Variants of Auto-Encoders

Zhifei Zhang

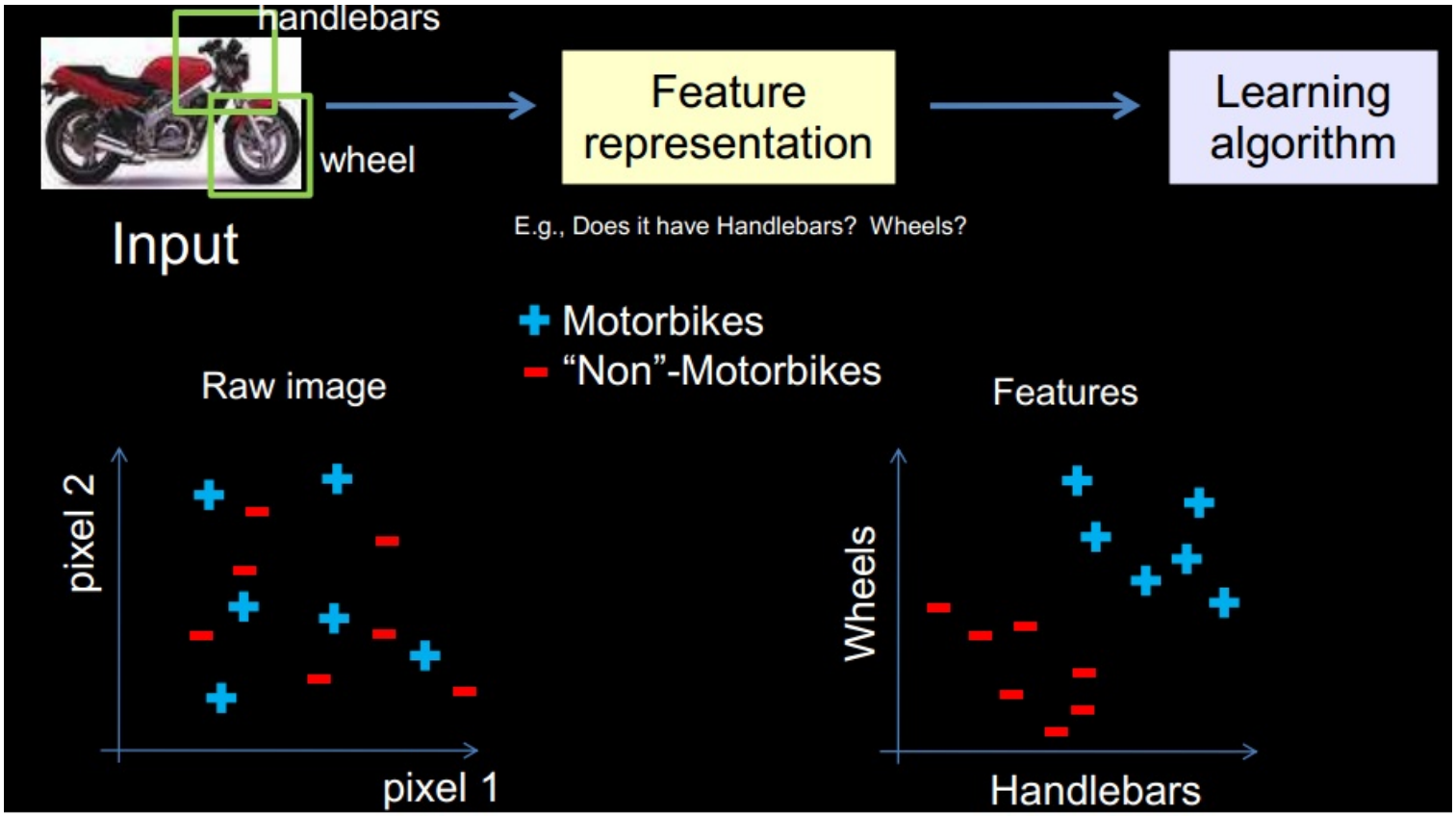
Department of Computer Science and Technology
Tongji University

Nov. 20, 2013

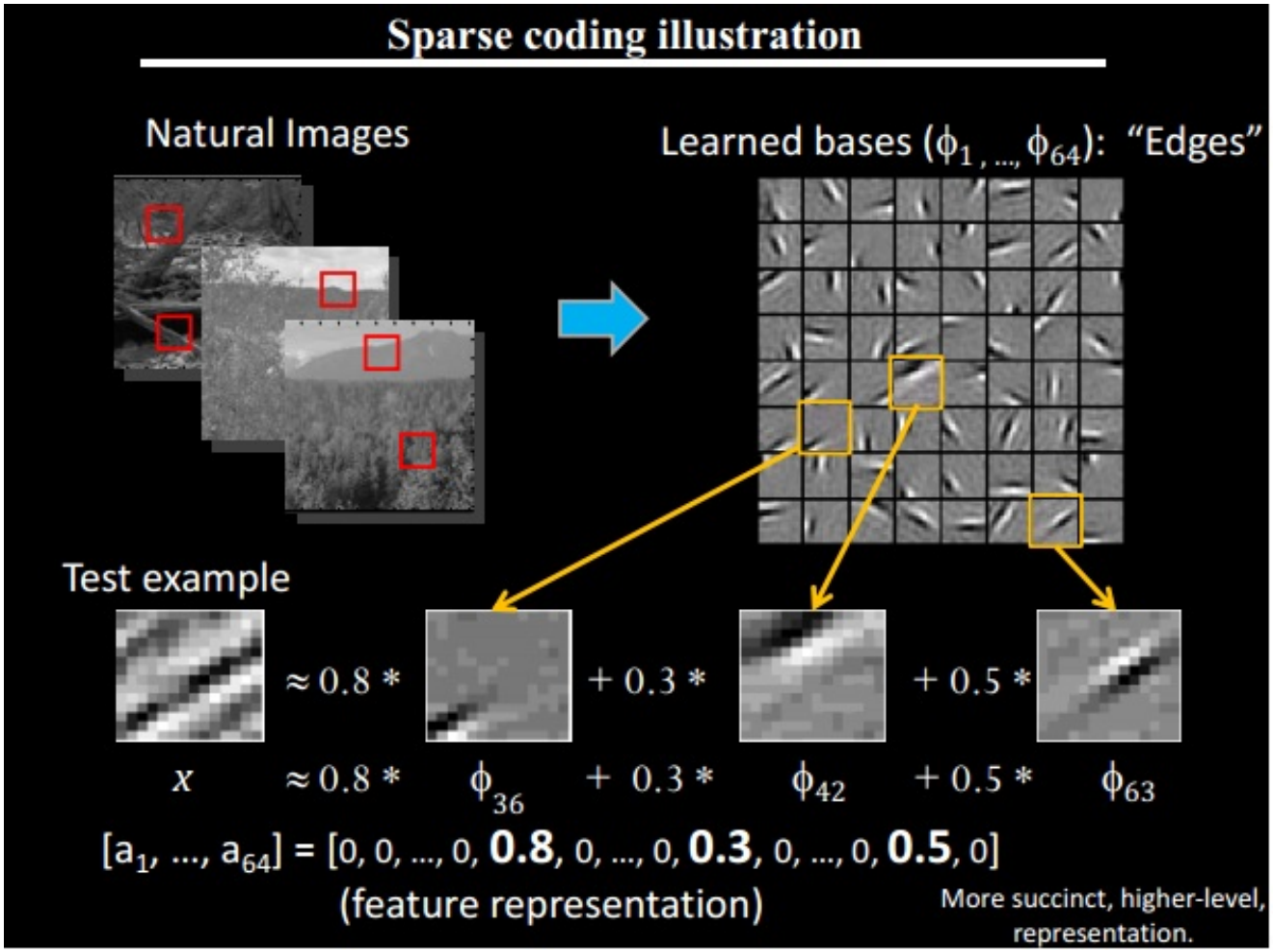
Outline

Outline

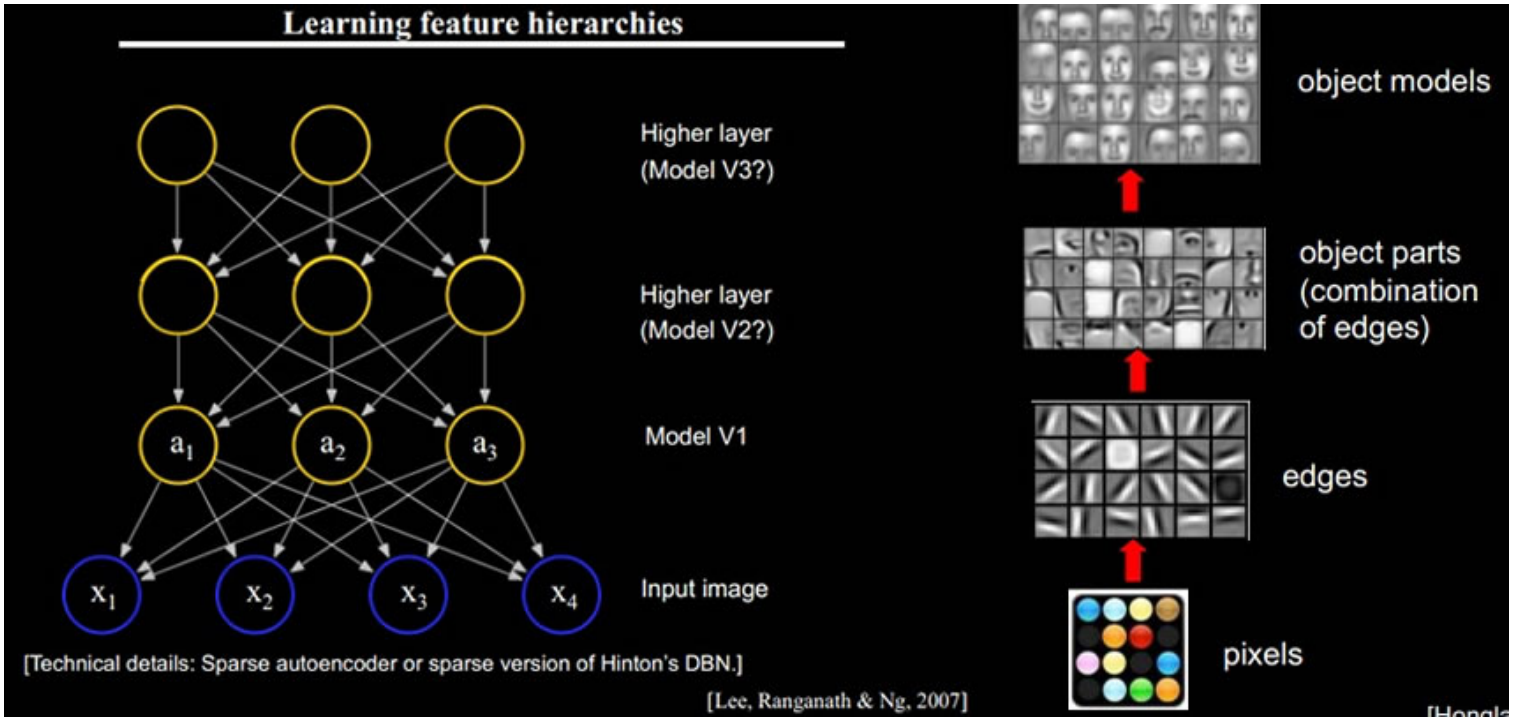
Granularity of Feature



Shallow Feature



Learning feature hierarchies



Topic Modelling

Outline

Signal Processing

- Nyquist-Shannon Sampling Theorem

$$f_s \geq 2B \tag{1}$$

- Compressed Sensing
  A paradigm shift that allows for the saving of time and space during the process of signal acquisition, while still allowing near perfect signal recovery when the signal is needed.
  It is possible to fully recover a signal from sampling points much fewer than that defined by the above sampling theorem.

Compressed Sensing

Given $\mathbf{x}$ of length N, only M measurements ($M < N$) is required to fully recover $\mathbf{x}$ when $\mathbf{x}$ is K-sparse ($K < M < N$).

Three essential criteria:

- Sparsity
- Incoherence
- Non-linear Reconstruction

The number of significant (strictly speaking, nonzero) components is relatively small compared to signal length.

Compressed Sensing

Given $\mathbf{x}$ of length N, only M measurements ($M < N$) is required to fully recover $\mathbf{x}$ when $\mathbf{x}$ is K-sparse ($K < M < N$).

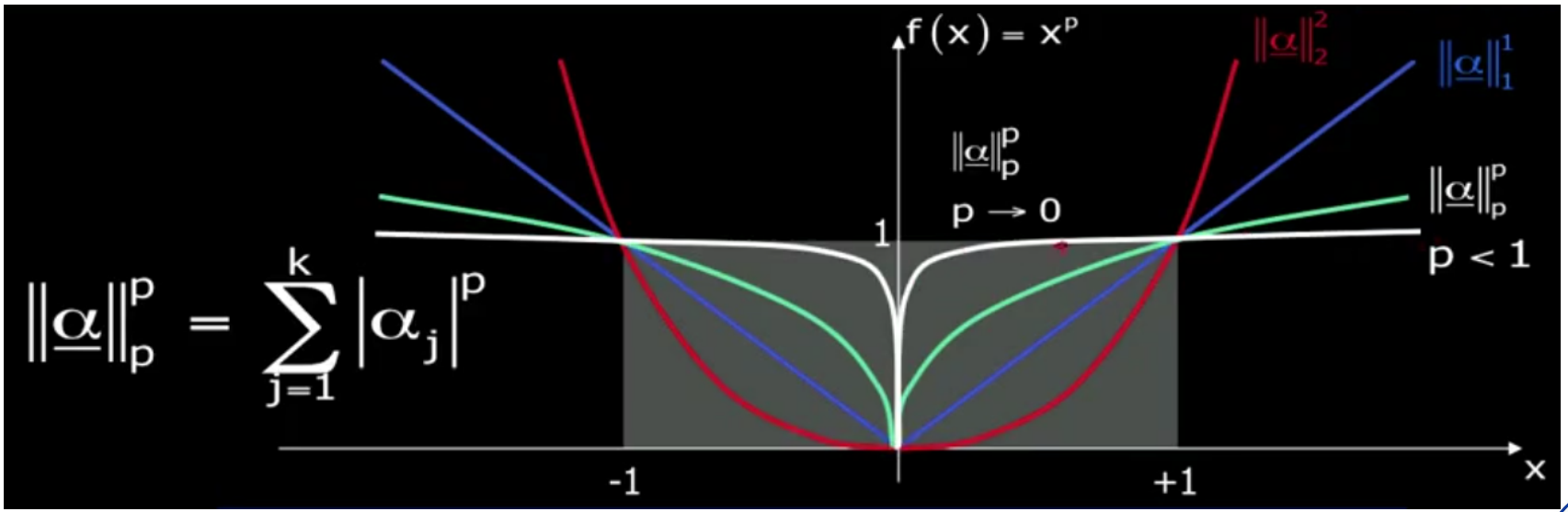
Three essential criteria:

- Sparsity
- Incoherence
- Non-linear Reconstruction

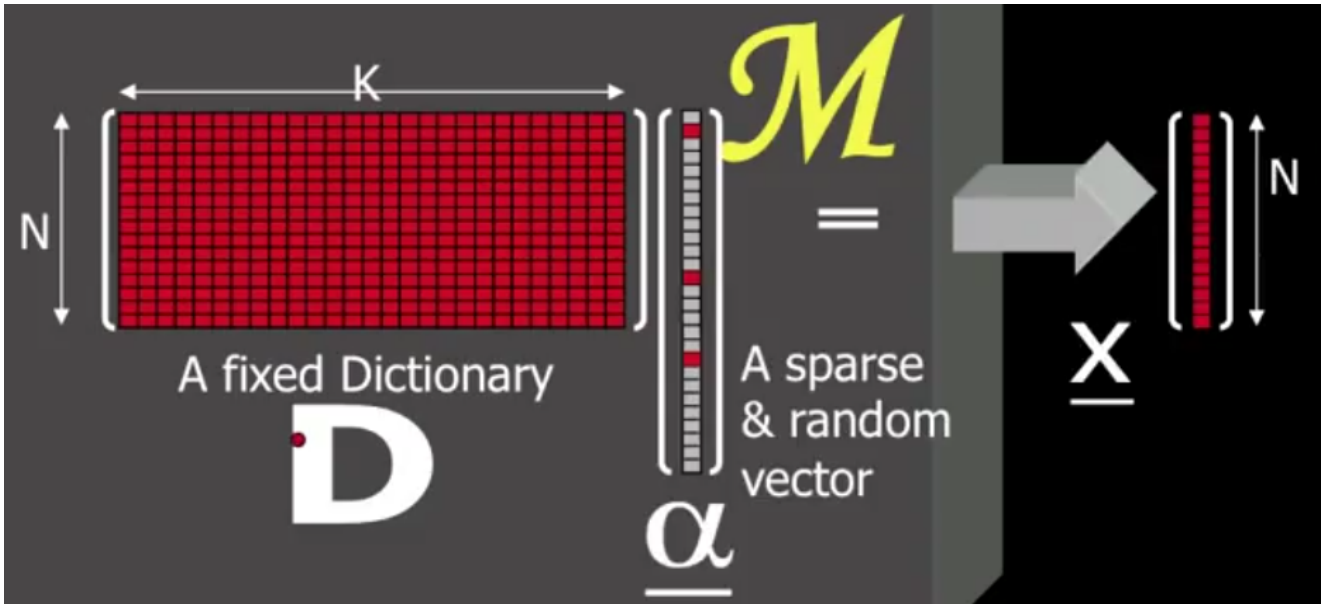
The number of significant (strictly speaking, nonzero) components is relatively small compared to signal length.

Sparsity Representation

- ℓ_p-Norm: $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{N} |x_i|^p \right)^{\frac{1}{p}}$
- ℓ_0-norm counts the number of non-zero components of $\mathbf{x}$.



Sparse Modelling



$$\min_{\mathbf{D},\alpha_j} \sum_{j=1}^{M} \|\mathbf{D}\alpha_j - \mathbf{x}_j\|_2^2 \quad \text{s.t.} \quad \|\alpha_j\|_0 < T \tag{2}$$

Uniqueness

$\mathbf{D}$ and $\mathbf{x}$ are known,

$$\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \mathbf{x} = \mathbf{D}\alpha \tag{3}$$

Spark: $\sigma = Spark(\mathbf{D})$ is the smallest number of columns that are linearly dependent.

e.g. $\begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$, $\sigma = 3$.

If we found a representation that satisfy $\frac{\sigma}{2} > \|\alpha\|_0$, then necessarily it is unique (the sparsest).

Sparse Coding

Problem Setting:

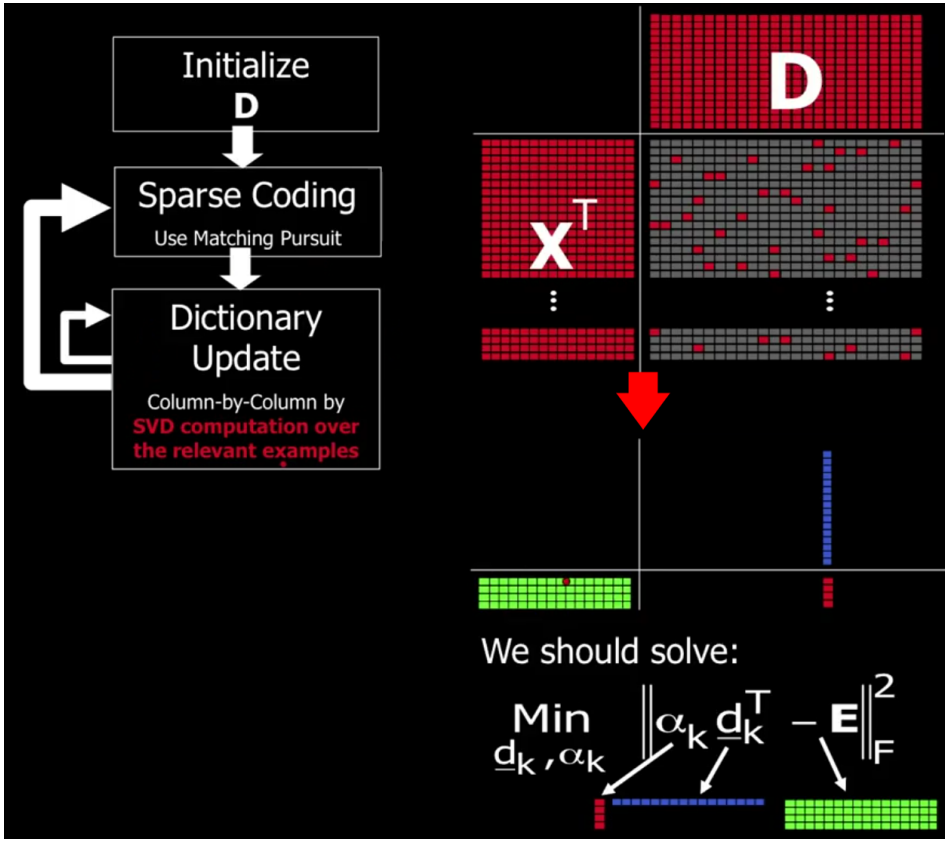
$$\min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \|\mathbf{D}\alpha - \mathbf{x}\|_2^2 \leq \varepsilon^2 \tag{4}$$

- Greedy Methods - Matching Pursuit (MP)
- Relaxation Methods - the Basis Pursuit (BP)

$$\min_{\alpha} \|\alpha\|_1 \quad \text{s.t.} \quad \|\mathbf{D}\alpha - \mathbf{x}\|_2 \leq \varepsilon \tag{5}$$

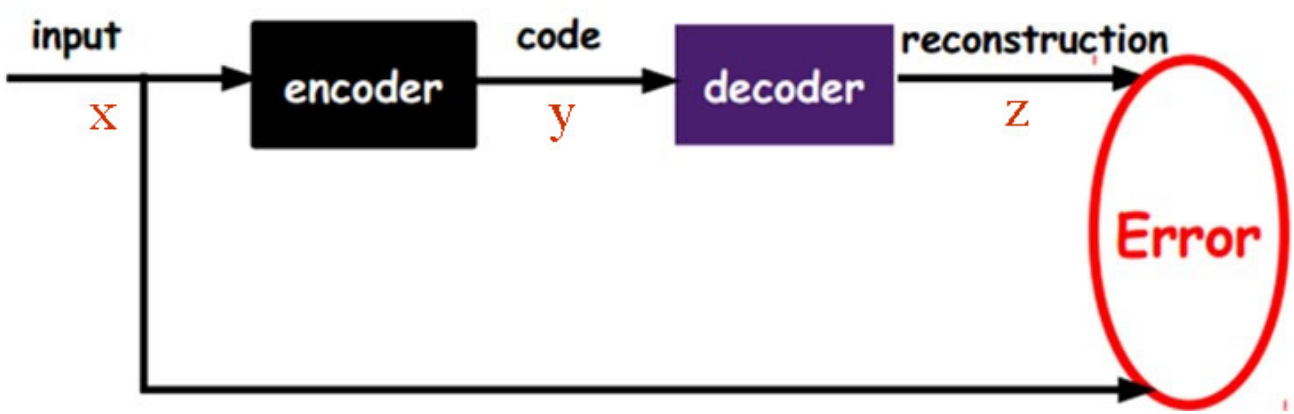
Dictionary Learning

K-SVD:



Outline

Basic Autoencoder



$\mathbf{x} \in [0,1]^d$, $\mathbf{y} \in [0,1]^{d'}$, $\mathbf{z} \in [0,1]^d$

Basic Autoencoder (Cont')

$$\mathbf{y} = f_\theta(\mathbf{x}) = s(\mathbf{W}\mathbf{x} + \mathbf{b}),\ \mathbf{z} = g_{\theta'}(\mathbf{y}) = s(\mathbf{W}'\mathbf{y} + \mathbf{b}')$$

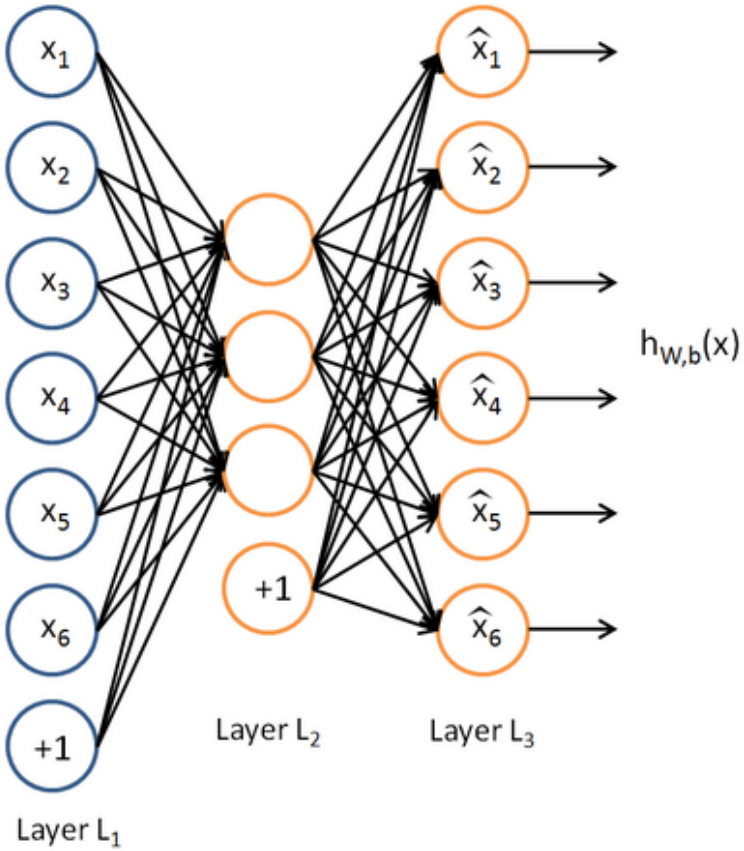
$$\theta^*, \theta'^* = \arg\min_{\theta,\theta'} \frac{1}{m} \sum_{i=1}^{m} L(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \tag{6}$$

where L is a loss function such as squared error $L(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|^2$. An alternative loss is reconstruction cross-entropy:

$$L_H(\mathbf{x}, \mathbf{z}) = H(B_\mathbf{x} || B_\mathbf{z}) = -\sum_{k=1}^{d} [x_k \log z_k + (1 - x_k) \log(1 - z_k)] \tag{7}$$

$$\theta^*, \theta'^* = \arg\min_{\theta,\theta'} E_{q^0(\mathbf{X})}[L_H(\mathbf{X}, g_{\theta'}(f_\theta(\mathbf{X})))] \tag{8}$$

Regularized Auto-Encoders



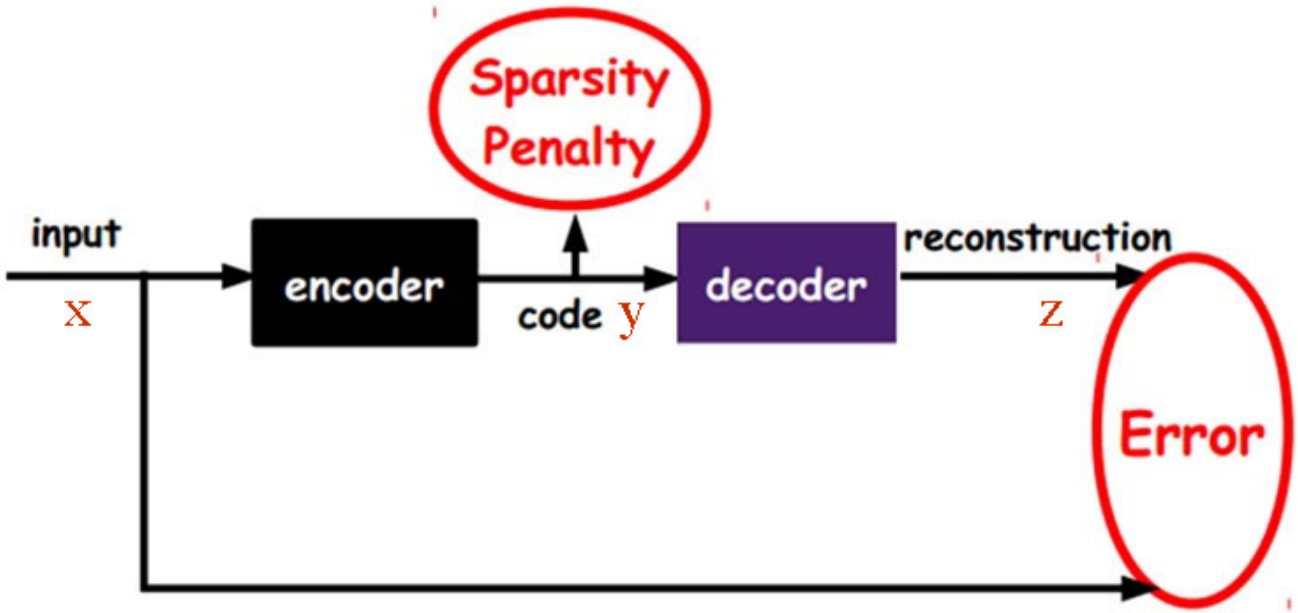
The simplest form of regularization is weight-decay which favors small weights.

$$J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m} J(W,b;x^{(i)},y^{(i)})\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2$$

$$= \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{W,b}(x^{(i)}) - y^{(i)}\right\|^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2$$

Outline

Sparse Auto-Encoders



Sparse Auto-Encoders (Cont')

$$J_{sparse}(\mathbf{W}, \mathbf{b}) = J(\mathbf{W}, \mathbf{b}) + \beta \sum_{j=1}^{s_2} \mathrm{KL}(\rho || \hat{\rho}_j) \tag{9}$$

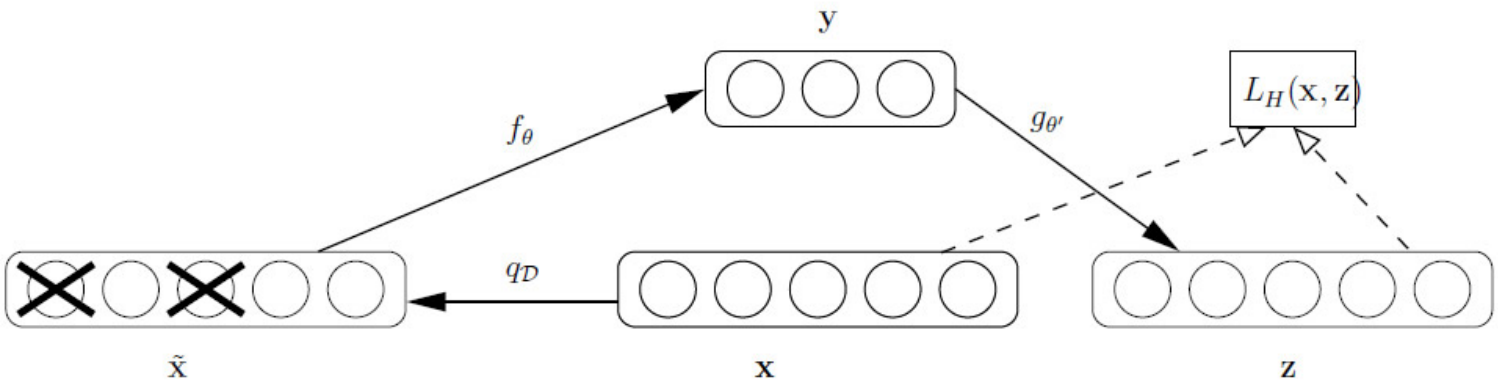
$$\mathrm{KL}(\rho || \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j} \tag{10}$$

where ρ is a sparsity parameter, which specifies our desired level of sparsity, typically a small value close to zero (say $\rho = 0.05$).

$$\hat{\rho}_j = \frac{1}{m} \sum_{i=1}^{m} [a_j^{(2)}(x^{(i)})] \tag{11}$$

If $\hat{\rho}_j$ is close to ρ, the hidden unit's activations must mostly be near 0.

Denoising Auto-Encoders



$\tilde{\mathbf{x}} \sim q_D(\tilde{\mathbf{x}}|\mathbf{x})$, the stochastic corruption process consists in randomly setting some of the inputs (as many of half on them) to zero.
$\mathbf{y} = f_\theta(\tilde{\mathbf{x}}) = s(\mathbf{W}\tilde{\mathbf{x}} + \mathbf{b})$, $\mathbf{z} = g_{\theta'}(\mathbf{y}) = s(\mathbf{W}'\mathbf{y} + \mathbf{b}')$

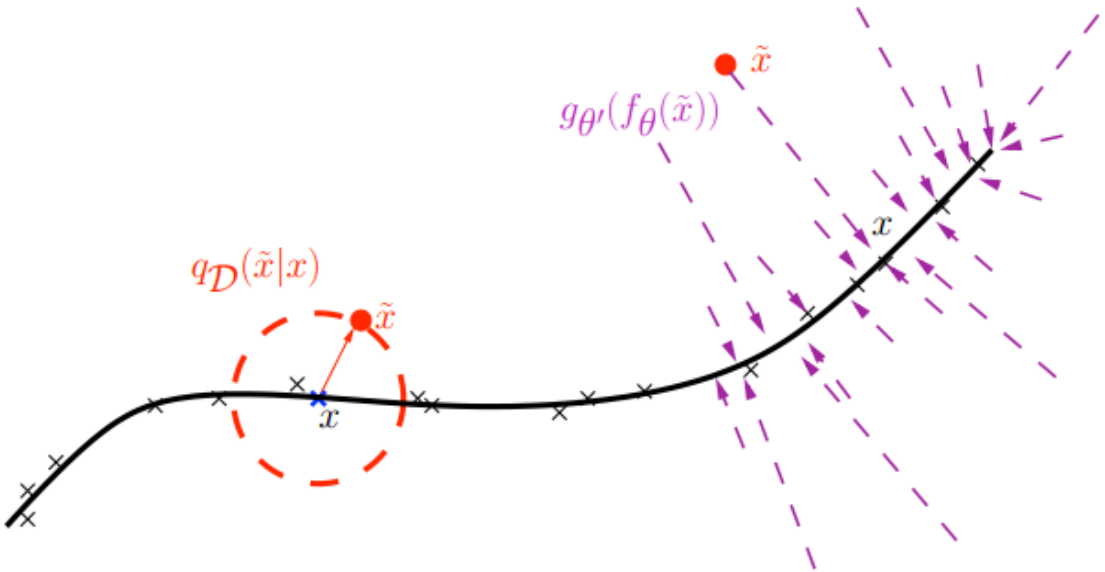
Denoising Auto-Encoders (Cont')

$$q^0(\mathbf{X}, \tilde{\mathbf{X}}) = q^0(\mathbf{X}) q_D(\tilde{\mathbf{X}}|\mathbf{X}) \delta_{f_\theta(\tilde{\mathbf{X}})}(\mathbf{Y}) \tag{12}$$

where $\delta_u(v)$ puts mass 0 when $u \neq v$.

$$\theta^*, \theta'^* = \arg\min_{\theta,\theta'} E_{q^0(\mathbf{X},\tilde{\mathbf{X}})}[L_H(\mathbf{X}, g_{\theta'}(f_\theta(\tilde{\mathbf{X}})))] \tag{13}$$

Manifold learning perspective:



Contractive Auto-Encoders

If input $\mathbf{x}$ is mapped by encoding function f to hidden representation h, this sensitivity penalization term is the sum of squares of all partial derivatives of the extracted features with respect to input dimensions:

$$\|J_f(\mathbf{x})\|_F^2 = \sum_{ij} \left(\frac{\partial h_j(\mathbf{x})}{\partial x_i} \right)^2 \tag{14}$$

which encourages the mapping to the feature space to be contractive in the neighborhood of the training data.

Contractive Auto-Encoders (Cont')

$$J_{CAE}(\theta, \theta') = \sum_{\mathbf{x} \in \mathbf{D}} \left(L(\mathbf{x}, g_{\theta'}(f_\theta(\mathbf{x}))) + \lambda \left\| J_f(\mathbf{x}) \right\|_F^2 \right) \tag{15}$$

In the case of a sigmoid nonlinearity, the penalty on the Jacobian norm has the following simple expression:

$$\left\| J_f(\mathbf{x}) \right\|_F^2 = \sum_{i=1}^{d'} \left(h_i (1 - h_i) \right)^2 \sum_{j=1}^{d} W_{ij}^2 \tag{16}$$

Relationship

- In the case of a linear encoder (i.e. when f is the identity function), CAEs and AEs+wd are identical.
- Sparse Auto-Encoders that output many close-to-zero features, are likely to correspond to a highly contractive mapping.
- Robustness to input perturbations was also one of the motivation of the denoising auto-encoder. CAEs explicitly encourage robustness of representation, whereas DAEs encourages robustness of reconstruction.

Saturating Auto-Encoders

A simple new regularizer for auto-encoders which encourages activations in the saturated regions of the corresponding activation function.
The auto-encoder function is defined as:

$$G(\mathbf{x}, \mathbf{W}) = \mathbf{W}' F(\mathbf{W}\mathbf{x} + \mathbf{b}) + \mathbf{b}' \tag{17}$$

$$L = \sum_{\mathbf{x} \in \mathbf{D}} \frac{1}{2} \left\| \mathbf{x} - (\mathbf{W}' F(\mathbf{W}\mathbf{x} + \mathbf{b}) + \mathbf{b}') \right\|^2 + \lambda \sum_{i=1}^{d'} f_c(\mathbf{W}_i \mathbf{x} + \mathbf{b}_i) \tag{18}$$

where F is the vector function that applies the scalar function f to each of its components. f will be designed to have the saturation regions.

$$f_c(z) = \inf_{z' \in \{z | f'(z) = 0\}} \left| z - z' \right| \tag{19}$$

$f_c(z)$ corresponds to the distance of z to one of the flat spots of $f(z)$.

Relationship

- The following equation adjusts the weights so as to push the activations into the low gradient (saturation) regions. CAEs indirectly encourage operation in the saturation regions.

$$\sum_{ij}\left(\frac{\partial h_i}{\partial x_j}\right)^2 = \sum_{i=1}^{d'}\left(f'(\sum_{j=1}^{d} W_{ij}x_j + b_i)^2 \left\|\mathbf{W}_i\right\|^2\right) \tag{20}$$

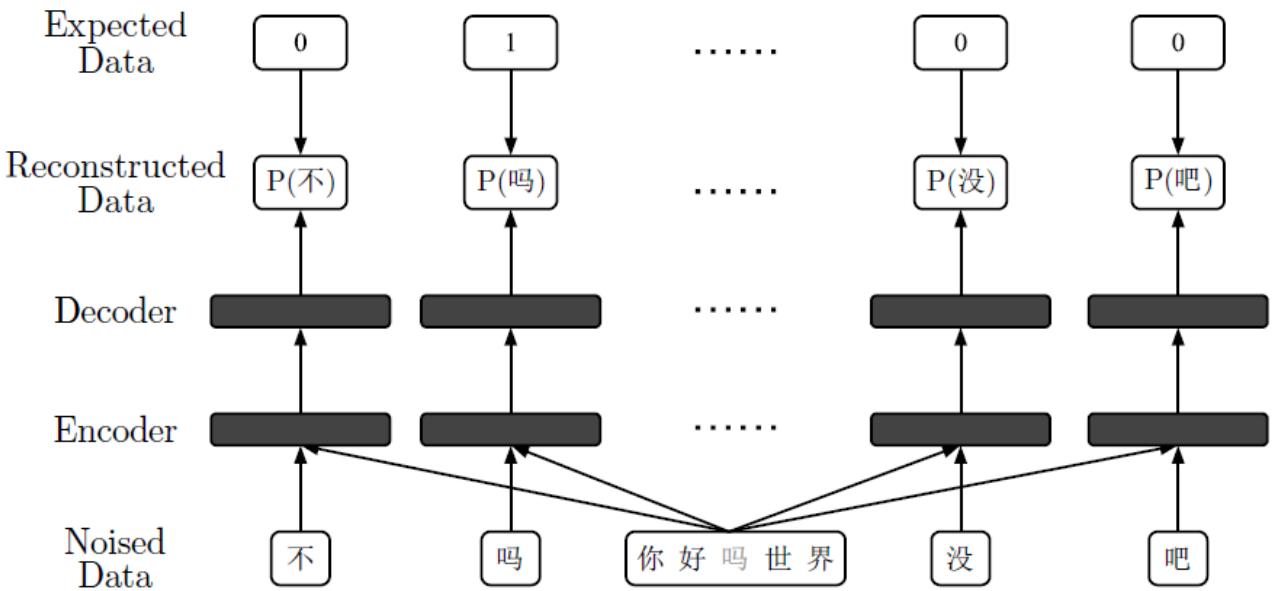
- The shrink function is particularly compatible with ℓ_1 minimization. SATAEs are a generalization of sparse auto-encoders.

$$shrink_c(x) = \begin{cases} abs(x) & |x| > \tau \\ 0 & \text{elsewhere} \end{cases} \tag{21}$$

Outline

Chinese Word Segmentation

Text Window Denoising Autoencoder: Building Deep Architecture for Chinese Word Segmentation (Wu et al., NLP&CC 2013)



中文词汇特征表示

基于自动编码器的中文词汇特征无监督学习(张开旭等, 中文信息学报 2013)

1 寻找有特异性的上下文词对

去苏州开会中，**苏州**所匹配的上下文词对为**<去,开会>**，可匹配地点词
本文自动找到1346个上下文词对，例如

<去,参加> <飞往,,> <届,国际> <从,引进> <在,出差> <在,警方>
<那么,。> <蛮,的> <非常,!> <很,地> <挺,的> <很,很>
<我们,了> 他们,了> <没有,过> <次,了> <地,了> <已,了> <上,了>
<名,说> <位,表示> <位,向> <位,认为> <位,这样>

2 用上下文词对在大规模语料中匹配目标词汇

使用清洗过的SogouT数据集，包含约260亿字符的文本。
使用THULAC工具进行快速分词，同时参考最优和部分次优分词结果
共得到52876个多字词的匹配结果

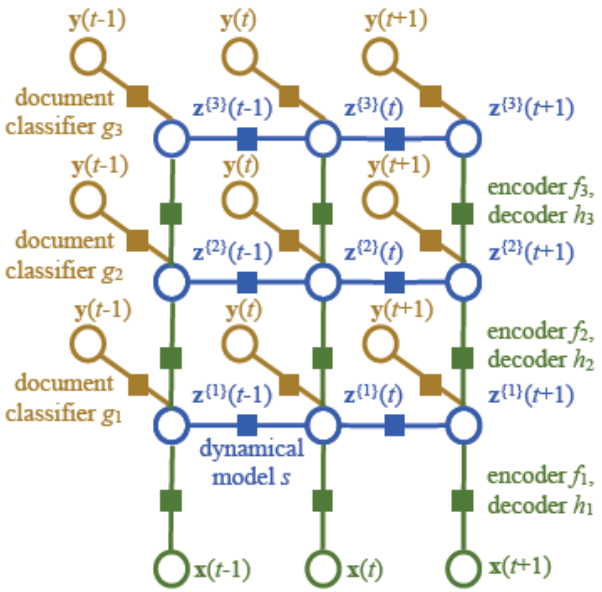
3 使用自动编码器进行降维、离散化、稀疏化，得到目标词汇的特征表示

自动编码器是一种神经网络，可将高维数匹配数据降维成50维稀疏01向量。对于每一个词，最终得到一个非零分量下标集合，表示其句法语义性质，例如

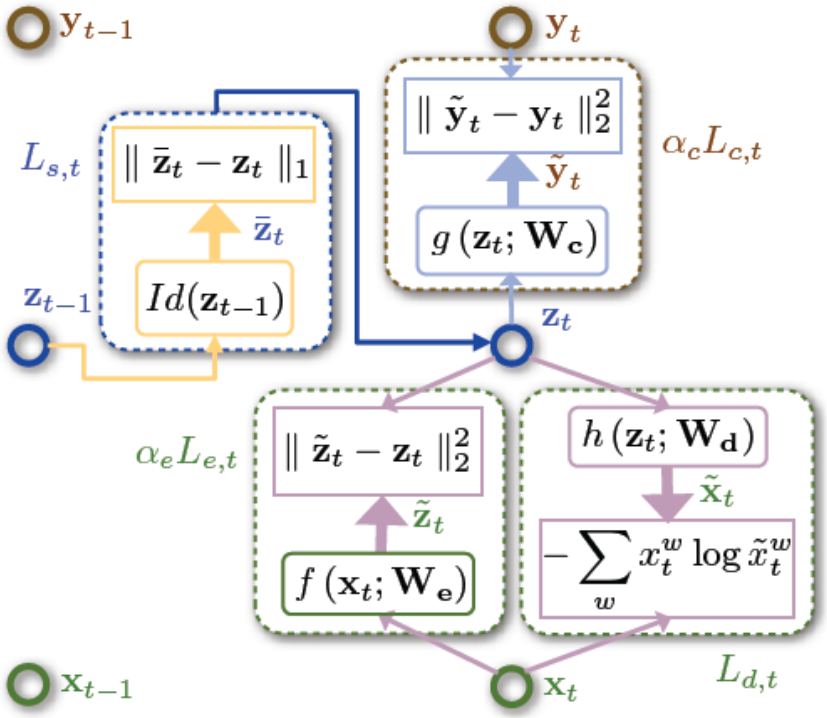
学生	1	4	5	7		11	12	14	15		17		21				37			46		
老师	1	4	5			11	12		15	16		19	21						45	46		49
学校	1		5	7		11	12											44		46	48	
公司	1		5	7	10	11	12		15					26	31	36		44		46	48	

Semantic Indexing

Dynamic Auto-Encoders for Semantic Indexing (Mirowski et al., NIPS Workshop 2010)



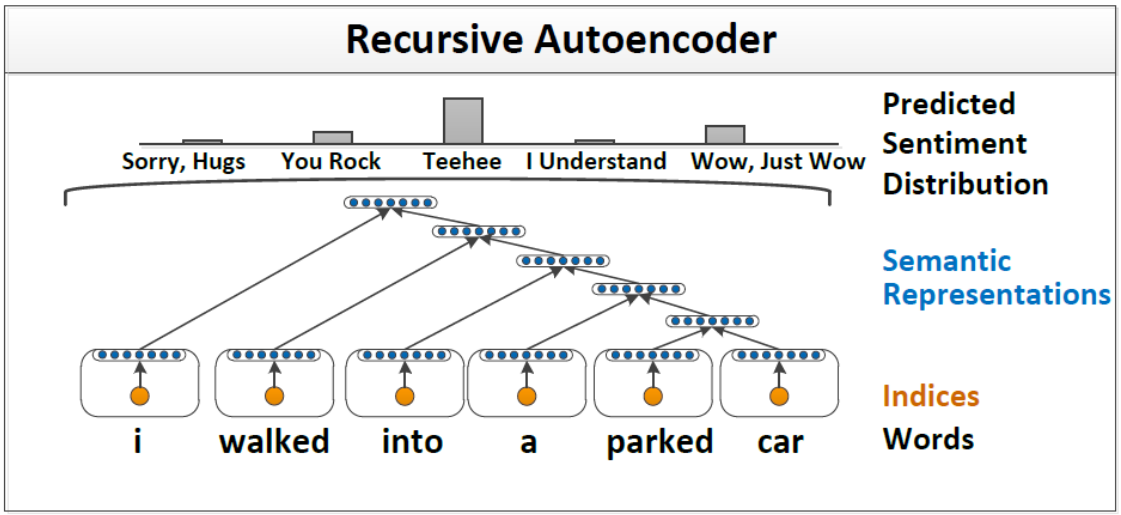
(a) 3-layer deep auto-encoder

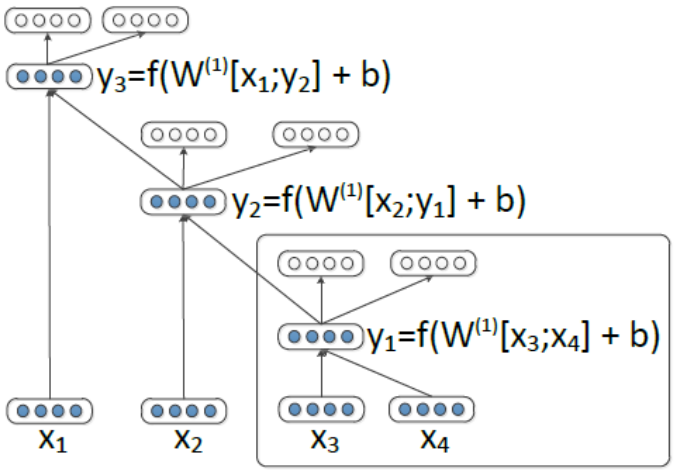


(b) Layer 1

Sentiment Analysis

Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions (Socher et al., EMNLP 2011)





Thank you !

Q & A?

Email:tjzhifei@163.com
Weibo:@同济张_志飞