Self-taught learning: Transfer learning from unlabelled data

Authors: Rajat Raina et al.
@Stanford University
Presenter: Shao-Chuan Wang

Some slides are from author's work

Self-taught learning: Transfer learning from unlabelled data

- Objective:
  - Use the unlabelled data to improve the performance on a classification task.

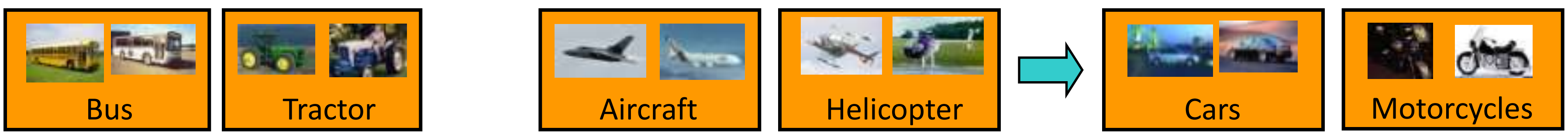
- Kea Ideas:
  - Relax the assumption about the unlabelled data.
  - Use unlabelled data to learn the best representation (dictionary)

Machine Learning Schemes

- Supervised learning

- Semi-supervised learning.

- Transfer learning.



- Next: Self-taught learning?

Self-taught Learning

- Labeled examples:

$$\{(x_l^{(i)}, y^{(i)})\}_{i=1}^{m} \qquad x_l^{(i)} \in R^n, y^{(i)} \in \{1, \ldots, T\}$$

- Unlabeled examples:

$$\{x_u^{(i)}\}_{i=1}^{k} \qquad x_u^{(i)} \in R^n, k >> m$$

- The unlabeled and labeled data:
  - **Need not share labels *y*.**
  - **Need not share a generative distribution.**

Advantage: Such unlabeled data is often easy to obtain.

Self-taught learning

Table 1. Details of self-taught learning applications evaluated in the experiments.

Domain	Unlabeled data	Labeled data	Classes	Raw features
Image classification	10 images of outdoor scenes	Caltech101 image classification dataset	101	Intensities in 14x14 pixel patch
Handwritten character recognition	Handwritten digits ("0"–"9")	Handwritten English characters ("a"–"z")	26	Intensities in 28x28 pixel character/digit image
Font character recognition	Handwritten English characters ("a"–"z")	Font characters ("a"/"A" – "z"/"Z")	26	Intensities in 28x28 pixel character image
Song genre classification	Song snippets from 10 genres	Song snippets from 7 *different* genres	7	Log-frequency spectrogram over 50ms time windows
Webpage classification	100,000 news articles (Reuters newswire)	Categorized webpages (from DMOZ hierarchy)	2	Bag-of-words with 500 word vocabulary
UseNet article classification	100,000 news articles (Reuters newswire)	Categorized UseNet posts (from "SRAA" dataset)	2	Bag-of-words with 377 word vocabulary

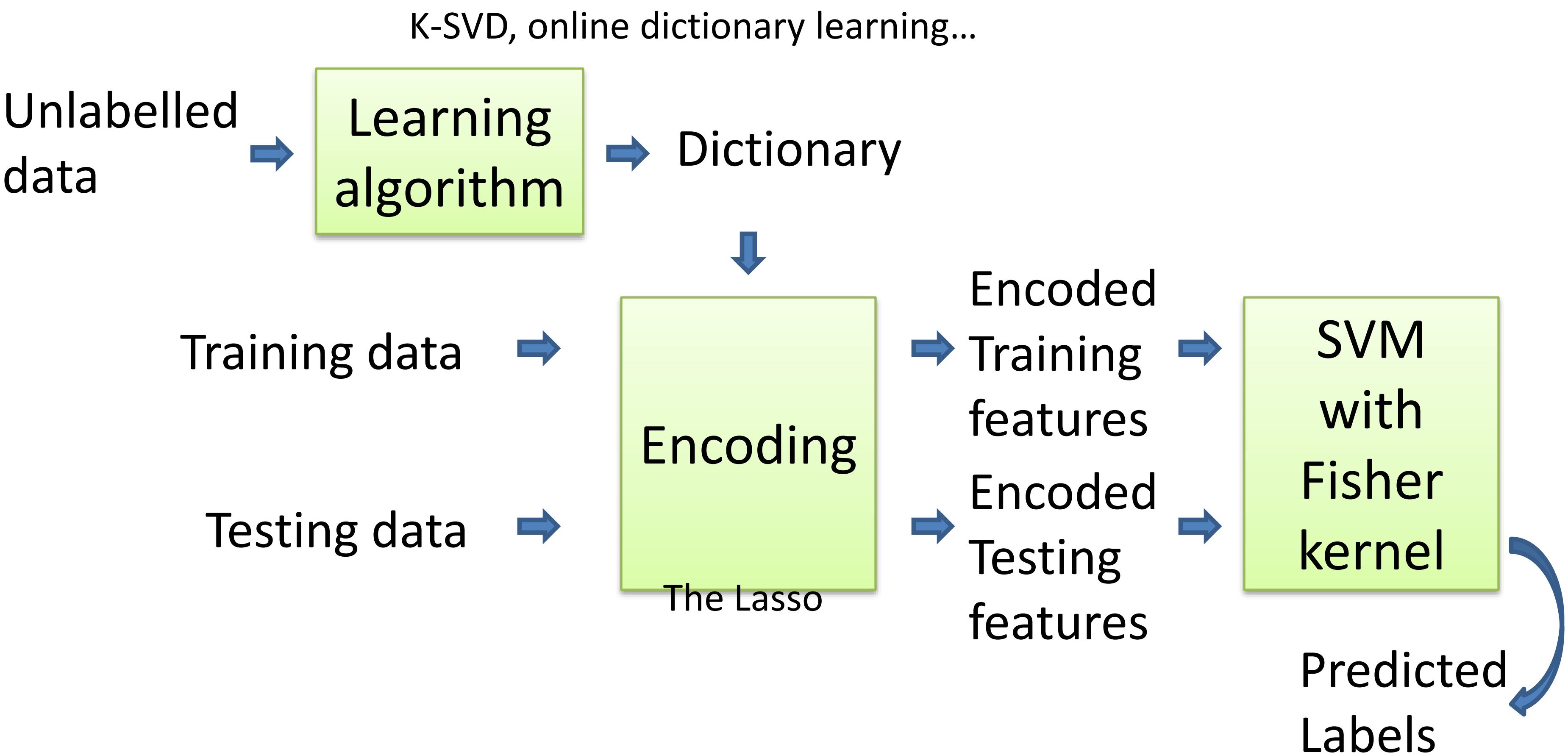
Use unlabelled data to learn the best representation.

Learning the structure

- Sparse coding: learning the dictionary
  - Encoding (L1-regularized least square problem)
    - Least angle regression (Efron et al. 2004) (very fast!!)
    - Feature-sign search (Honglak Lee et al. 2006)
    - Coordinate descent (Friedman et al. 2008)
  - Update dictionary(L2-constrained least square problem)
    - K-SVD (Aharon et al. 2006)
    - Online dictionary learning (Mairal et al. 2009)

$$\min_{d_j, \alpha_j^{(i)}} \underbrace{\sum_i \| x_u^{(i)} - \sum_j \alpha_j^{(i)} d_j \|_2^2}_{\text{Reconstruction error}} + \lambda \underbrace{\sum_i \| \alpha^{(i)} \|_1}_{\text{Sparsity penalty}}$$

Self-taught learning: flow



SVM with Fisher kernel

- Fisher kernel

$$U_x = \nabla_d \log P(x, \alpha \mid d) \qquad K(X_i, X_j) = U_x^{\ T} I^{-1} U_x$$

$$x = \hat{x} + \hat{r} \qquad \hat{x} = D\alpha$$

- In Bayesian view, $\hat{r} \sim \exp(-\|x - \hat{x}\|_2^2)$

$$P(\alpha) \sim \exp(-\lambda \sum_j |\alpha_j|)$$ Laplace prior

$$P(x, \alpha \mid d) = P(x \mid d, \alpha) P(\alpha) \propto \exp(-\|x - \hat{x}\|_2^2) \exp(-\lambda \sum_j |\alpha_j|)$$

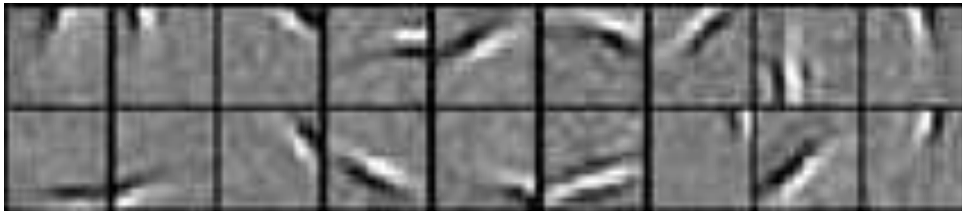
$$U_x = \nabla_d \log P(x, \alpha \mid d) = C \nabla_d (\|x - \hat{x}\|_2^2 + \lambda \sum_j |\alpha_j|)$$

$$\boxed{K(X_i, X_j) = (\alpha^{(i)^T} \alpha^{(j)}) \cdot (r^{(i)^T} r^{(j)})}$$

Example atoms

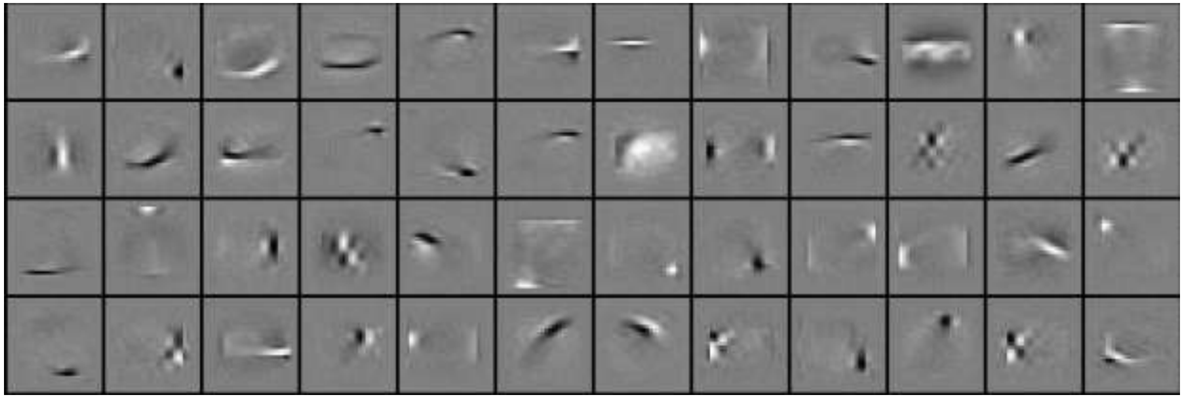
Natural images.

"edges"



Handwritten characters.

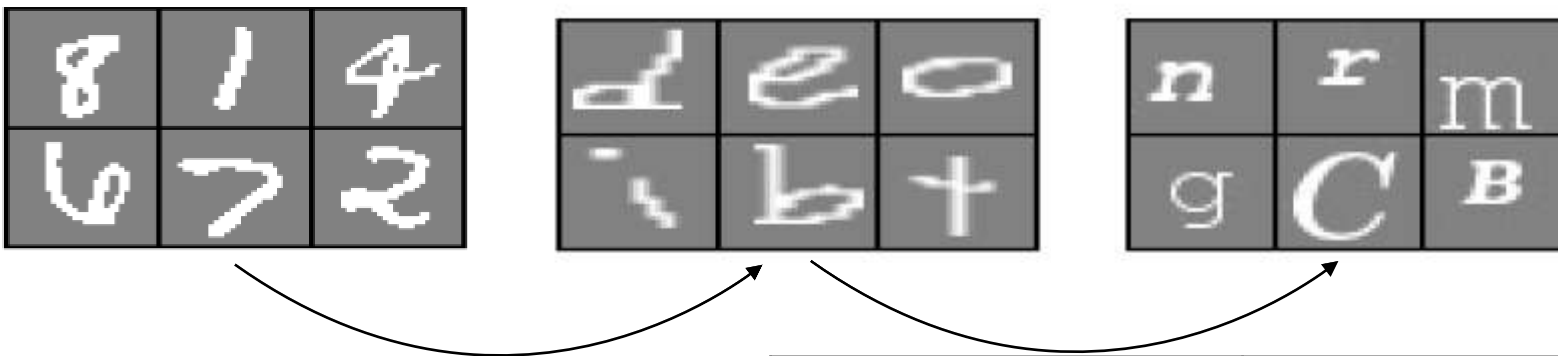
"strokes"



Sparse representation gives a higher level representation

Result: Character recognition

Digits | Handwritten English | English font



Raw	54.8%
PCA	54.8%
Sparse coding	**58.5%**

Handwritten English classification

(20 labeled images per handwritten character)

Bases learnt on digits

8.2% error reduction

Raw	17.9%
PCA	14.5%
Sparse coding	16.6%
Sparse coding + Raw	**20.2%**

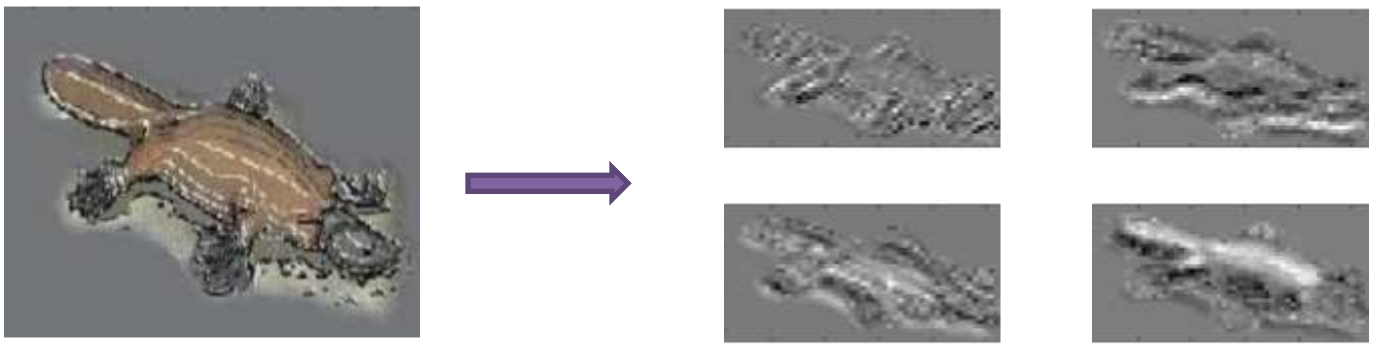
English font classification

(20 labeled images per font character)

Bases learnt on handwritten English

2.8% error reduction

Image classification



Baseline	16%
PCA	37%
Sparse coding	**47%**

(15 labeled images per class)

36.0% error reduction

Other reported results:
Fei-Fei et al, 2004: 16%
Berg et al., 2005: 17%
Holub et al., 2005: 40%
Serre et al., 2005: 35%
Berg et al, 2005: 48%

Zhang et al., 2006: **59%**
Lazebnik et al., 2006: 56%

Discussion

- Why sparse coding? Can this idea (learning structure from unlabelled image) be applied to other encoding schemes?
  - Probably because the nonlinear coding from x to alpha? Emulation of “end-stopping” phenomenon. (A feature is maximally activated by edges of only a specific orientation and length)

Summary

- Self-taught learning: Unlabeled data does not share the labels of the classification task.

- Use unlabeled data to discover features.
- Use sparse coding to construct an easy-to-classify, "higher-level" representation.