Two-way Multimodal Online Matrix Factorization

JORGE A. VANEGAS

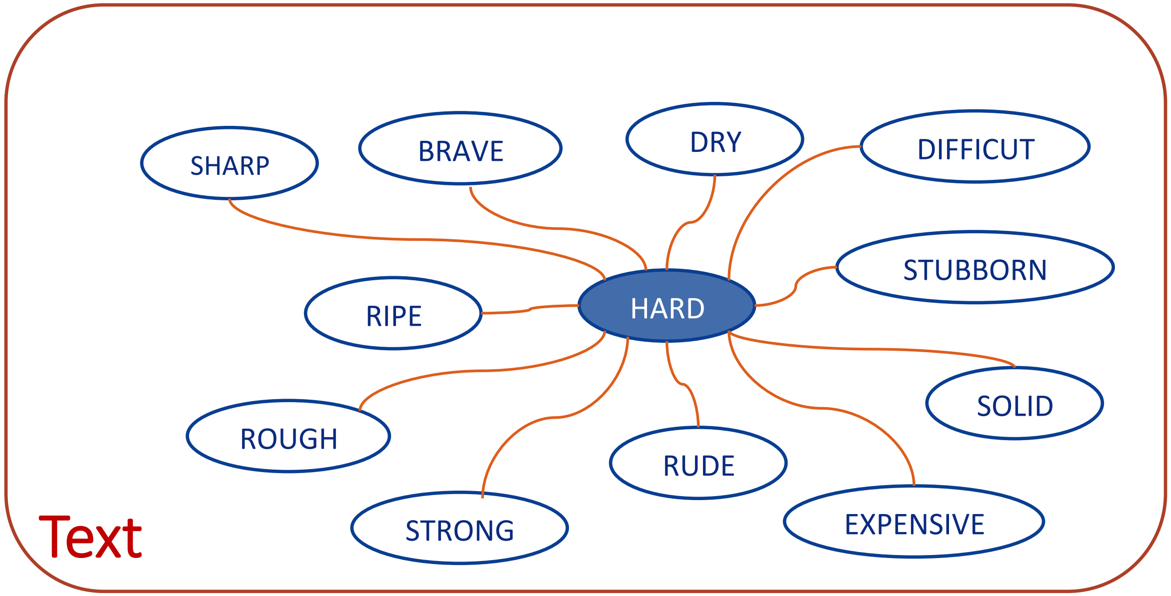
MindLab Research Group, Universidad Nacional de Colombia, BogotÁ, Colombia

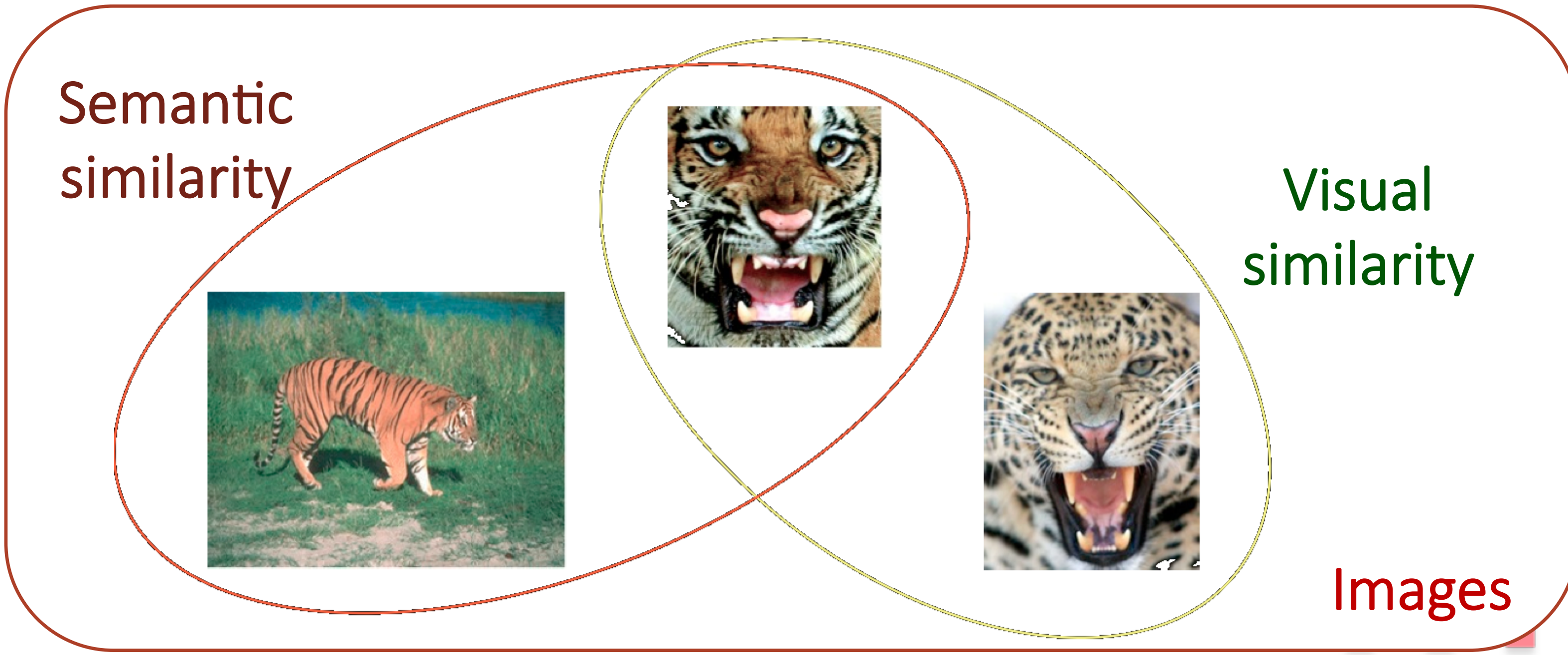
Outline

1. Introduction
2. Two-way Multimodal Online Matrix Factorization for Multi-label Annotation
3. Semi-supervised Dimensionality Reduction via Multimodal Matrix Factorization

Semantic gap

Synonymy and polysemy





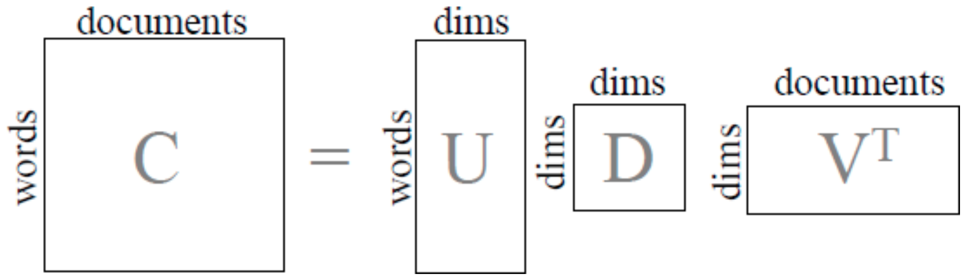
Dimensionality reduction

Eliminate the **redundancy** and the **noise** present in the manifold structure of the original high dimensional feature representation.

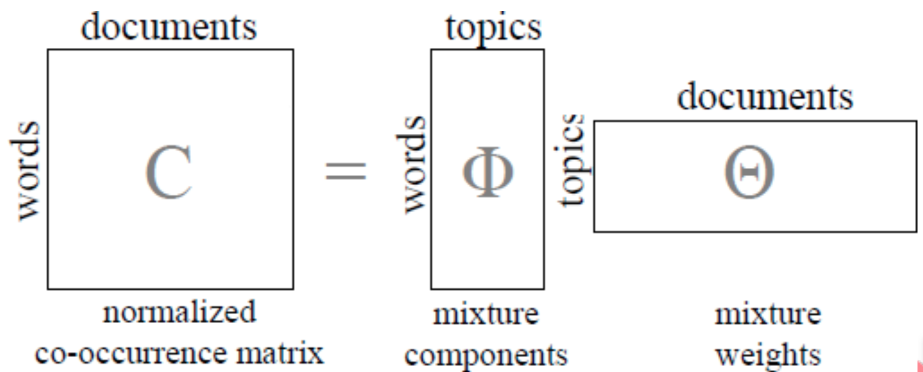
Tackles the **curse of dimensionality** by compressing the representation in a more expressive reduced set of variables.

Semantic representation via matrix factorization

- Latent Semantic Analysis (LSA)
  [Dumai et al. 2004]



- Nonnegative Matrix Factorization (NMF)
  [Lee et al. 1999]



Two-way Multimodal Online Matrix Factorization for Multi-label Annotation (ICPRAM)

Jorge A. Vanegas

MindLAB Research Group - Universidad Nacional de Colombia

November 24, 2015

Multi-label Annotation

The multi-label annotation problem arises in situations such as object recognition in images where we want to automatically find the objects present in a given image.
The solution consists in learning a classification model able to assign one or many labels to a particular sample.

Two-way Multimodal Matrix Factorization

$$\mathbf{X} \xrightarrow{f} \mathbf{H} \xrightarrow{g} \mathbf{X}$$

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^r,$$

$$g : \mathbb{R}^r \rightarrow \mathbb{R}^n$$

$$\text{where } n \gg r$$

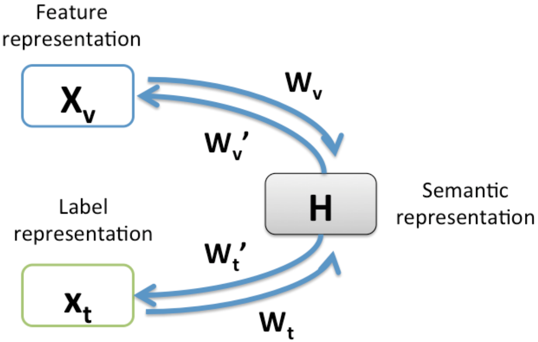
Reconstruction of the original feature representation:

$$X_v \approx W_v^{'} W_v X_v \tag{1}$$

Reconstruction of the original label representation:

$$X_t \approx W_t^{'} W_t X_t \tag{2}$$

Two-way Multimodal Matrix Factorization



$$X_t = W_t' W_v X_v$$

Optimization problem

$$
\begin{aligned}
L\left(X_v, X_t, W_v, W_v^{'}, W_t, W_t^{'}\right) = \\
\alpha \left\| X_v - W_v^{'} W_v X_v \right\|_F^2 + \\
+ (1-\alpha) \left\| X_t - W_t^{'} W_t X_t \right\|_F^2 \\
+ \delta \left\| X_t - W_t^{'} W_v X_v \right\|_F^2 \\
+ \beta \left(\| W_v \|_F^2 + \left\| W_v^{'} \right\|_F^2 + \| W_t \|_F^2 + \left\| W_t^{'} \right\|_F^2 \right) \qquad (3)
\end{aligned}
$$

α controls the relative importance between the reconstruction of the instance representation and the label representation.

δ controls the relative importance of the mapping between instance features and label information

β controls the relative importance of the regularization terms, which penalizes large values for the Frobenius norm of the transformation matrices.

Online learning

Loss function:

$$Q(z, w) = \ell(f_w(x), y)$$

Gradient descent:

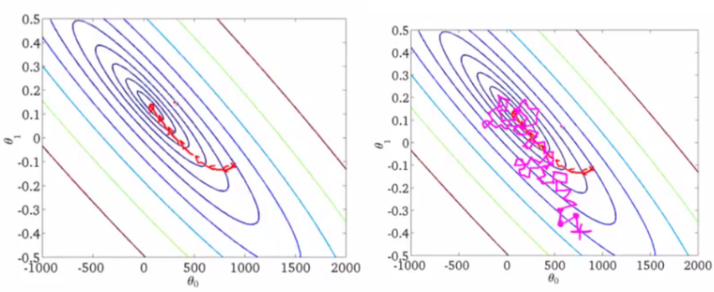
$$w_{t+1} = w_t - \gamma \frac{1}{n} \sum_{i=1}^{n} \nabla_w Q(z_i, w_t),$$

Stochastic gradient descent:

$$w_{t+1} = w_t - \gamma_t \nabla_w Q(z_t, w_t).$$

Online learning

- The learning rate ϒ can be either constant or gradually decaying
- “Generally" move in the direction of the global minimum, but not always
- Never actually converges like batch gradient descent does, but ends up wandering around some region close to the global minimum In practice, this isn't a problem



Two-way multimodal online matrix factorization algorithm

input r:latent space size, γ_0: initial step size, *epochs*: number of epochs,
$X_v \in \mathbb{R}^{n\times l}$, $X_t \in \mathbb{R}^{m\times l}$, α, δ, β

Random initialization of transformation matrices:

$$W_v'^{(0)} = \mathrm{random\_matrix}\,(r, n)$$
$$W_v^{(0)} = \mathrm{random\_matrix}\,(n, r)$$
$$W_t'^{(0)} = \mathrm{random\_matrix}\,(r, m)$$
$$W_t^{(0)} = \mathrm{random\_matrix}\,(m, r)$$

for $i = 1$ **to** *epochs* **do**

for $j =$ **1 to** l **do**

$\tau = i \times j$

$x_v^{(\tau)}, x_t^{(\tau)} \leftarrow \mathrm{sample\_without\_replacement}(X_v, X_t)$

Compute gradients:

$$g_{W_v'}^{(\tau)} = \nabla_{W_v'} L\left(x_v^{(\tau)}, x_t^{(\tau)}, W_v^{(\tau)}, W_v'^{(\tau)}, W_t^{(\tau)}, W_t'^{(\tau)}\right)$$
$$g_{W_v}^{(\tau)} = \nabla_{W_v} L\left(x_v^{(\tau)}, x_t^{(\tau)}, W_v^{(\tau)}, W_v'^{(\tau)}, W_t^{(\tau)}, W_t'^{(\tau)}\right)$$
$$g_{W_t'}^{(\tau)} = \nabla_{W_t'} L\left(x_v^{(\tau)}, x_t^{(\tau)}, W_v^{(\tau)}, W_v'^{(\tau)}, W_t^{(\tau)}, W_t'^{(\tau)}\right)$$
$$g_{W_t}^{(\tau)} = \nabla_{W_t} L\left(x_v^{(\tau)}, x_t^{(\tau)}, W_v^{(\tau)}, W_v'^{(\tau)}, W_t^{(\tau)}, W_t'^{(\tau)}\right)$$

...

Two-way multimodal online matrix factorization algorithm

...

Update term calculation using momentum:

$$\triangle W_v^{'(\tau)} = -\gamma^{(\tau)} g_{W_v'}^{(\tau)} + p \triangle W_v^{'(\tau-1)}$$
$$\triangle W_v^{(\tau)} = -\gamma^{(\tau)} g_{W_v}^{(\tau)} + p \triangle W_v^{(\tau-1)}$$
$$\triangle W_t^{'(\tau)} = -\gamma^{(\tau)} g_{W_t'}^{(\tau)} + p \triangle W_t^{'(\tau-1)}$$
$$\triangle W_t^{(\tau)} = -\gamma^{(\tau)} g_{W_t}^{(\tau)} + p \triangle W_t^{(\tau-1)}$$

Update transformation matrices:

$$W_v^{'(\tau+1)} = W_v^{'(\tau)} + \triangle W_v^{'(\tau)}$$
$$W_v^{(\tau+1)} = W_v^{(\tau)} + \triangle W_v^{(\tau)}$$
$$W_t^{'(\tau+1)} = W_t^{'(\tau)} + \Delta W_t^{'(\tau)}$$
$$W_t^{(\tau+1)} = W_t^{(\tau)} + \Delta W_t^{(\tau)}$$

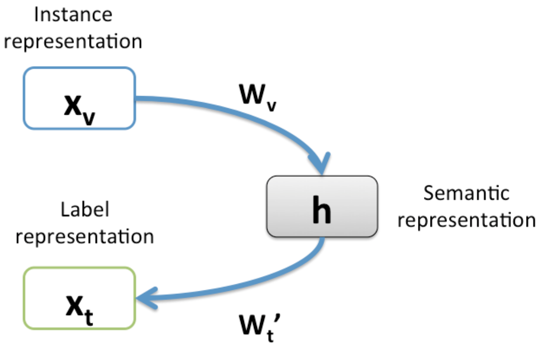
end for

end for

return $W_v^{'(N)}, W_v^{(N)}, W_t^{'(N)}, W_t^{(N)}$

Prediction

$$\tilde{x}_t = W_t' W_v x_v$$



Annotation

- The transformation of the input features generates an $m-$dimensional vector with an smoothed label representation
  - The final decision to assign a label would be taken by defining a threshold
  - assign 1 to the $j-th$ label if $\tilde{x_{t,j}} \geqq threshold$, or we can assign 1 to the top$-k$ labels with the highest values in the vector.

Implementation details

- **Pylearn2** library [1]
- theano[2]

[1] http://deeplearning.net/software/pylearn2/
[2] http://deeplearning.net/software/theano/

Experiments and Results

- 80% of the images for training
- the remaining 20% for test
- Were compared against 8 MLLSE[3] algorithms:
  - OVA: One-versus-all
  - CCA: Canonical Correlation Analysis
  - CS: Compressed Sensing
  - PLST: Principal Label Space Transform
  - MME: Multilabel max-margin embedding
  - ANMF, MNMF, OMMF

[3] multi-label latent space embedding

Datasets

Dataset	Corel5k	Bibtex	MediaMill
Labels	374	159	101
Features	500	1,836	120
Label cardinality	3,522	2,402	4.376
Examples	5,000	7,395	43,907

The method was evaluated on 3 standard multilabel datasets distributed by the mulan framework authors (http://mulan.sourceforge.net/datasets.html)

F-Measure for each method

Performance of each method in terms of f-measure

Method	Corel5k	Bibtex	MediaMill
OVA	0.112	0.372	—
CCA	0.150	0.404	—
CS	0.086 (50)	0.332 (50)	—
PLST	0.074 (50)	0.283 (50)	—
MME	0.178 (50)	0.403 (50)	0.199 (350)
ANMF	0.210 (30)	0.297 (140)	0.496 (350)
MNMF	0.240 (35)	0.376 (140)	0.510 (350)
OMMF	0.263 (40)	**0.436 (140)**	0.503 (350)
Our Method	**0.283 (100)**	0.422 (300)	**0.540 (300)**

Sebastian Otálora-Montenegro et al. [1]

Conclusions and Future Work

- We presented a novel multi-label annotation method which learns a mapping between the original sample representation and labels by finding a common semantic representation.
- We propose a model that finds a mapping from the sample representation space to a semantic space, and simultaneously finds a back-projection from the semantic space to the original space.
- This method is formulated as an online learning algorithm allowing to deal with large collections.
- One important limitation is that the method assumes linear dependencies between the data modalities

References

Sebastian Otálora-Montenegro, Santiago A. Pérez-Rubiano, and Fabio A. González.
Online matrix factorization for space embedding multilabel annotation.
In *CIARP (1)*, volume 8258 of *Lecture Notes in Computer Science*, pages 343–350. Springer, 2013.

Sunho Park and Seungjin Choi.
Max-margin embedding for multi-label learning.
Pattern Recognition Letters, 34(3):292 – 298, 2013.

Semi-supervised Dimensionality Reduction via Multimodal Matrix Factorization

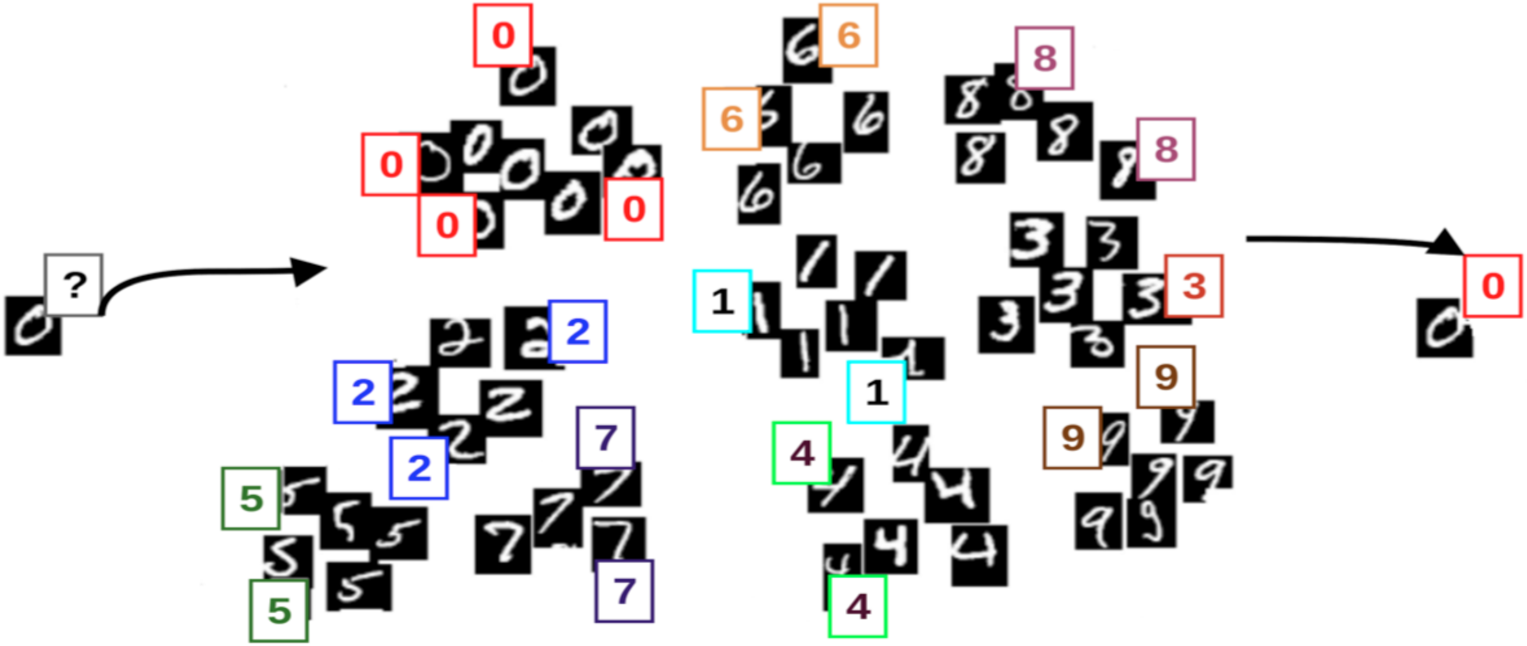
VIVIANA BELTRÁN, JORGE A. VANEGAS, FABIO A. GONZÁLEZ

MindLab Research Group, Universidad Nacional de Colombia, BogotÁ, Colombia

Motivation

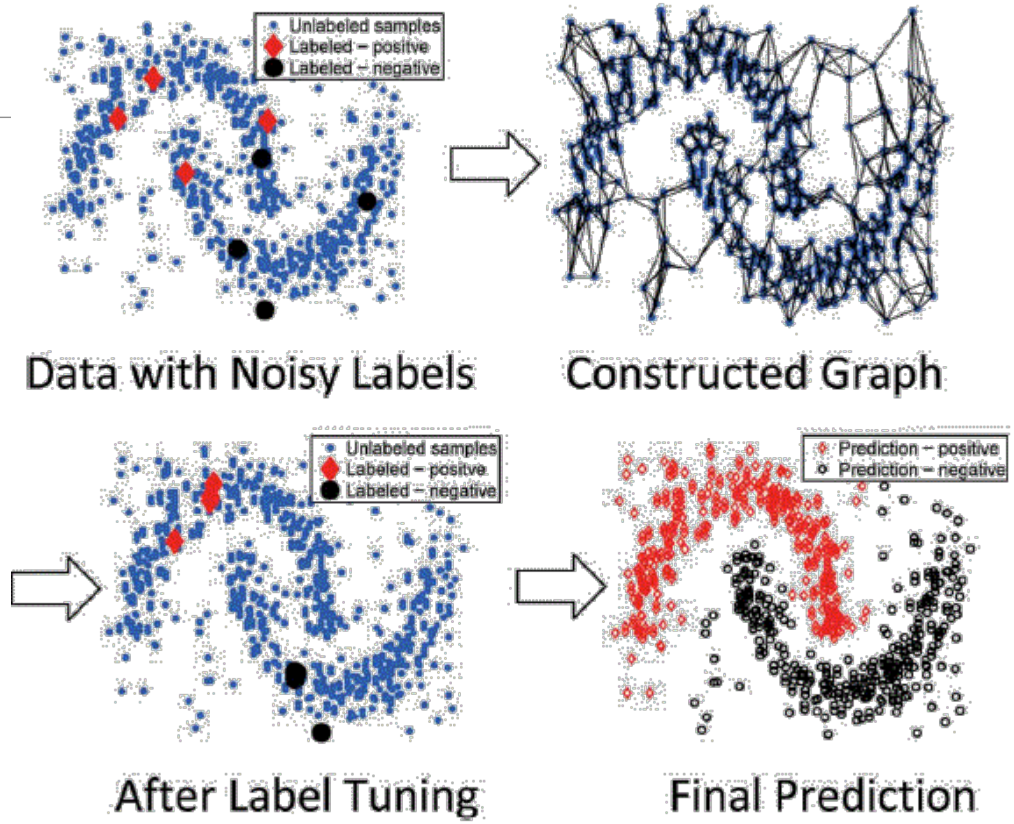
Challenges	Proposed Solution
High dimensional data	Low dimensional embedding representation
Reduced number of labeled data	Semi–supervised learning via matrix factorization
Huge amount of unstructured data: • Massive unlabeled examples available	• Stochastic gradient descent • GPU implementation

Semi-supervised learning



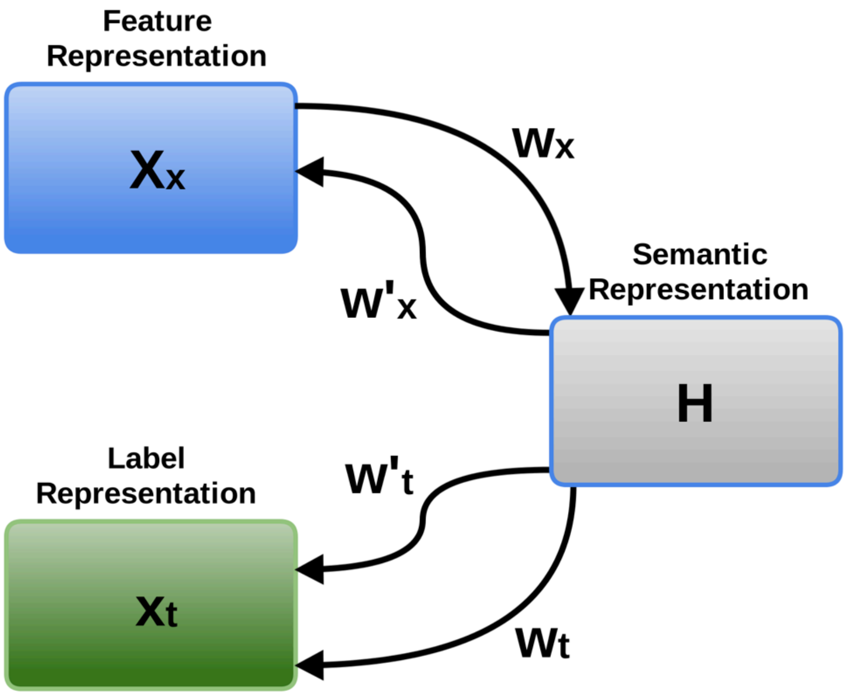
Manifold learning for classification

Semi-supervised learning



Annotated instances are used to maximize the discrimination between classes, but also, non-annotated instances can be exploited to estimate the intrinsic manifold structure of the data.

Model



$$X \xrightarrow{f} H \xrightarrow{g} X$$

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^r,$$

$$g : \mathbb{R}^r \rightarrow \mathbb{R}^n$$

n>>r

$$X_x \approx W_x' W_x X_x$$

$$X_t \approx W_t' W_t X_t$$

Model

$$L = \alpha \sum_{i=1}^{k} \left\| x_i - W_x' W_x x_i \right\|_F^2 + (1-\alpha) \sum_{i=1}^{l} \left\| t_i - W_t' W_t t_i \right\|_F^2$$
$$+ \delta \sum_{i=1}^{l} \left\| t_i - W_t' W_x x_i \right\|_F^2 + \beta \left(\|W_v\|_F^2 + \left\| W_v' \right\|_F^2 + \|W_t\|_F^2 + \left\| W_t' \right\|_F^2 \right)$$

***X**$_i$* feature vector of the *i-th* instance in the data collection *X*
t$_i$ binary label vector of the *i-th* instance
k instances for training
l labeled instances

k >> l

Experiments

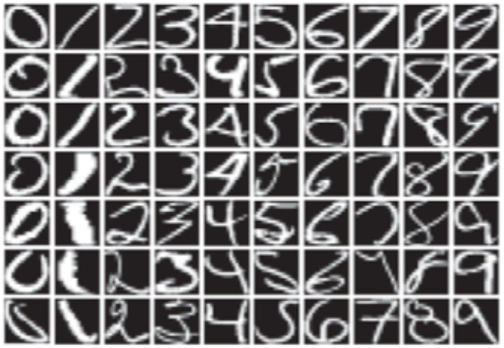
Experimental Setup:

- Parameter exploration by using **5-fold cross validation**.
- **Average classification accuracies** for 10 runs evaluated by using **1-KNN** setup similar as in [Zhao et al. 2014].
- Linear **supervised, semi-supervised and unsupervised** dimensionality reduction methods as baselines.

Datasets

MNIST

USPS



LETTERS

COVTYPE

Datasets

Datasets partitions

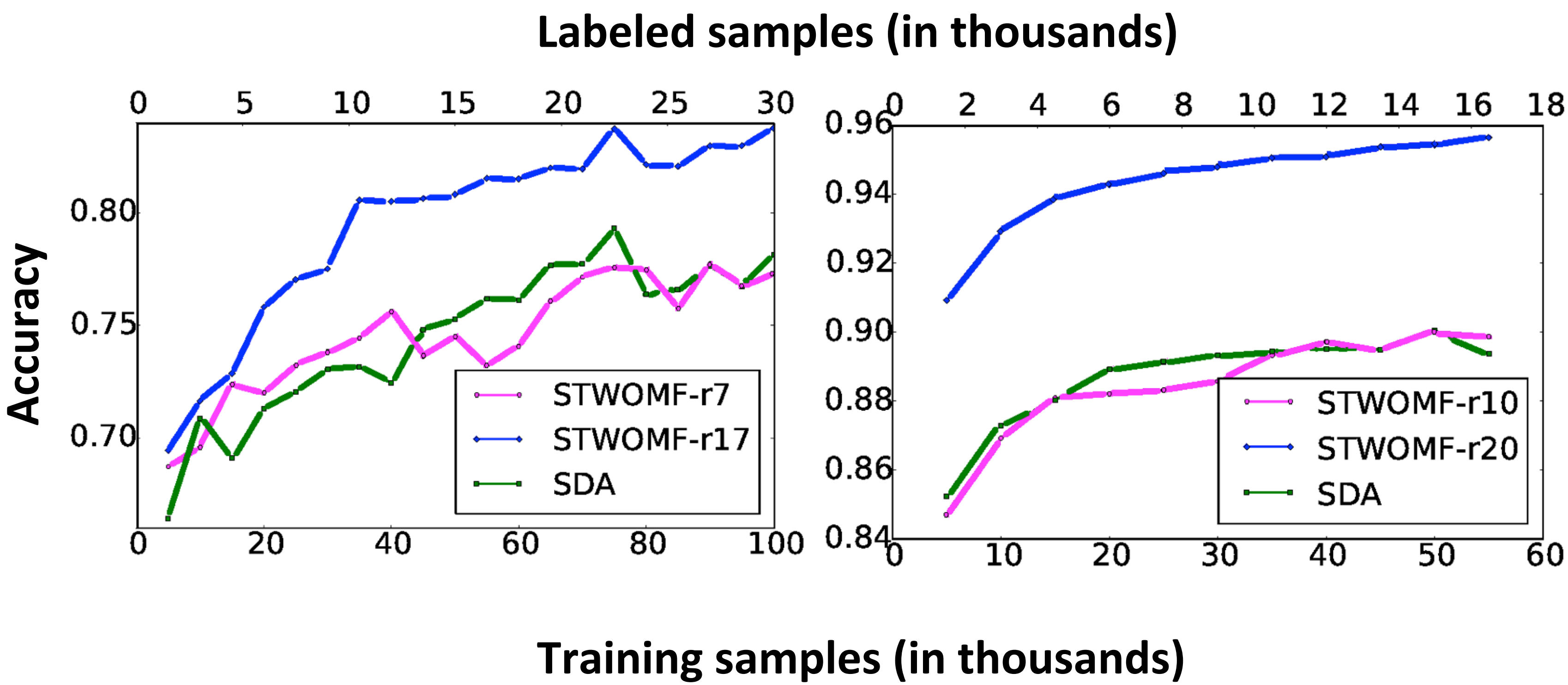
Dataset	Original dataset partitions		Low-scale partitions		Large-scale evaluation		#Dim	#Class
	Train	Test	Train	Test	Train	Test		
Covtype	581012		8000	8000	100000	2000	54	7
MNIST	60000	10000	8000	8000	60000	10000	784	10
Letters	20000		8000	8000	–		16	26
USPS	4649	4649	4649	4649	–		256	10

Classification accuracy for different percentages of annotated instances

METHOD		STWOMF r=C	STWOMF r=C+10	SDA	LDA	SRDA	PCA r=C
COVTYPE	100%	0.725	**0.770**	**0.735**	0.708	0.698	0.707
	60%	**0.720**	**0.755**	0.719	0.704	0.685	0.683
	30%	0.686	**0.712**	**0.687**	**0.707**	0.653	0.639
MNIST	100%	0.882	**0.939**	0.870	**0.897**	0.856	0.874
	60%	0.864	**0.930**	0.870	**0.890**	0.833	0.863
	30%	0.848	**0.916**	0.850	**0.881**	0.786	0.842
LETTERS	100%	**0.946**	**0.946**	**0.950**	0.699	0.936	0.940
	60%	**0.933**	**0.923**	**0.940**	0.694	0.919	0.913
	30%	**0.905**	0.885	**0.917**	0.680	**0.893**	0.872
USPS	100%	0.936	**0.966**	0.925	**0.943**	0.921	0.930
	60%	0.927	**0.957**	0.917	**0.939**	0.906	0.921
	30%	0.910	**0.942**	0.903	**0.926**	0.884	0.903

Avg. Classification Accuracy

Covtype **Mnist**



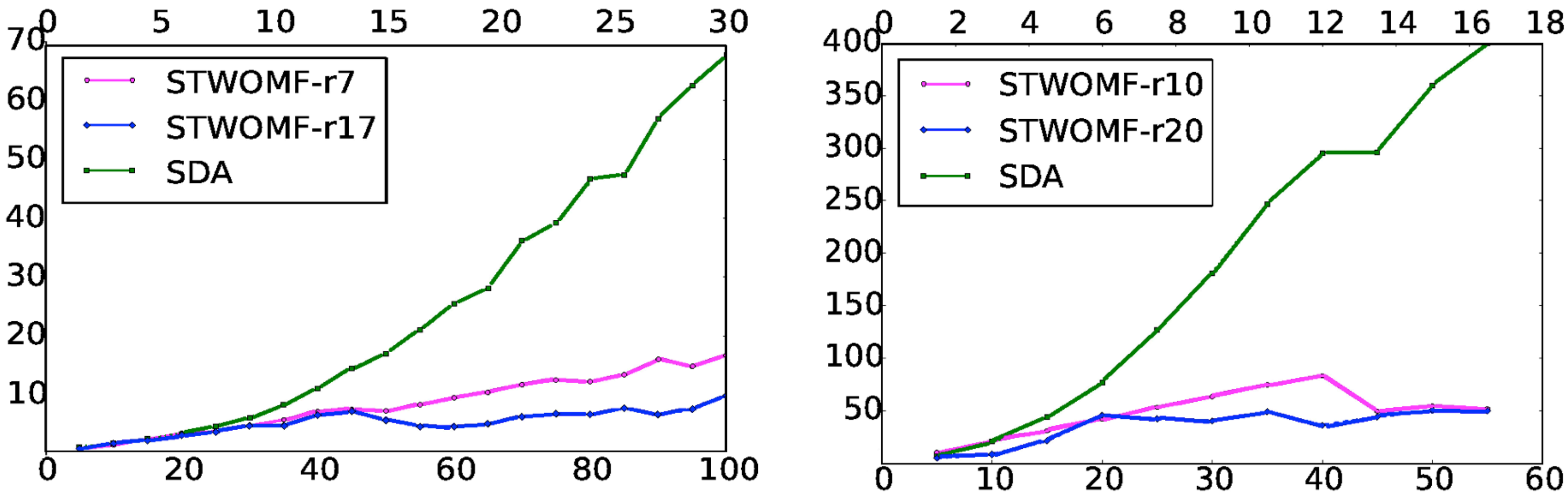
For all training sizes only 30% of instances are annotated

Avg. Training time

Covtype **Mnist**

Labeled samples (in thousands)

Training time in seconds



Training samples (in thousands)

For all training sizes only 30% of instances are annotated

Conclusions

We presented a method whose main characteristics are:

- Modeling a semantic low-space representation
- Preserving the separability of the original classes
- Ability to exploit unlabeled instances for modeling the manifold structure of the data
- Online formulation

References

[Zhao et al. 2014] Mingbo Zhao, Zhao Zhang, Tommy WS Chow, and Bing Li. A general soft label based linear discriminant analysis for semi-supervised dimensionality reduction. Neural Networks, 55:83–97, 2014.