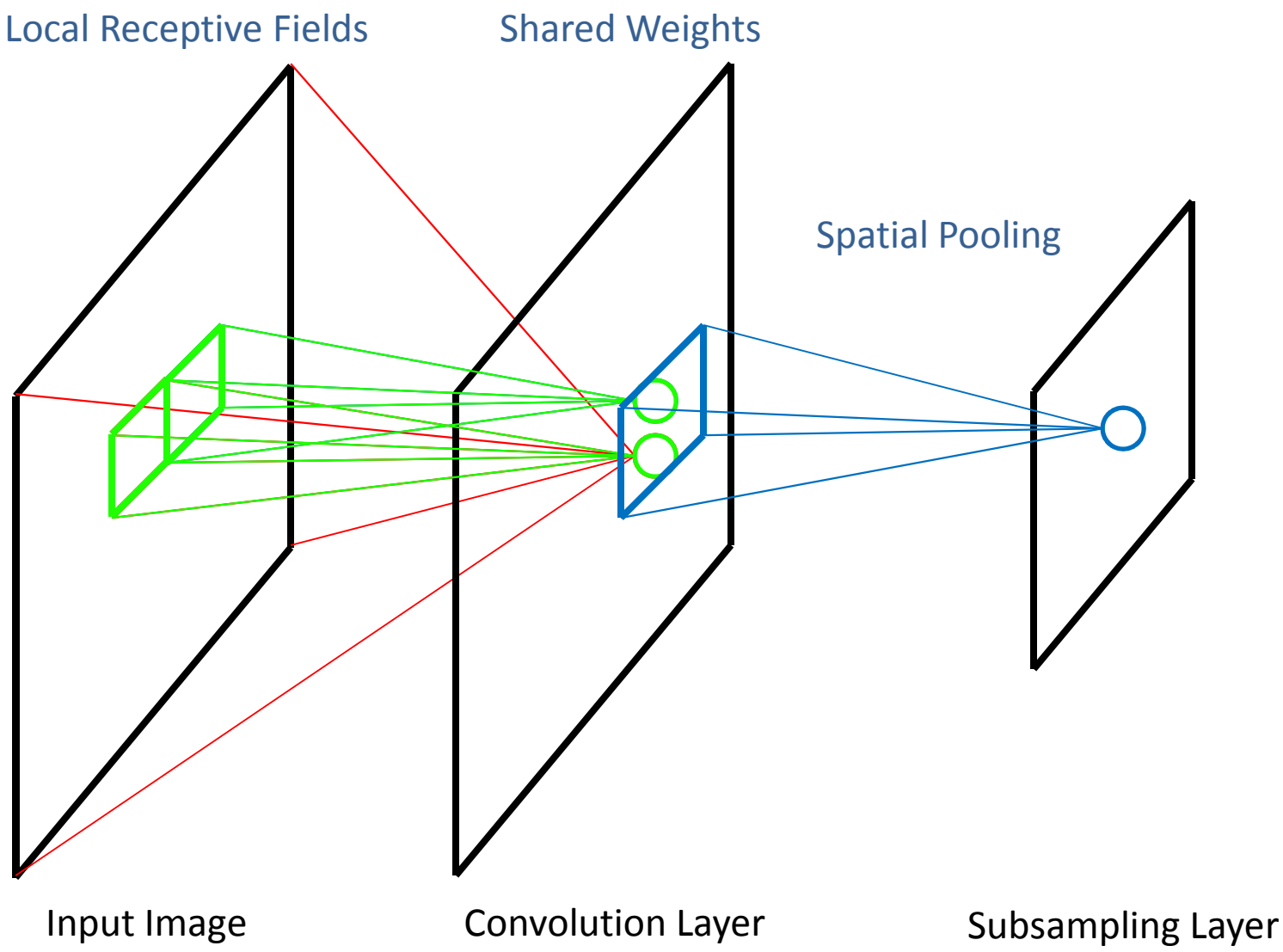
Outline

- Deep learning
  - Greedy layer-wise training (for supervised learning)
  - Deep belief nets
  - Stacked denoising auto-encoders
  - Stacked predictive sparse coding
  - Deep Boltzmann machines
- Applications
  - Vision
  - Audio
  - Language

Applications

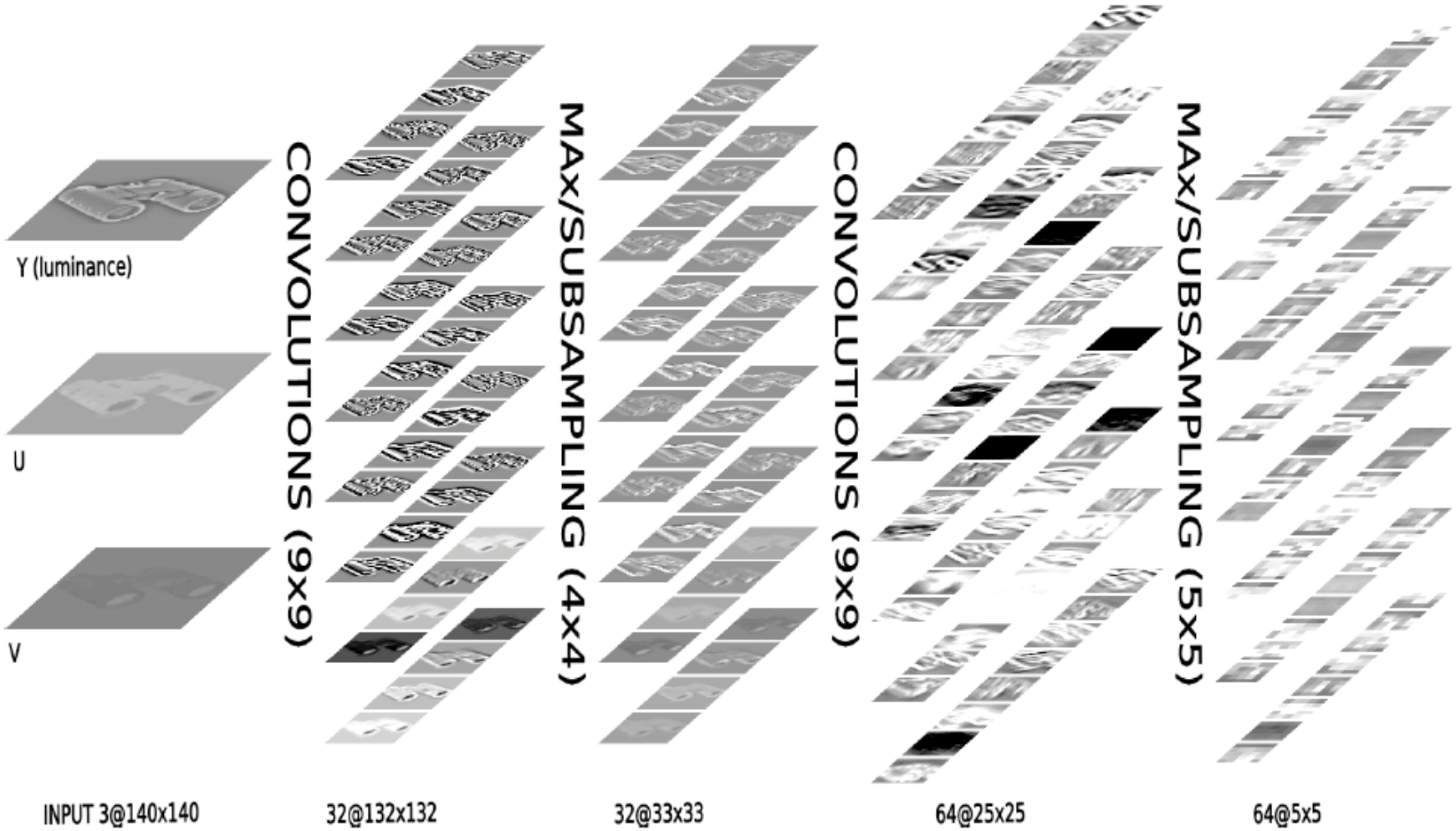
- Cannot always apply one of our deep learning methods to get best results (yet).
  - Scale (e.g., 32x32 RGB image = 3072 inputs)
  - Domain knowledge (e.g., invariance)
  - Other design choices (e.g., nonlinearities)
  - Tuning (e.g., penalties, optimization algorithm)

Convolutional Neural Networks



Deep Convolutional Architectures

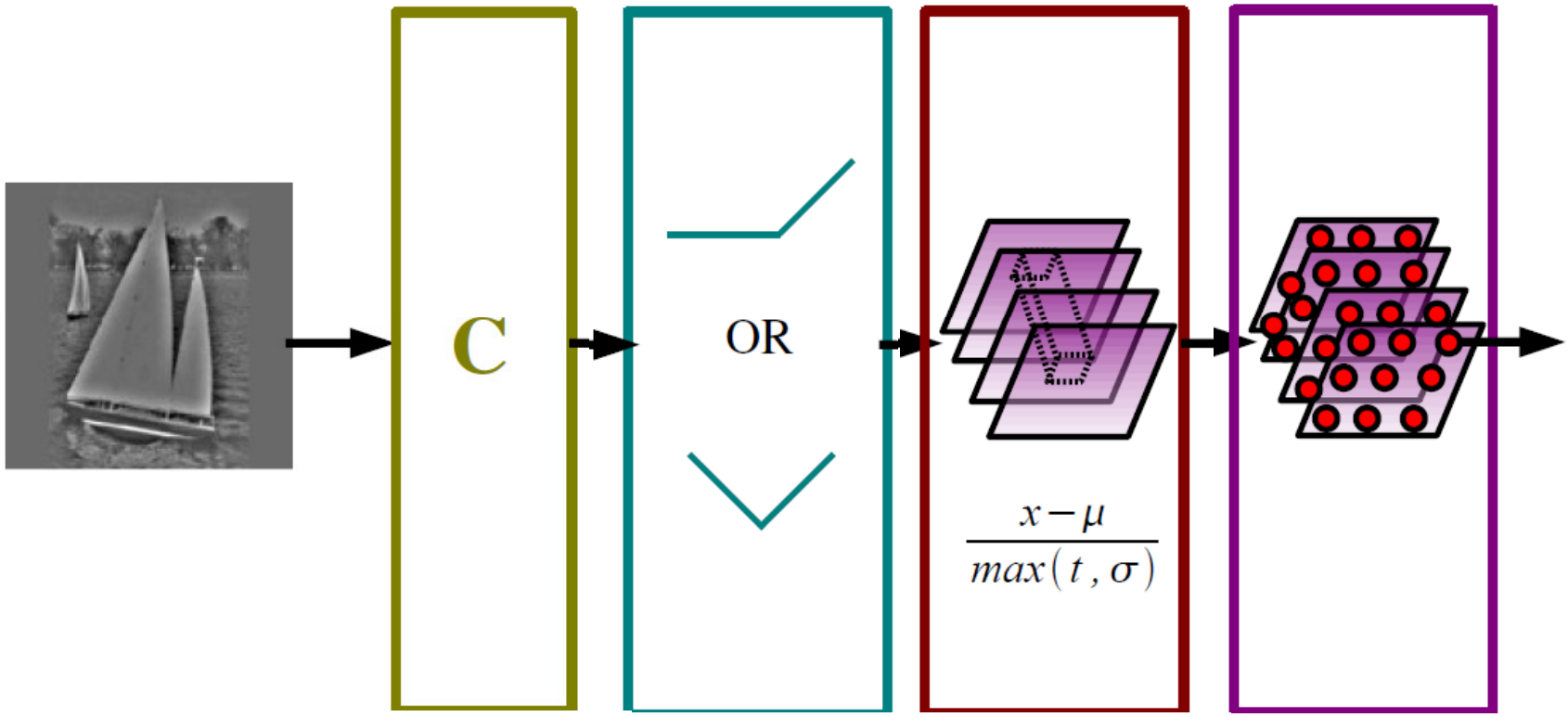
Many state-of-the-art results. (E.g., 0.27% on MNIST by Ciresan et. al, 2011)



Nonlinearities and pooling

(Jarrett et al., 2009)

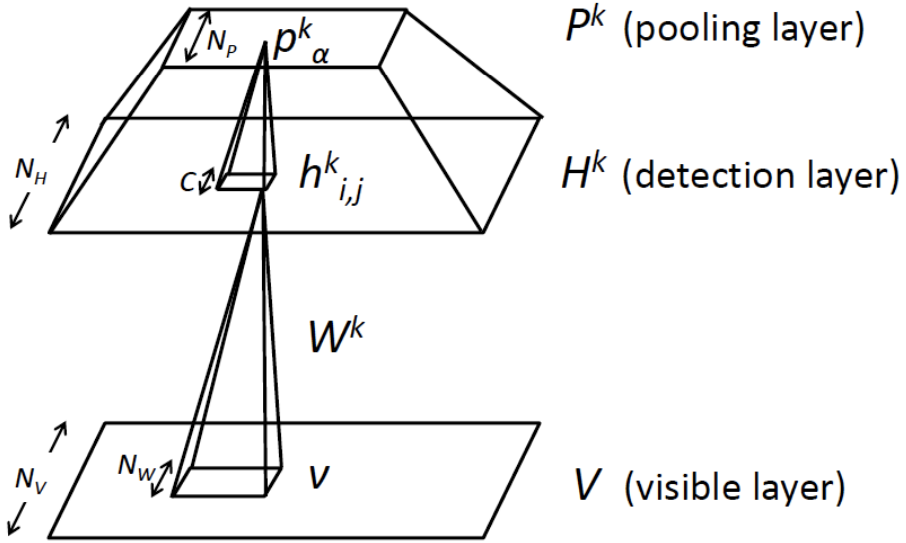
- Details of feature processing stages matter.



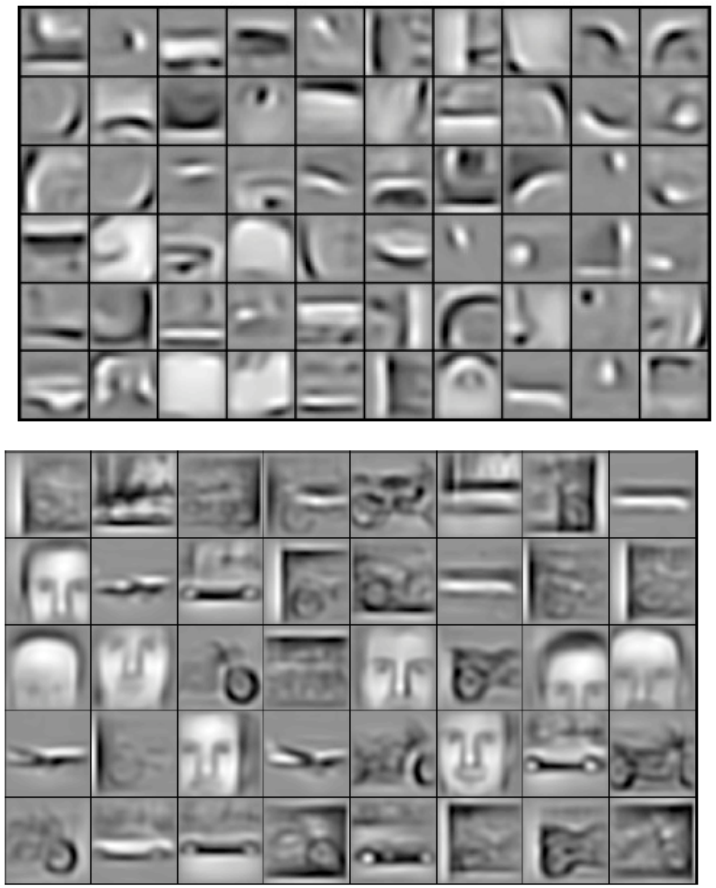
(Scherer et al., 2010; Boureau et. al 2010; Coates et al., 2011)

Convolutional DBNs

Convolutional RBM: Generative training of convolutional structures (with probabilistic max-pooling)



faces, cars, airplanes, motorbikes

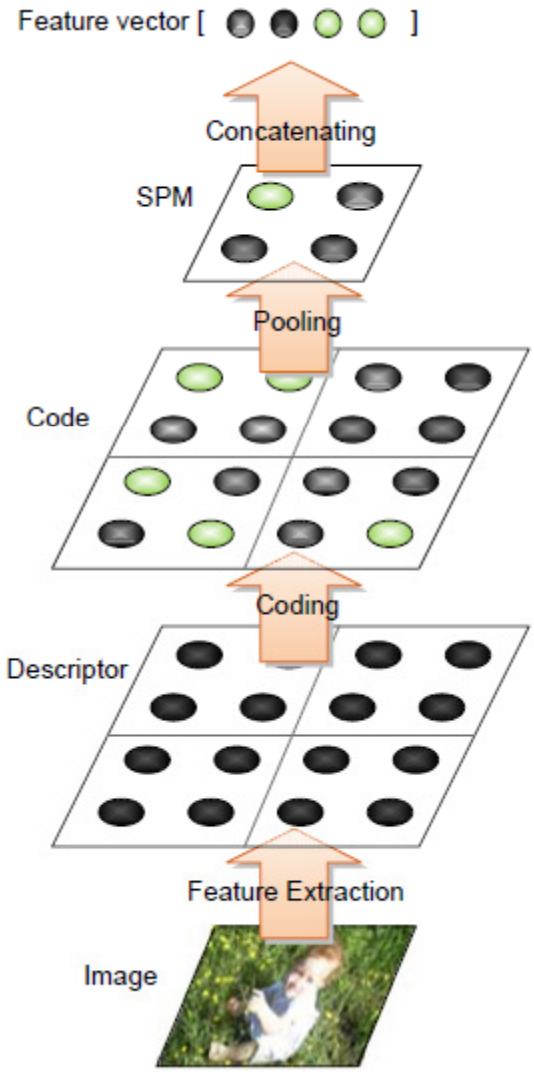


(Lee et al, 2009; Desjardins and Bengio, 2008; Norouzi et al., 2009; Masci et al., 2011)

Spatial Pyramid Structure

(Yang et al., 2009)

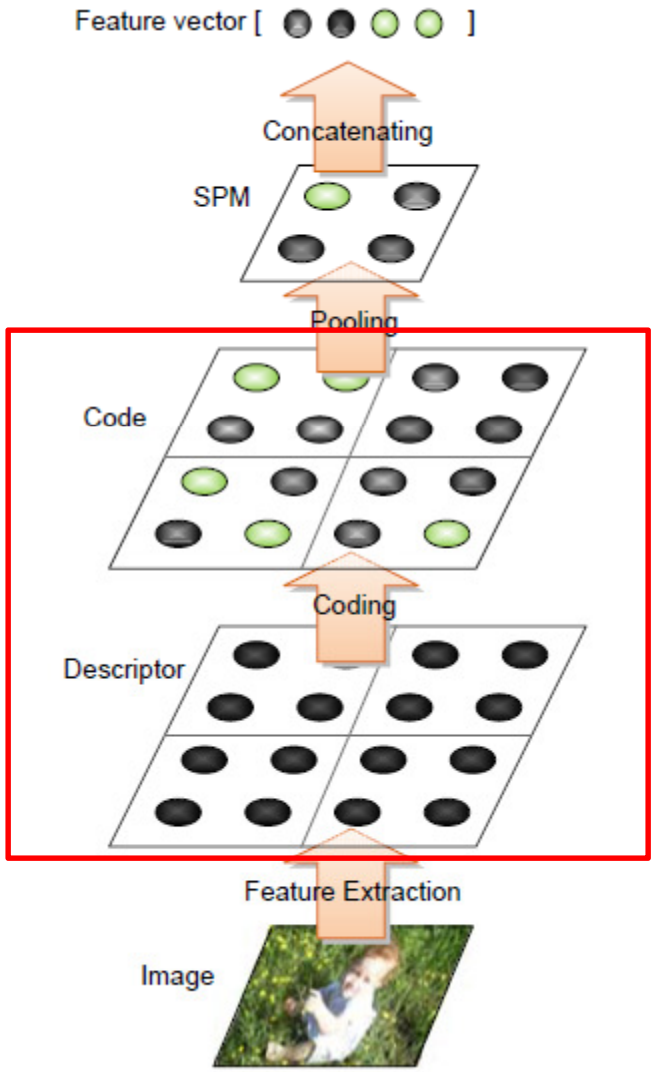
- **Descriptor Layer:** detect and locate features, extract corresponding descriptors (e.g. SIFT)
- **Code Layer:** code the descriptors
  - Vector Quantization (VQ): each code has only one non-zero element
  - Soft-VQ: small group of elements can be non-zero
- **SPM layer:** pool codes across subregions and average/normalize into a histogram



Improving the coding step

(Yang et al., 2009)

- Modify the Coding step to produce better feature representations
  - Sparse coding
    - Sparse like VQ
    - Multiple non-zeros (more expressive)
  - Local Coordinate coding

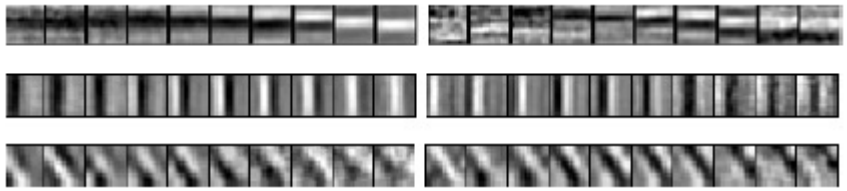


Experimental results

- Competitive performance to other state-of-the-art methods using a single type of features on object recognition benchmarks
- E.g.: Caltech 101 (30 examples per class)
  - Using pixel representation: ~65% accuracy (Jarret et al., 2009; Lee et al., 2009; and many others)
  - Using SIFT representation: 73~77% accuracy (Yang et al., 2009; Jarret et al., 2009, Boureau et al., 2011, and many others)

Video?

- Extend your favorite method to spatio-temporal receptive fields:



(Le et al., 2011)

	KTH	Hollywood2	UCF	YouTube
Best Prior	92.1%	50.9%	85.6%	71.2%
Le et al. 2011	**93.9%**	**53.3%**	**86.5%**	**75.8%**

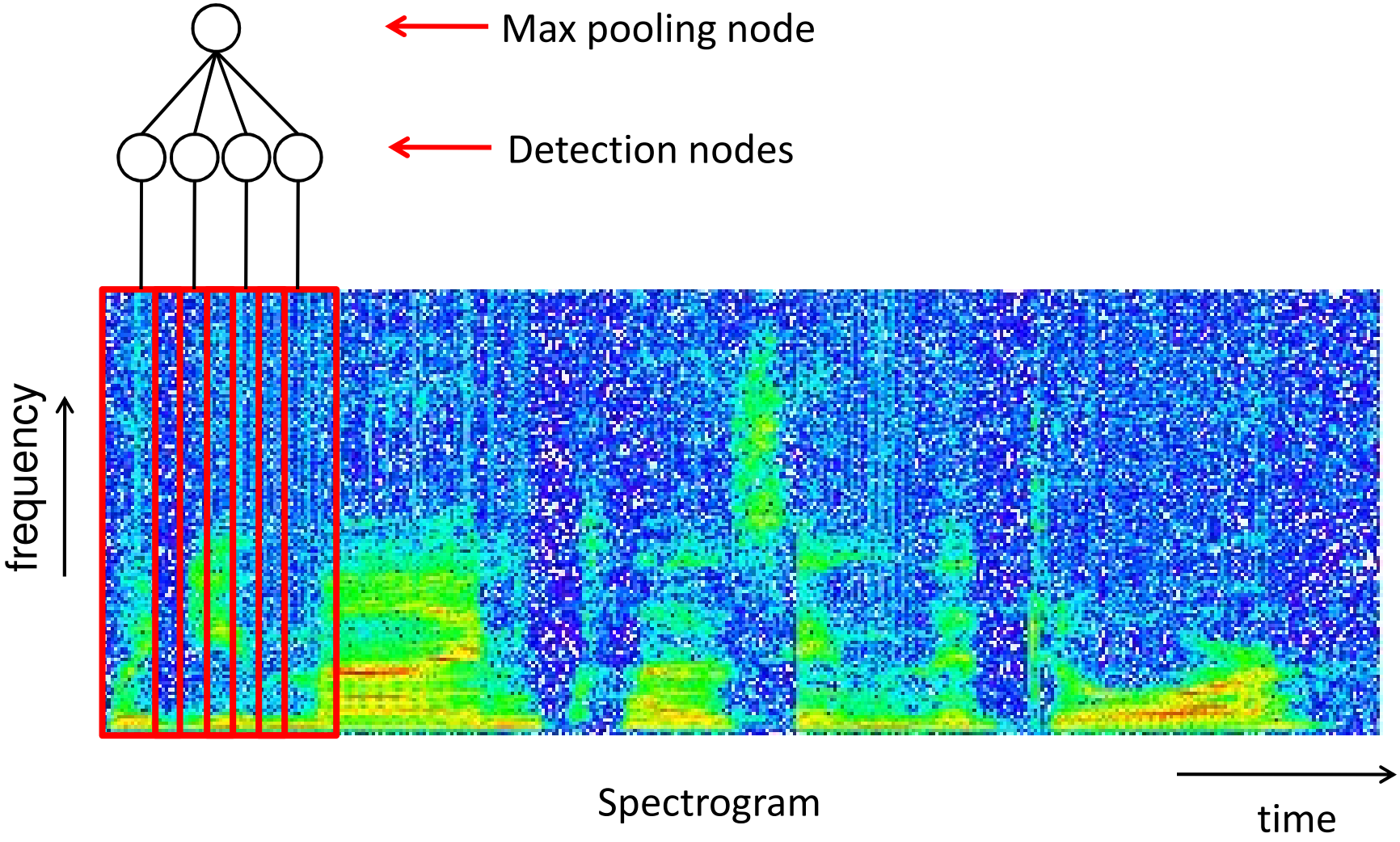
Also: Taylor et al., 2010

Outline

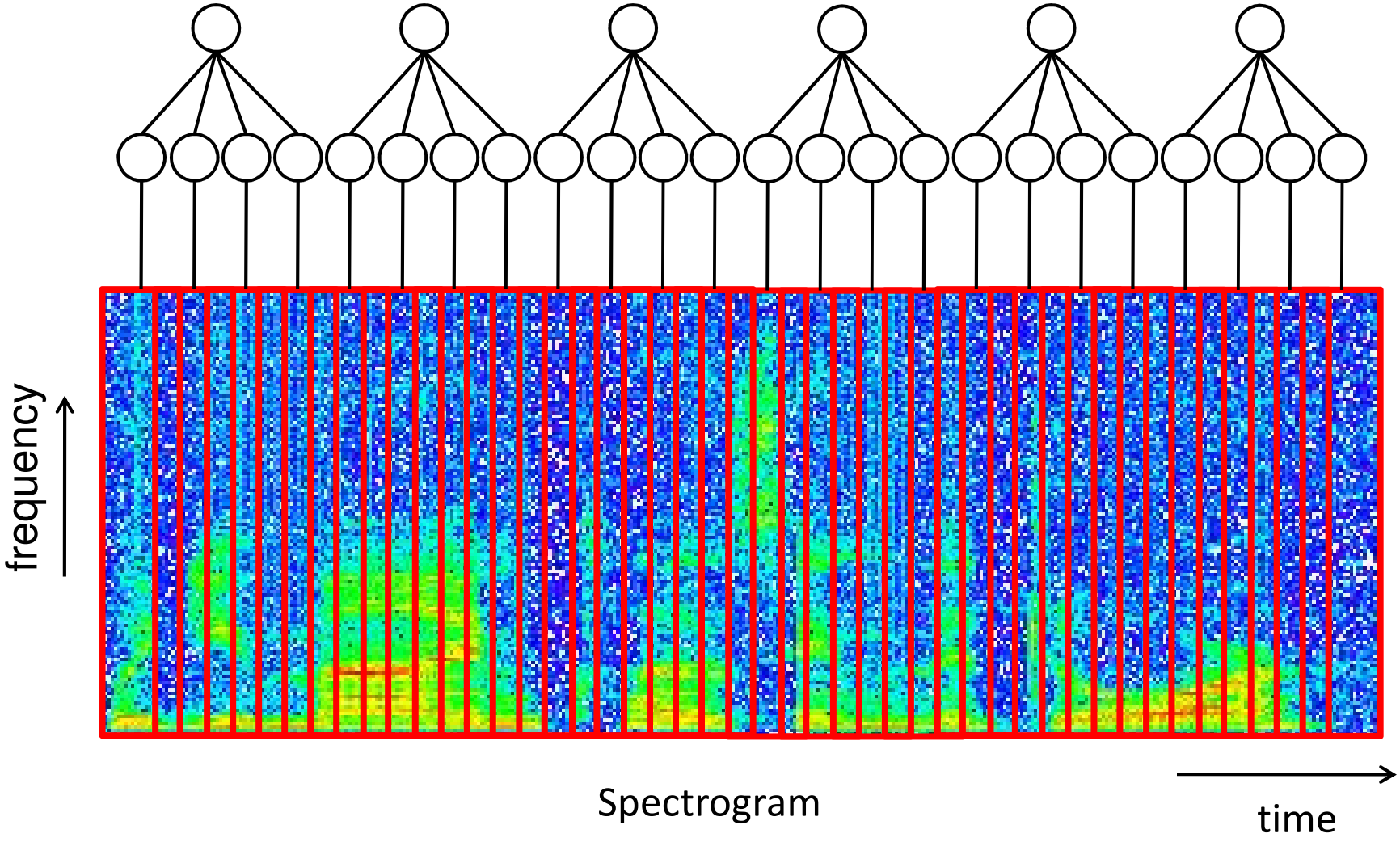
- Deep learning
  - Greedy layer-wise training (for supervised learning)
  - Deep belief nets
  - Stacked denoising auto-encoders
  - Stacked predictive sparse coding
  - Deep Boltzmann machines
- Applications
  - Vision
  - Audio
  - Language

Convolutional DBN for audio

(Lee et al., 2009)

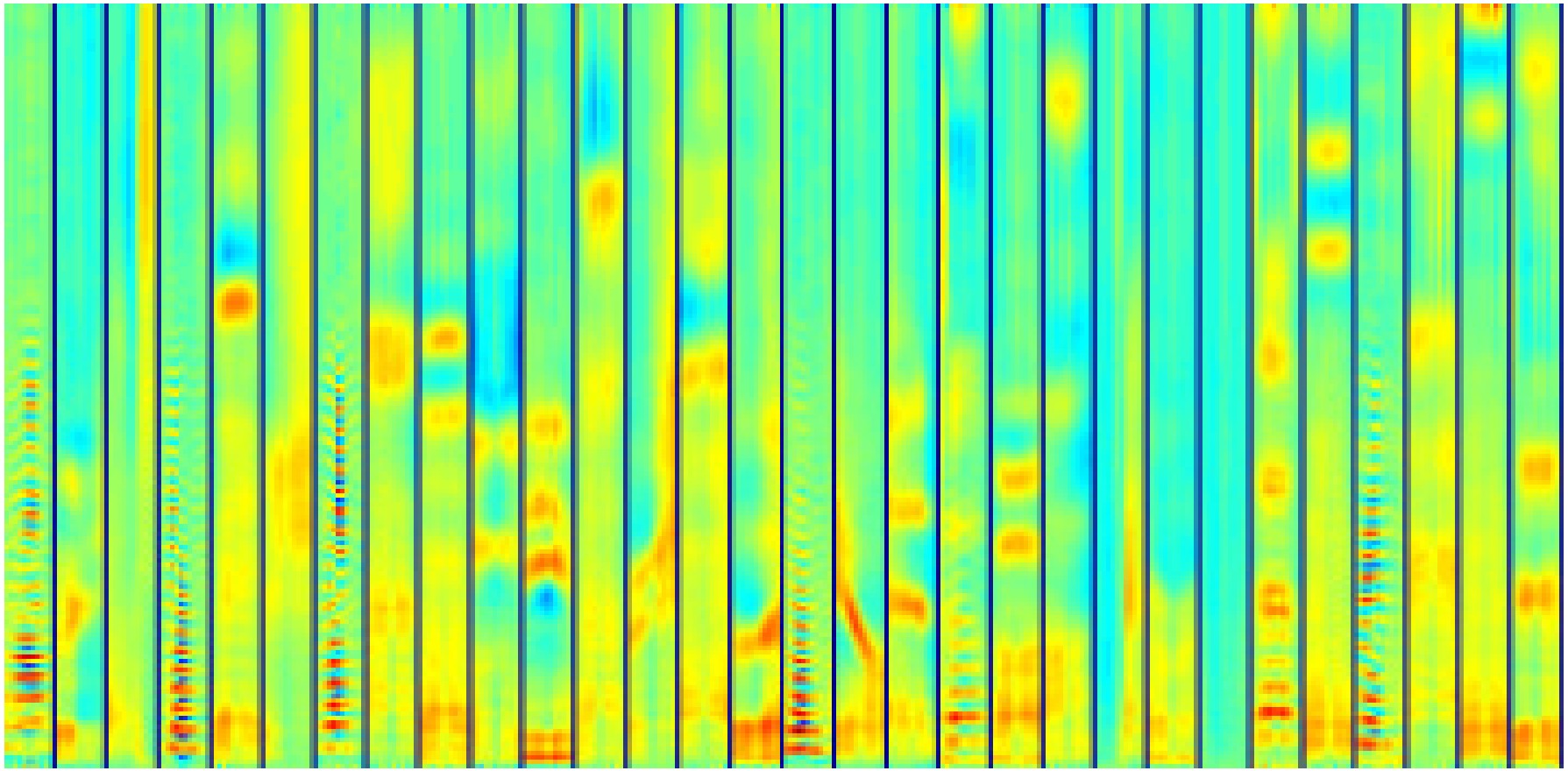


Convolutional DBN for audio



CDBNs for speech

Trained on unlabeled TIMIT corpus



Learned first-layer bases

Experimental Results

(Lee et al., 2009)

- Speaker identification

TIMIT Speaker identification	Accuracy
Prior art (Reynolds, 1995)	99.7%
Convolutional DBN	**100.0%**

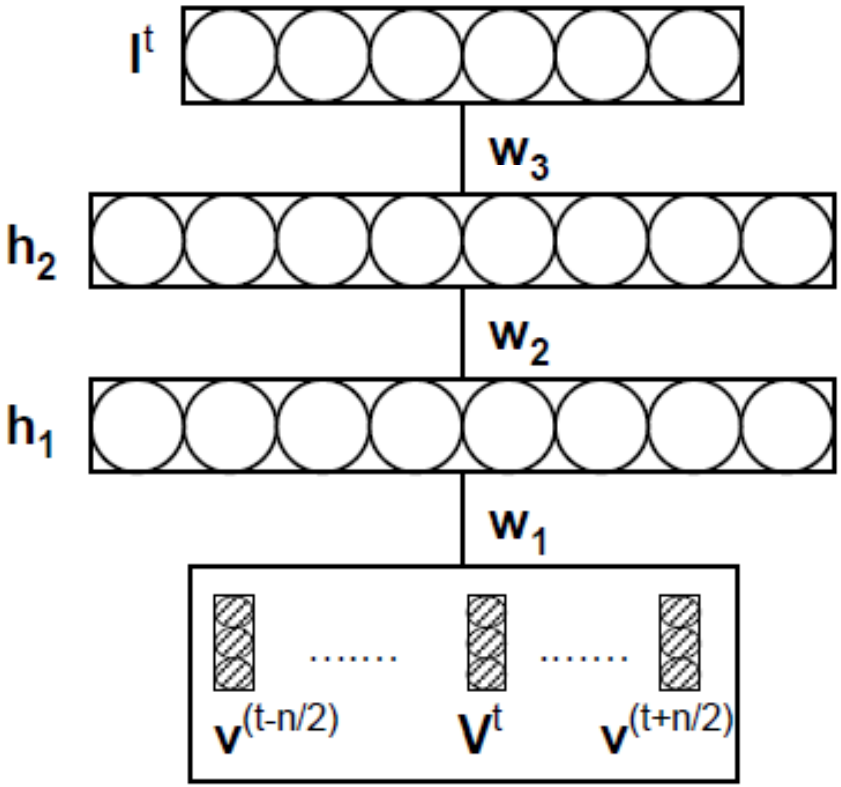
- Phone classification

TIMIT Phone classification	Accuracy
Clarkson et al. (1999)	77.6%
Gunawardana et al. (2005)	78.3%
Sung et al. (2007)	78.5%
Petrov et al. (2007)	78.6%
Sha & Saul (2006)	78.9%
Yu et al. (2009)	79.2%
Convolutional DBN	**80.3%**

Phone recognition using DBNs

(Dahl et al., 2010; Mohamed et al., 2011)

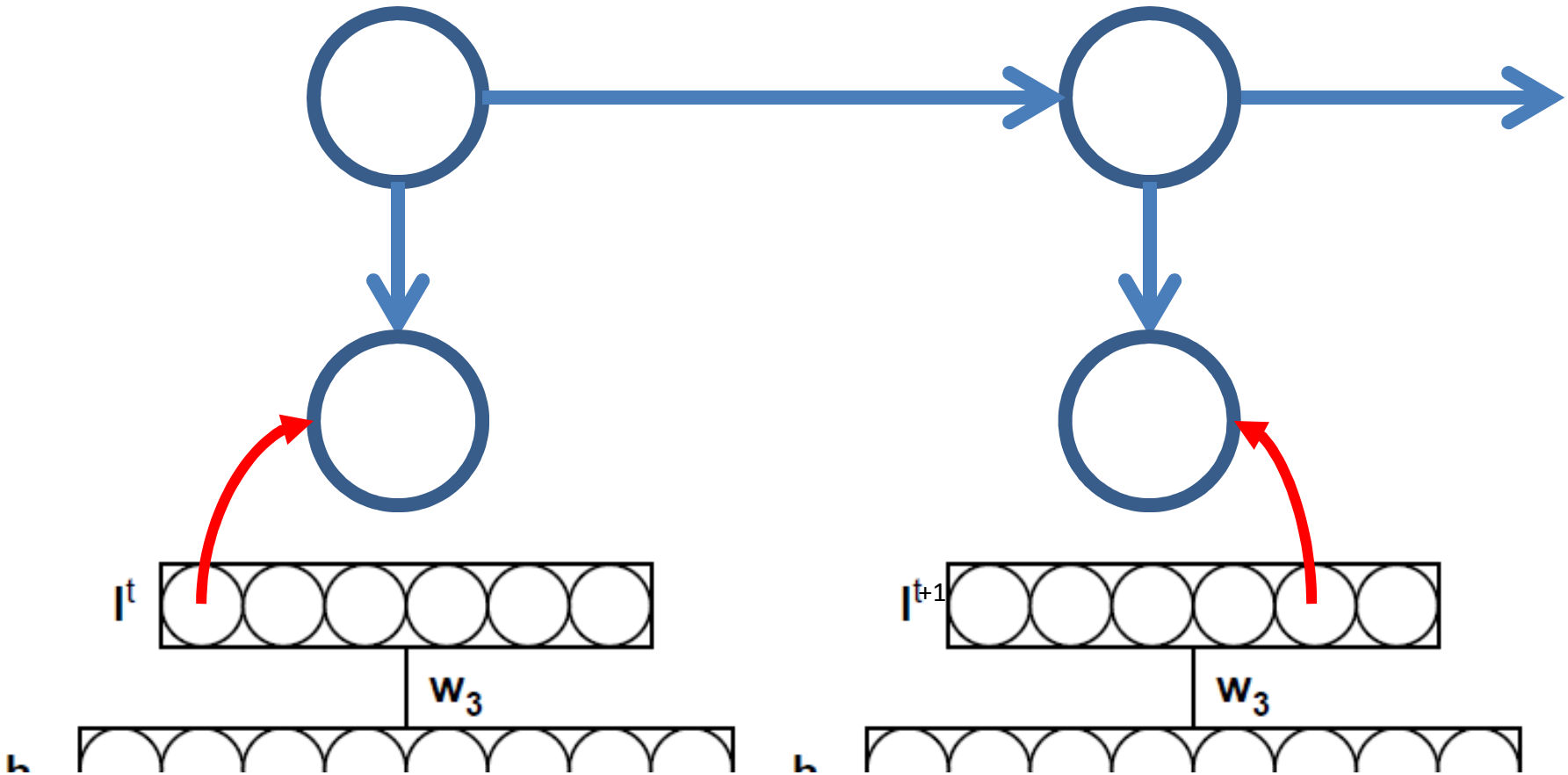
- Pre-training RBMs followed by fine-tuning with back propagation



Phone recognition using DBNs

(Dahl et al., 2010; Mohamed et al., 2011)

- Can use predicted labels to recognize phones in stream of audio with HMM.



Phone recognition results

(Dahl et al., 2010)

Method	PER
Stochastic Segmental Models	36.0%
Conditional Random Field	34.8%
Large-Margin GMM	33.0%
CD-HMM	27.3%
Augmented conditional Random Fields	26.6%
Recurrent Neural Nets	26.1%
Bayesian Triphone HMM	25.6%
Monophone HTMs	24.8%
Heterogeneous Classifiers	24.4%
Deep Belief Networks(DBNs)	**23.0%**
Triphone HMMs discriminatively trained w/ BMMI	22.7%
Deep Belief Networks (with mcRBM feature extraction)	**20.5%**

Outline

- Deep learning
  - Greedy layer-wise training (for supervised learning)
  - Deep belief nets
  - Stacked denoising auto-encoders
  - Stacked predictive sparse coding
  - Deep Boltzmann machines
- Applications
  - Vision
  - Audio
  - Language

Language modeling

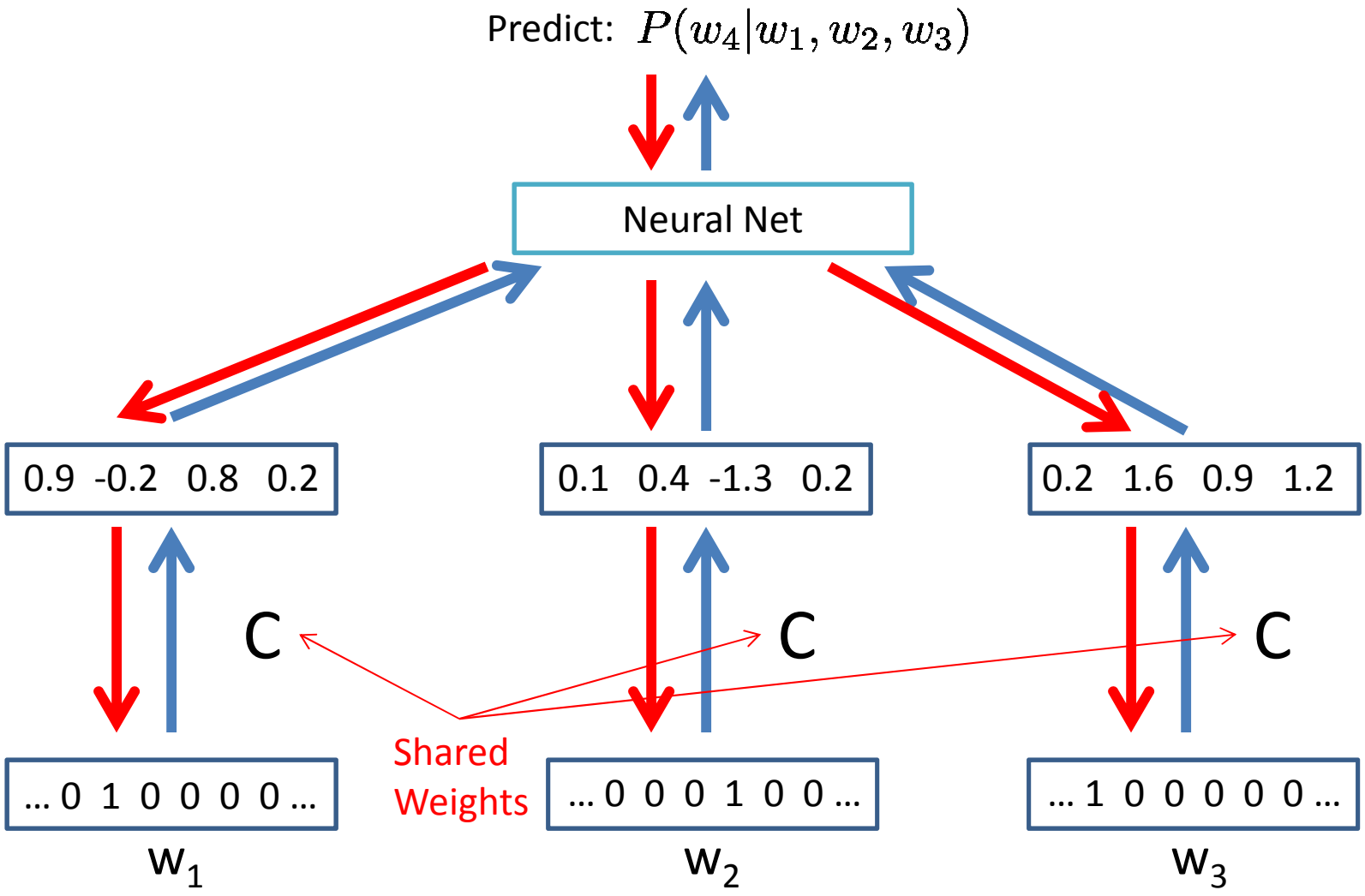
- Language Models
  - Estimating the probability of the next word w given a sequence of words
- Baseline approach in NLP
  - N-gram models (with smoothing):

$$P(w_i|w_{i-(n-1)},\ldots,w_{i-1}) = \frac{count(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{count(w_{i-(n-1)},\ldots,w_{i-1})}$$

- Deep Learning approach
  - Bengio et al. (2000, 2003): via Neural network
  - Mnih and Hinton (2007): via RBMs

Predicting the next word

Bengio et al. (2000, 2003)

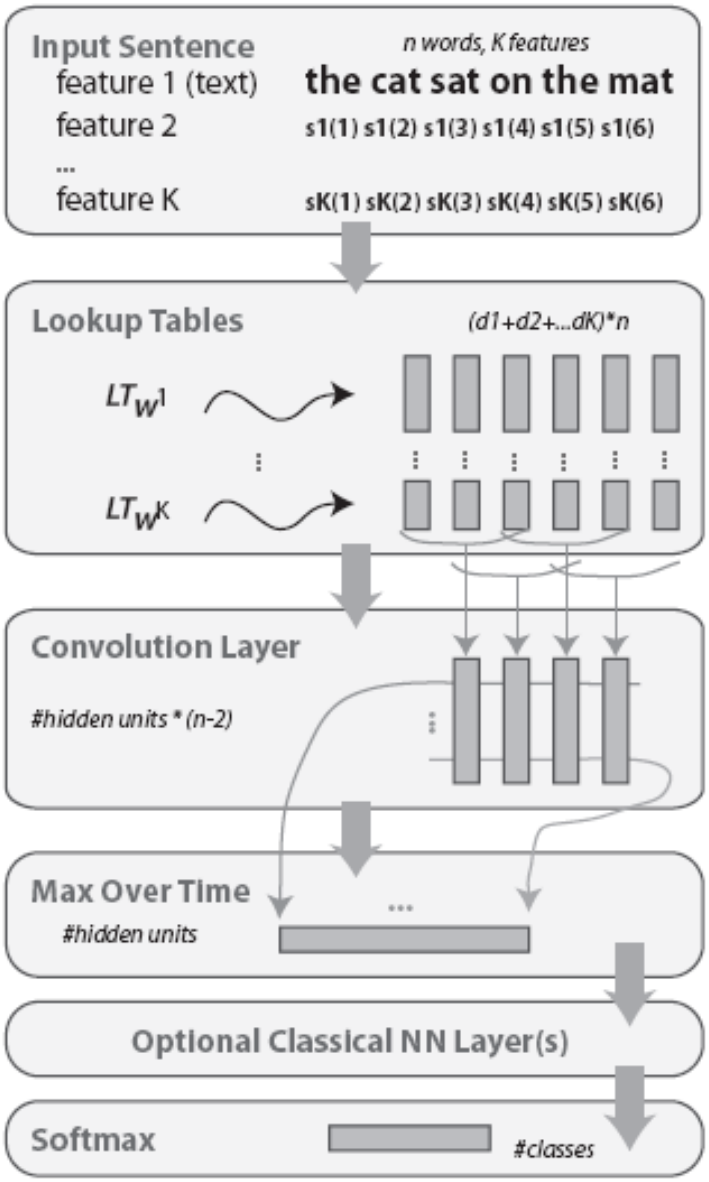


A unified architecture for NLP

(Collobert and Weston, 2009)

- Main idea: use as unified architecture for NLP
  - Deep Neural Network
  - Trained jointly with different tasks (feature sharing and multi-task learning)
- Shows the generality of the architecture

General Deep Architecture for NLP



(Collobert and Weston, 2009)

Basic features (e.g., word, capitalization, relative position)

Embedding by lookup table

Convolution (higher-level features that fold in context)

Max pooling

Supervised learning

Other NLP tasks

(Collobert and Weston, 2009)

- Part-Of-Speech Tagging (POS)
  - mark up the words in a text (corpus) as corresponding to a particular tag
    - E.g. Noun, adverb, ...
- Chunking
  - Also called shallow parsing
  - In the view of phrase: Labeling phrase to syntactic constituents
    - E.g. NP (noun phrase), VP (verb phrase), …
  - In the view of word: Labeling word to syntactic role in a phrase
    - E.g. B-NP (beginning of NP), I-VP (inside VP), …

Other NLP tasks

(Collobert and Weston, 2009)

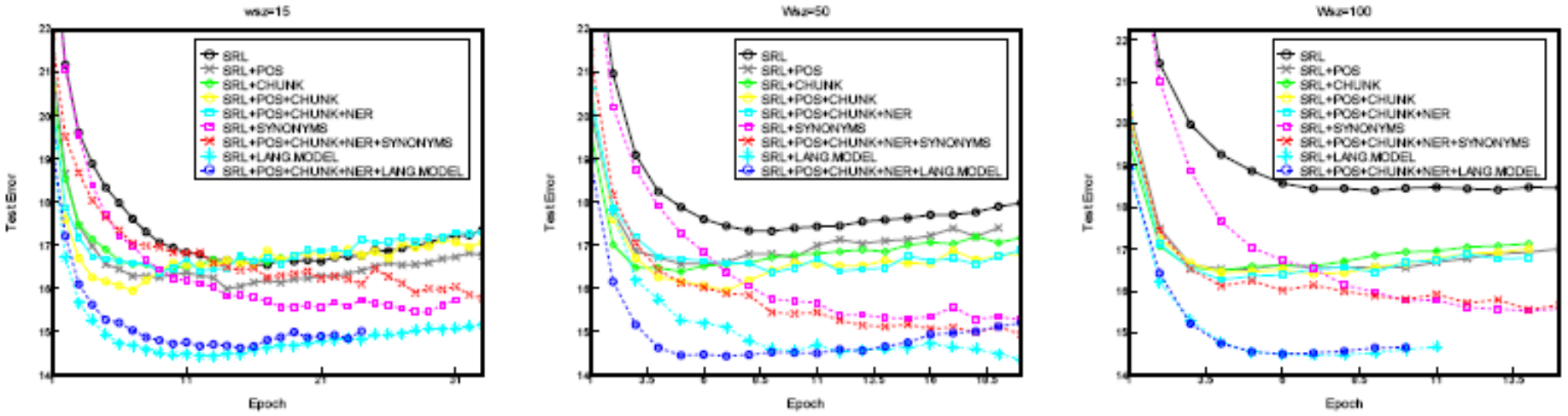
- Named Entity Recognition (NER)
  - In the view of thought group: Given a stream of text, determine which items in the text map to proper names
  - E.g., labeling "atomic elements" into "PERSON", "COMPANY", "LOCATION"
- Semantic Role Labeling (SRL)
  - In the view of sentence: giving a semantic role to a syntactic constituent of a sentence
  - E.g. $[\text{John}]_{\text{ARG0}}$ $[\text{ate}]_{\text{REL}}$ $[\text{the apple}]_{\text{ARG1}}$ (Proposition Bank)
    - An Annotated Corpus of Semantic Roles (Palmer et al.)

Results

(Collobert and Weston, 2009)

- MTL improves SRL's performance

	wsz=15	wsz=50	wsz=100
SRL	16.54	17.33	18.40
SRL + POS	15.99	16.57	16.53
SRL + Chunking	16.42	16.39	16.48
SRL + NER	16.67	17.29	17.21
SRL + Synonyms	15.46	15.17	15.17
SRL + Language model	14.42	14.30	14.46
SRL + POS + Chunking	16.46	15.95	16.41
SRL + POS + NER	16.45	16.89	16.29
SRL + POS + Chunking + NER	16.33	16.36	16.27
SRL + POS + Chunking + NER + Synonyms	15.71	14.76	15.48
SRL + POS + Chunking + NER + Language model	14.63	14.44	14.50



Summary

- Training deep architectures
  - Unsupervised pre-training helps training deep networks
  - Deep belief nets, Stacked denoising auto-encoders, Stacked predictive sparse coding, Deep Boltzmann machines
- Deep learning algorithms and unsupervised feature learning algorithms show promising results in many applications
  - vision, audio, natural language processing, etc.