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 feature sharing transfer algorithm based on joint regularization.
- An efficient algorithm for training jointly sparse classifiers in high dimensional feature spaces.

Outline

- **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 intermediate latent representations: [Thrun 1996, Baxter 1997, Caruana 1997, Argyriou 2006, Amit 2007]
  - Learning priors over parameters: [Raina 2006, Lawrence et al. 2004]
  - Learning relevant shared features via joint sparse regularization: [Torralba 2004, Obozinsky 2006]

Feature Sharing Framework:

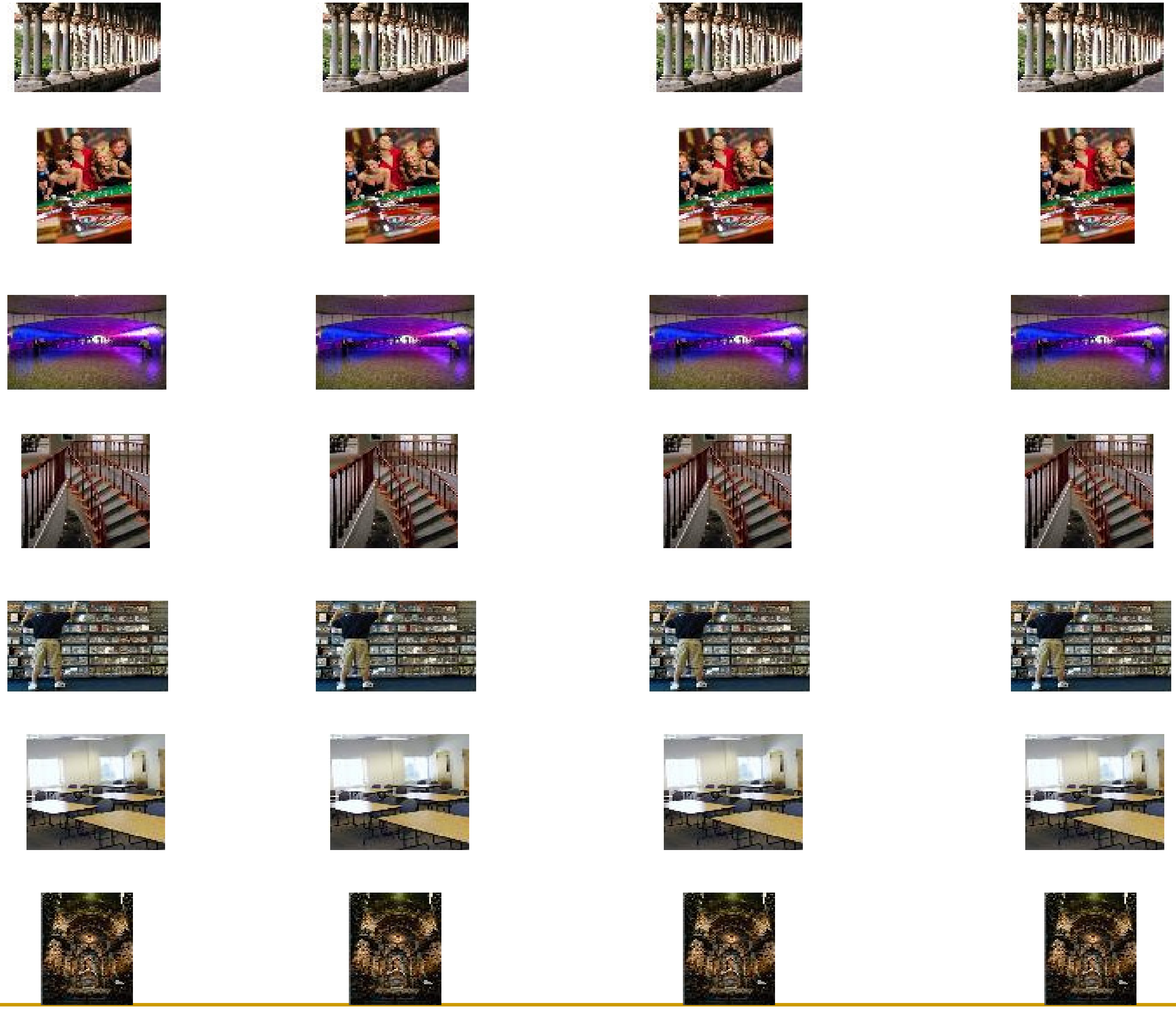
- Work with a rich representation:
  - Complex features, high dimensional space
  - Some of them will be very discriminative (hopefully)
  - Most will be irrelevant

- Related problems may share relevant features.

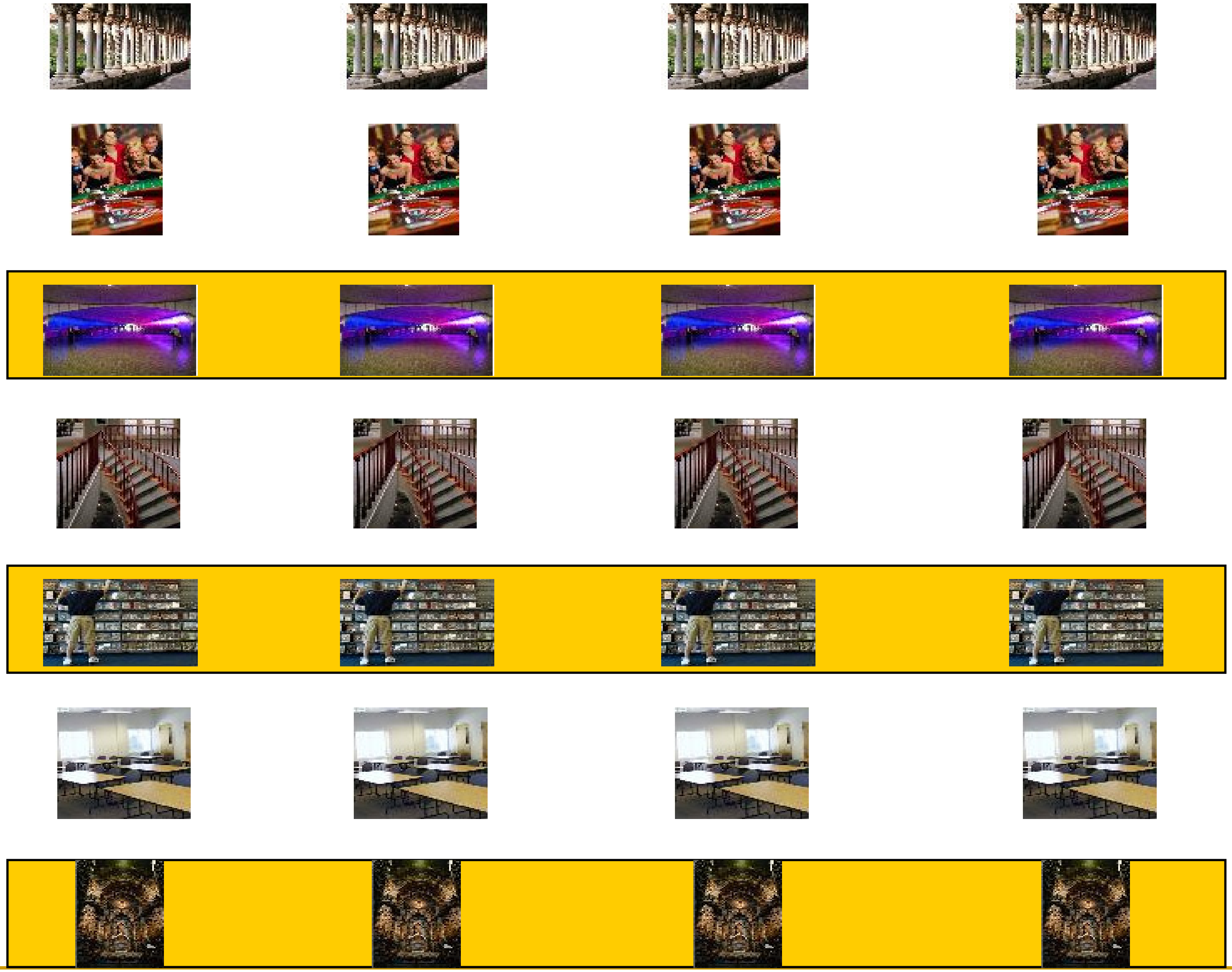
- If we knew the relevant features we could:
  - Learn from fewer examples
  - Build more efficient classifiers

- We can train classifiers from related problems together using a regularization penalty designed to promote joint sparsity.

Church	Airport	Grocery Store	Flower-Shop



Church | Airport | Grocery Store | Flower-Shop



Related Formulations of Joint Sparse Approximation

- 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 $L_{1\text{-}2}$ joint penalty and developed a blockwise boosting scheme based on Boosted-Lasso.

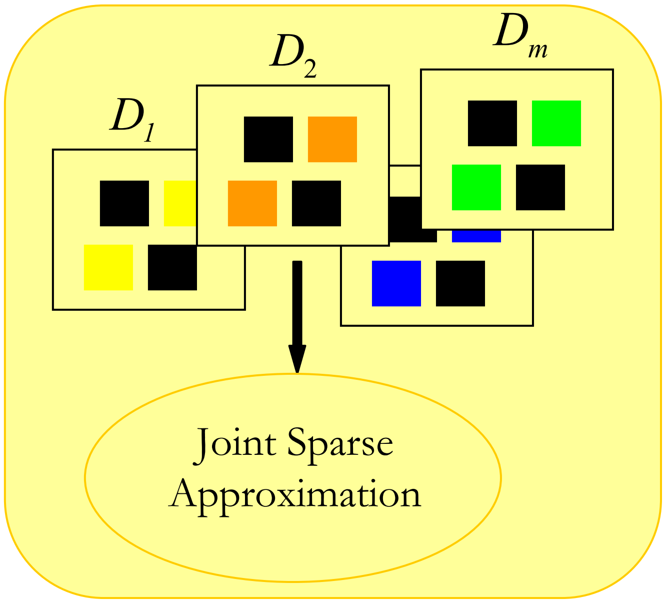
.

Our Contribution

A new model and optimization algorithm for training jointly sparse classifiers in high dimensional feature spaces.

- Previous approaches to joint sparse approximation (Torralba et al., 2004, Obozinski et al.,2006;) have relied on greedy coordinate descent methods.
- We propose a simple an efficient global optimization algorithm with guaranteed convergence rates $O\left(\frac{1}{\varepsilon^2}\right)$
- 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.
- We show that our algorithm can successfully recover jointly sparse solutions.

Notation



Collection of Tasks

$$\mathbf{D} = \{D_1, D_2, \ldots, D_m\}$$

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

$$\mathbf{x} \in \Re^d \quad y \in \{+1, -1\}$$

$$W \Rightarrow \begin{matrix} w_{1,1} & w_{1,2} & \cdots & w_{1,m} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,m} \\ \vdots & \ddots & \ddots & \vdots \\ w_{d,1} & w_{d,2} & \cdots & w_{d,m} \end{matrix}$$

Single Task Sparse Approximation

- 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:

$$\arg\min_{\mathbf{w}} \underbrace{\sum_{(x,y)\in D} l(f(x),y)}_{\text{Classification error}} + Q\underbrace{\sum_{j=1}^{d}}_{L_1 \text{ penalizes non-sparse solutions}} |w_j|$$

- 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

$$f_k(x) = \mathbf{w}_k \cdot x$$

$$\arg\min_{\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_m} \underbrace{\sum_{k=1}^{m} \frac{1}{|D_k|} \sum_{(x,y)\in D_k} l\,(f_k(x), y)}_{\text{Average Loss on training set k}} + \underbrace{Q\,\mathrm{R}(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m)}_{\text{penalizes solutions that utilize too many features}}$$

Joint Regularization Penalty

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

$$\begin{matrix} W_{1,1} & \boxed{W_{1,2}} & \cdots & W_{1,m} \\ W_{2,1} & W_{2,2} & \cdots & W_{2,m} \\ \vdots & \ddots & \ddots & \vdots \\ W_{d,1} & W_{d,2} & \cdots & W_{d,m} \end{matrix}$$

Coefficients for for feature 2

Coefficients for classifier 2

$$\mathrm{R}(W) = \# non - zero - rows$$

- 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_∞ norm on each row promotes non-sparsity on the rows.	$\longrightarrow$	Share features
An L_1 norm on the maximum absolute values of the coefficients across tasks promotes sparsity.	$\longrightarrow$	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 $L_{1\text{-}\infty}$ norm we can rewrite our objective function as:

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

- 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:

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

- Max value constraints:

$for: k = 1:m$ and $for: i = 1:d$

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

- Slack variables constraints:

$for: k = 1:m$ and $for: j = 1:|D_k|$

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

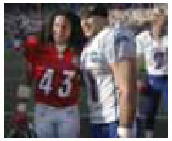
$$\varepsilon_j^k \ge 0$$

Outline

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

Setting: Asymmetric Transfer

SuperBowl

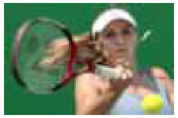


Sharon

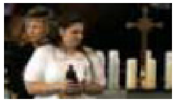
Danish Cartoons

Academy Awards

Australian Open



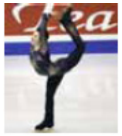
Trapped Miners



Golden globes

Grammys

Figure Skating

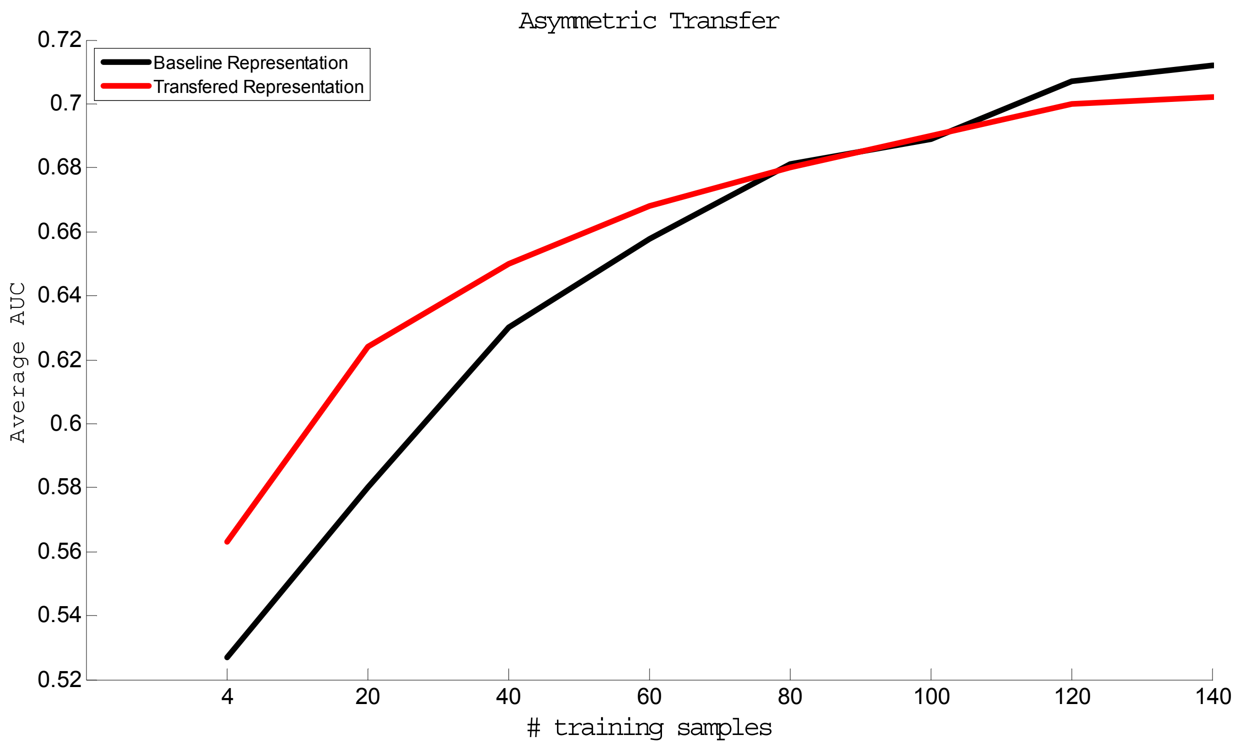


Iraq

- Train a classifier for the 10th held out topic using the relevant features **R** only.

- Learn a representation using labeled data from 9 topics.
- Learn the matrix **W** using our transfer algorithm.
- Define the set of relevant features to be: $R = \{r : \max_k (| w_{rk} |) > 0\}$

Results



Asymmetric Transfer
Baseline Representation
Transfered Representation
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.58
0.56
0.54
0.52
Average AUC
4
20
40
60
80
100
120
140
training samples

An efficient training algorithm

- 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.

Outline

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

$L_{1\text{-}\infty}$ Regularization: Constrained Convex Optimization Formulation

$$\arg\min_{\mathbf{W}} \sum_{k=1}^{m} \frac{1}{|D_k|} \sum_{(x,y)\in D_k} l\,(f_k(x), y)$$ A convex function

$$s.t. \sum_{i=1}^{d} \max_k(|W_{ik}|) \le C$$ 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

Snapshot of the idea:

- We map the projection to a simpler problem which involves finding new maximums for each feature across tasks and using them to truncate the original matrix.
- The total mass removed from a feature across tasks should be the same for all features whose coefficients don't become zero.

Euclidean Projection into the $L_{1\text{-}\infty}$ ball

$$
\begin{aligned}
\mathbf{P}_{\mathbf{1},\infty}: \quad \min_{B,\boldsymbol{\mu}} \quad & \frac{1}{2}\sum_{i,j}(B_{i,j} - A_{i,j})^2 \\
\text{s.t.} \quad & \forall i,j \;\; B_{i,j} \le \mu_i \\
& \sum_i \mu_i = C \\
& \forall i,j \;\; B_{i,j} \ge 0 \\
& \forall i \;\; \mu_i \ge 0
\end{aligned}
$$

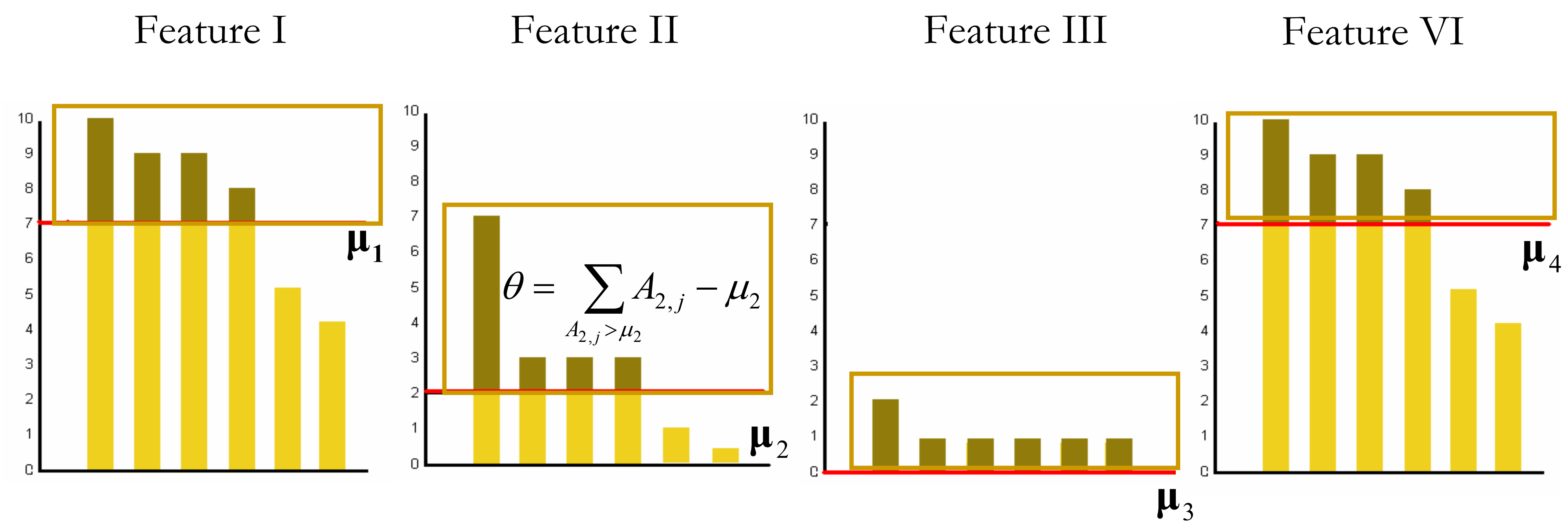
Characterization of the solution

Lemma 1 *Let* μ *be the optimal maximums of problem* $\mathrm{P}_{1,\infty}$. *The optimal matrix* B *of* $\mathrm{P}_{1,\infty}$ *satisfies that:*

$$
\begin{aligned}
A_{i,j} \geq \mu_i &\implies B_{i,j} = \mu_i \\
A_{i,j} \leq \mu_i &\implies B_{i,j} = A_{i,j} \\
\mu_i = 0 &\implies B_{i,j} = 0
\end{aligned}
$$

Characterization of the solution

Lemma 2 *At the optimal solution of* $\mathrm{P}_{1,\infty}$ *there exists a constant* $\theta \geq 0$ *such that for every* i*: either (a)* $\mu_i > 0$ *and* $\sum_j (A_{i,j} - B_{i,j}) = \theta$*; or (b)* $\mu_i = 0$ *and* $\sum_j A_{i,j} \leq \theta$*.*



Mapping to a simpler problem

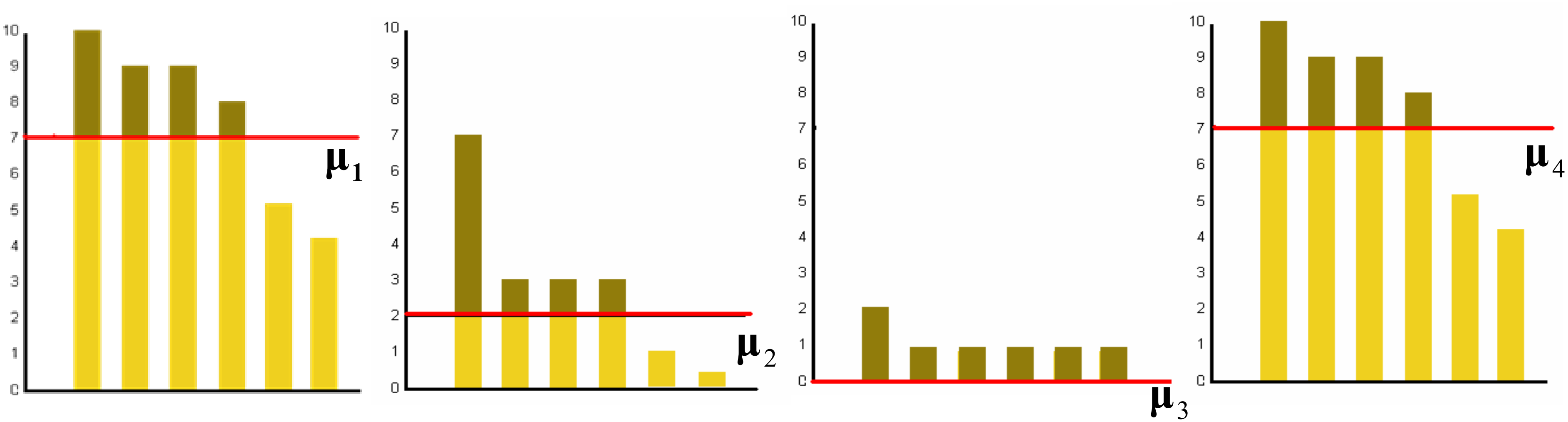
- We can map the projection problem to the following problem which finds the optimal maximums **μ**:

$$
\begin{aligned}
\mathrm{M}_{\mathbf{1},\infty}: \quad & \text{find} \quad \boldsymbol{\mu}\ ,\ \theta \\
& \text{s.t.} \quad \sum_i \mu_i = C \\
& \sum_{j:A_{i,j} \geq \mu_i} (A_{i,j} - \mu_i) = \theta\ ,\ \forall i \text{ s.t. } \mu_i > 0 \\
& \sum_j A_{i,j} \leq \theta\ ,\ \forall i \ \text{ s.t. } \ \mu_i = 0 \\
& \forall i \ \ \mu_i \geq 0 \ ; \ \ \theta \geq 0
\end{aligned}
$$

Lemma 3 *For a matrix A and a constant $C < ||A||_{1,\infty}$, there is a unique solution $\boldsymbol{\mu}^*, \theta^*$ to the problem* $\mathrm{M}_{1,\infty}$.

Efficient Algorithm for: $M_{1,\infty}$, in pictures

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



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: $O(dm\log(dm))$

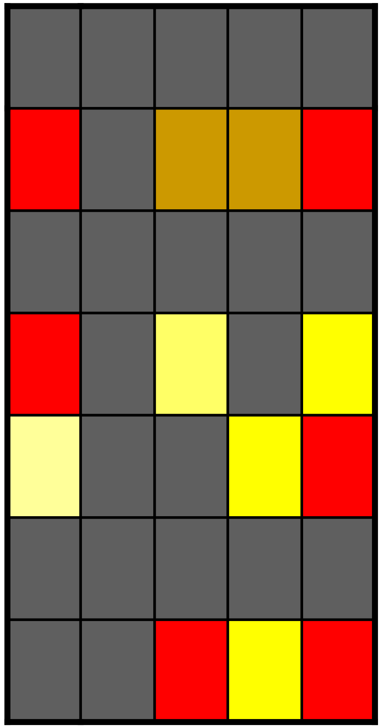
- Notice that we only need to consider non-zero entries of **A,** so the computational cost is dominated by the number of non-zero.

Outline

- 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**:



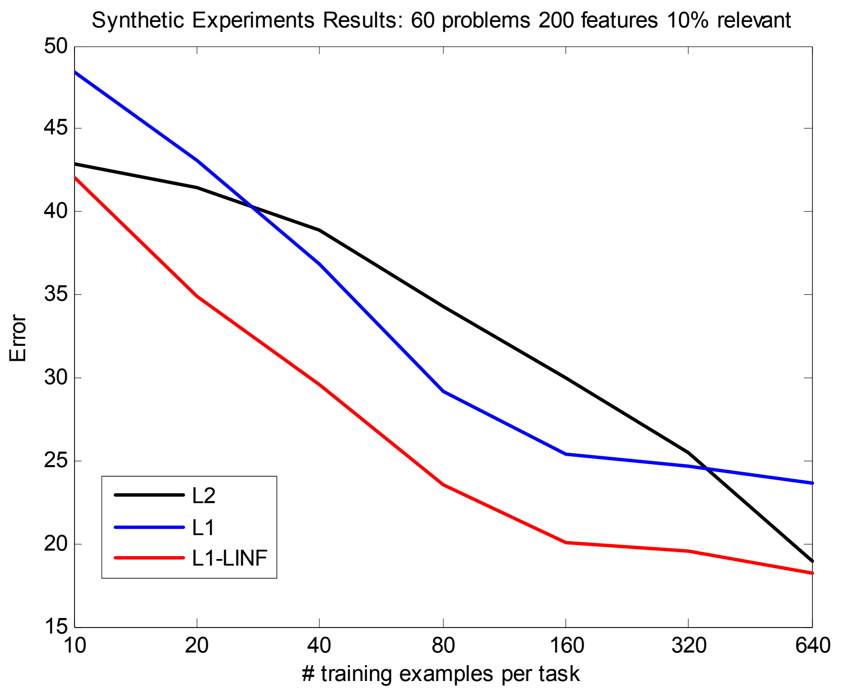
- For every task we generate pairs: (x_i^k, y_i^k)

  where $y_i^k = sign(w_k^t x_i^k)$

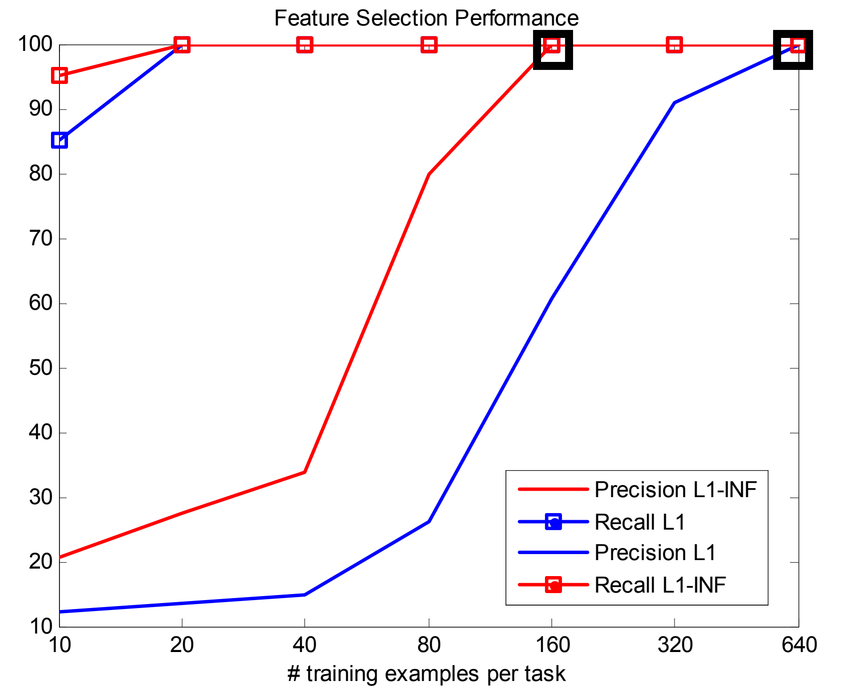
- We compared three different types of regularization (i.e. projections):
  - $L_{1-\infty}$ projection
  - L2 projection
  - L1 projection

Synthetic Experiments

Test Error



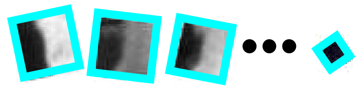
Performance on predicting relevant features

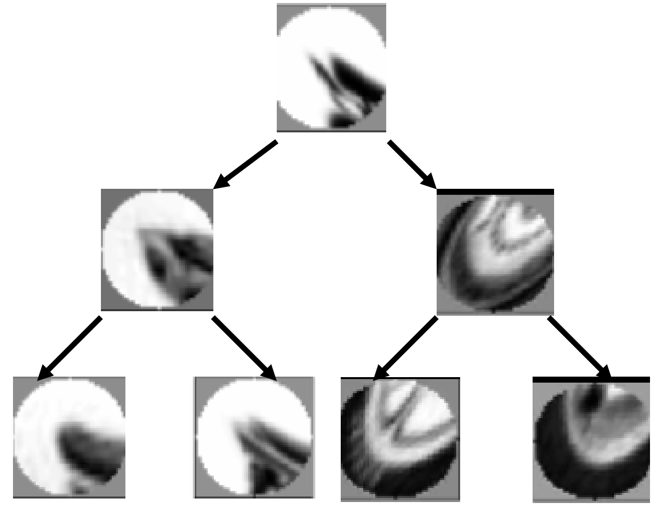


Dataset: Image Annotation

- 40 top content words
- Raw image representation: Vocabulary Tree (Grauman and Darrell 2005, Nister and Stewenius 2006)
- 11000 dimensions

Experiments: Vocabulary Tree representation





- Find patches
- Map each patch to a feature vector.
- Perform hierarchical k-means

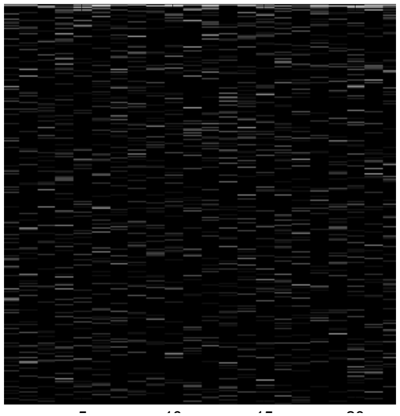
- To compute a representation for an image:
- Find patches.
- Map each patch to its closest cluster in each level.

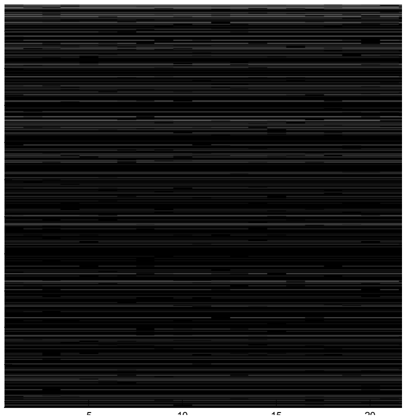
$$x = [\#c_1^1, \#c_2^1, \ldots, c_{p_1}^1, \ldots.\#c_{1,}^l \#c_2^l, \ldots., \#c_{p_l}^l]$$

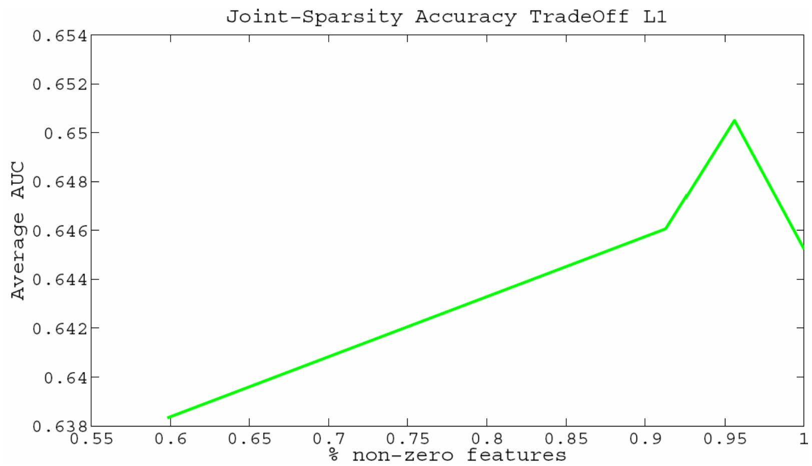
Results

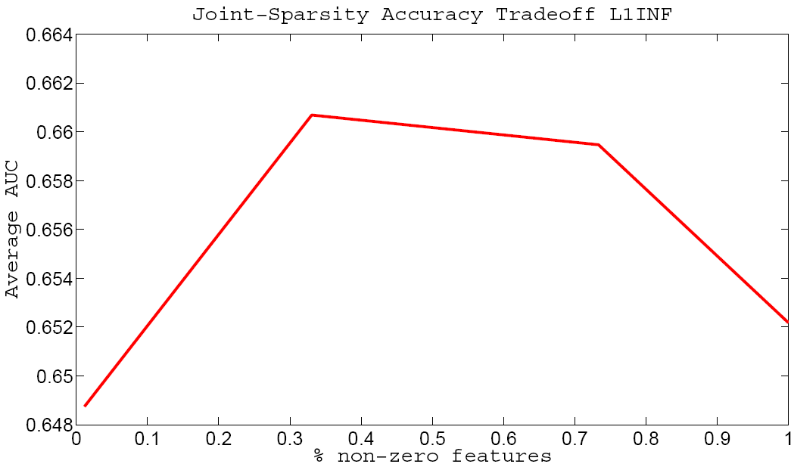

Performance Comparison
L1INF
L1
L2
Average AUC
0.74
0.73
0.72
0.71
0.7
0.69
0.68
0.67
0.66
0.65
0.64
4800
9600
19200
38400
63408
training samples

Results

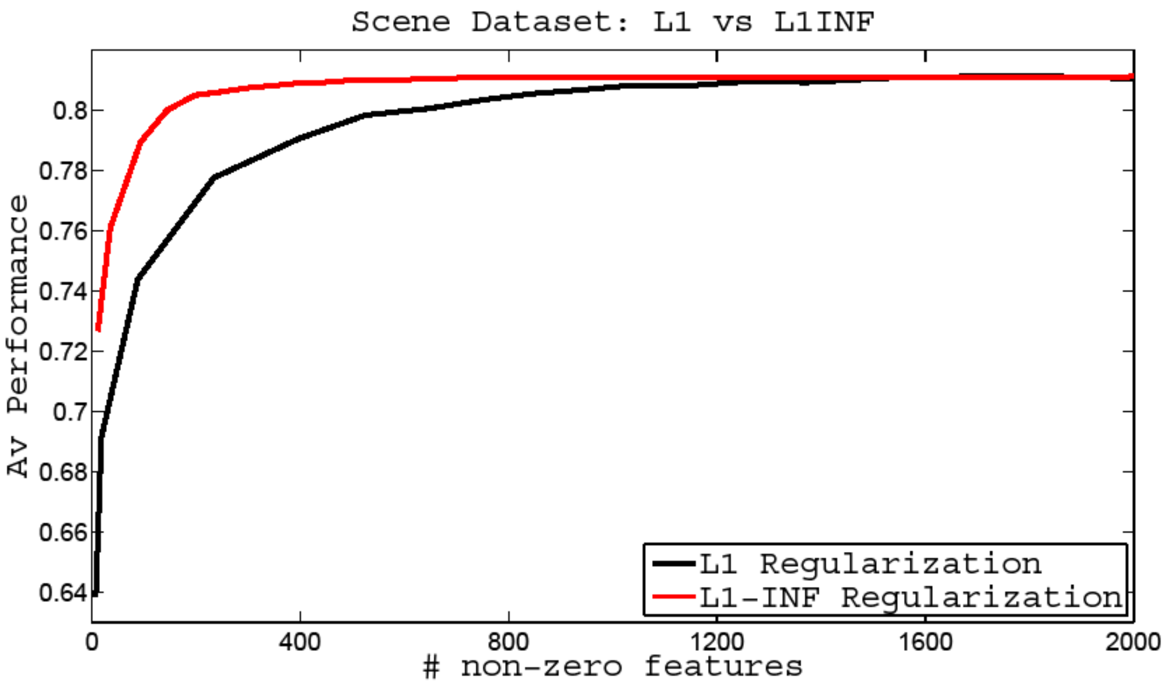





Joint-Sparsity Accuracy TradeOff L1
Average AUC
% non-zero features
0.654
0.652
0.65
0.648
0.646
0.644
0.642
0.64
0.638
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1


Joint-Sparsity Accuracy Tradeoff L1INF
Average AUC
% non-zero features
0.664
0.662
0.66
0.658
0.656
0.654
0.652
0.65
0.648
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Results:



Scene Dataset: L1 vs L1INF
Av Performance
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0
400
800
1200
1600
2000
non-zero features
L1 Regularization
L1-INF Regularization

Summary of Thesis Contributions

- We presented a method that learns efficient image representations using unlabeled images + meta-data.

- We developed a feature sharing transfer based on performing a joint loss minimization over the training sets of related tasks with a shared regularization.

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

- We propose a simple an efficient global optimization algorithm for training joint models with $L_{1-\infty}$ constraints.

- We provide a tool that makes implementing a joint sparsity regularization penalty as easy and almost as efficient as implementing the standard L1 and L2 penalties.

- We show the performance of our transfer algorithm 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!