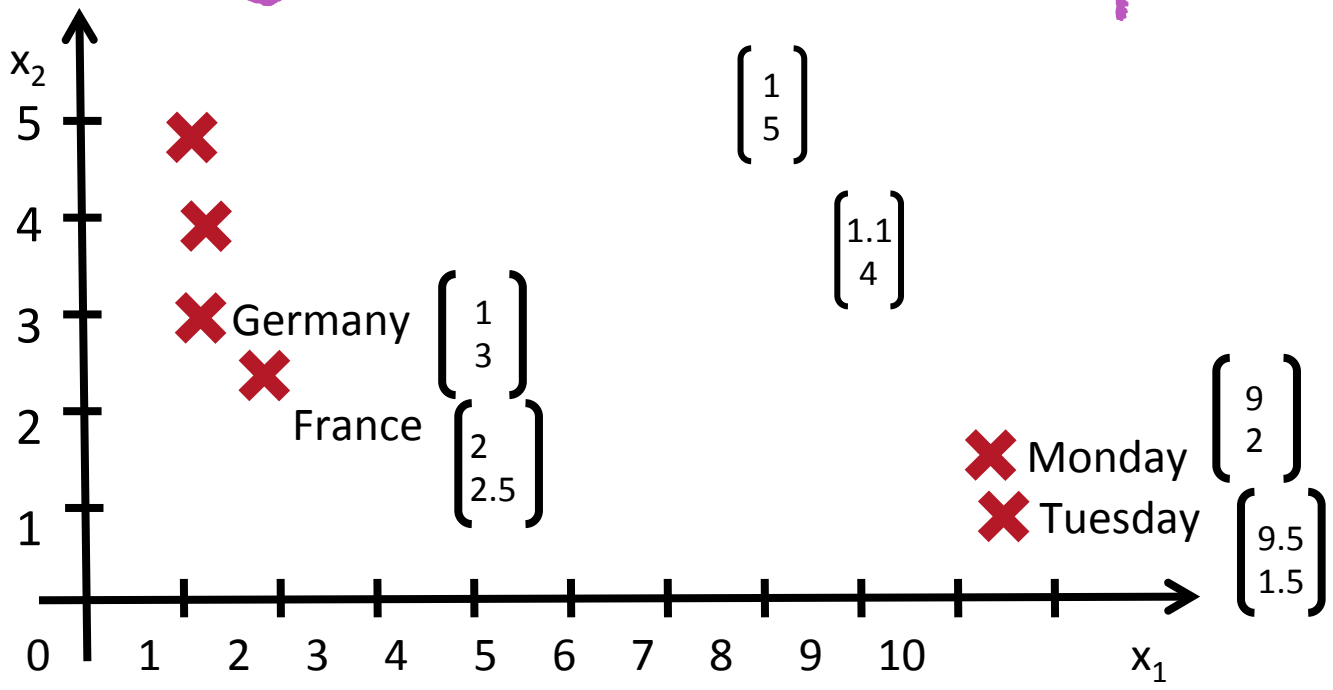
Recursive Neural Networks

Building on Word Vector Space Models



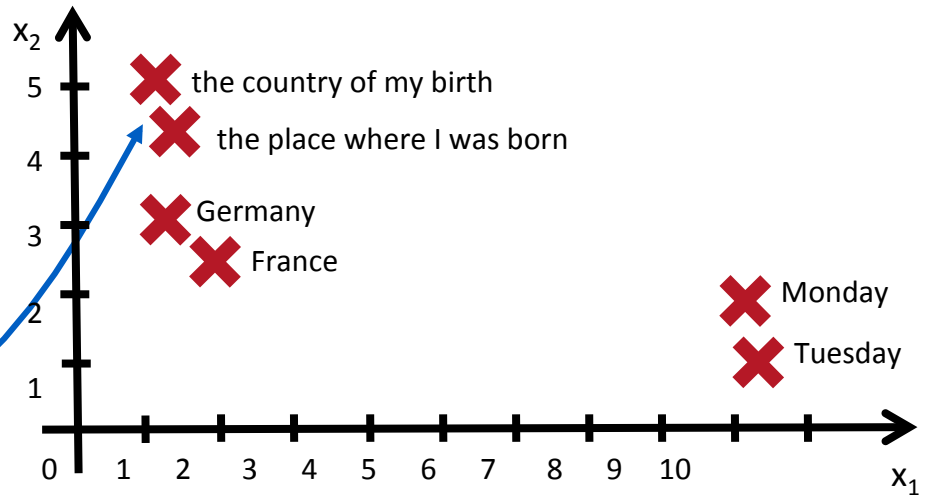
the country of my birth
the place where I was born

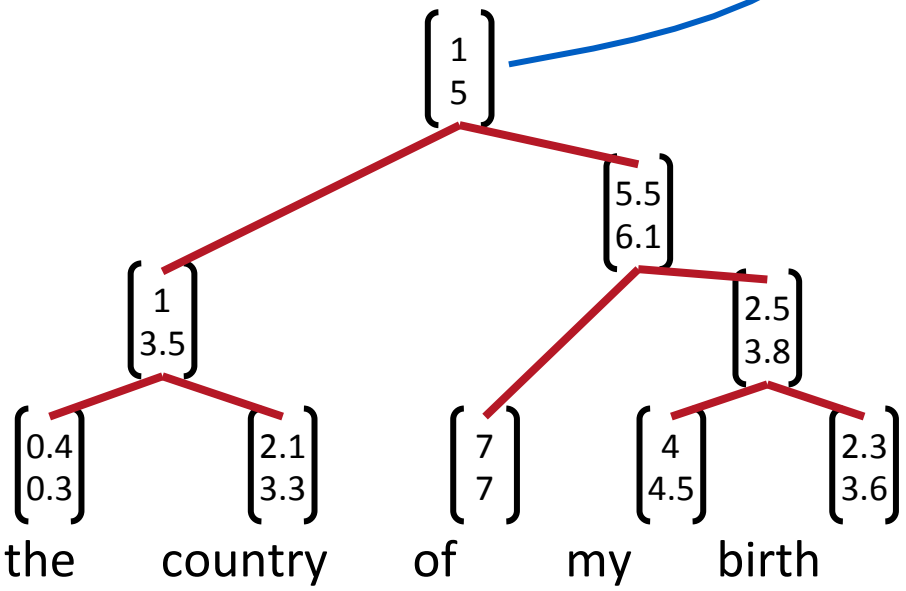
But how can we represent the meaning of longer phrases?

By mapping them into the same vector space!

How should we map phrases into a vector space?

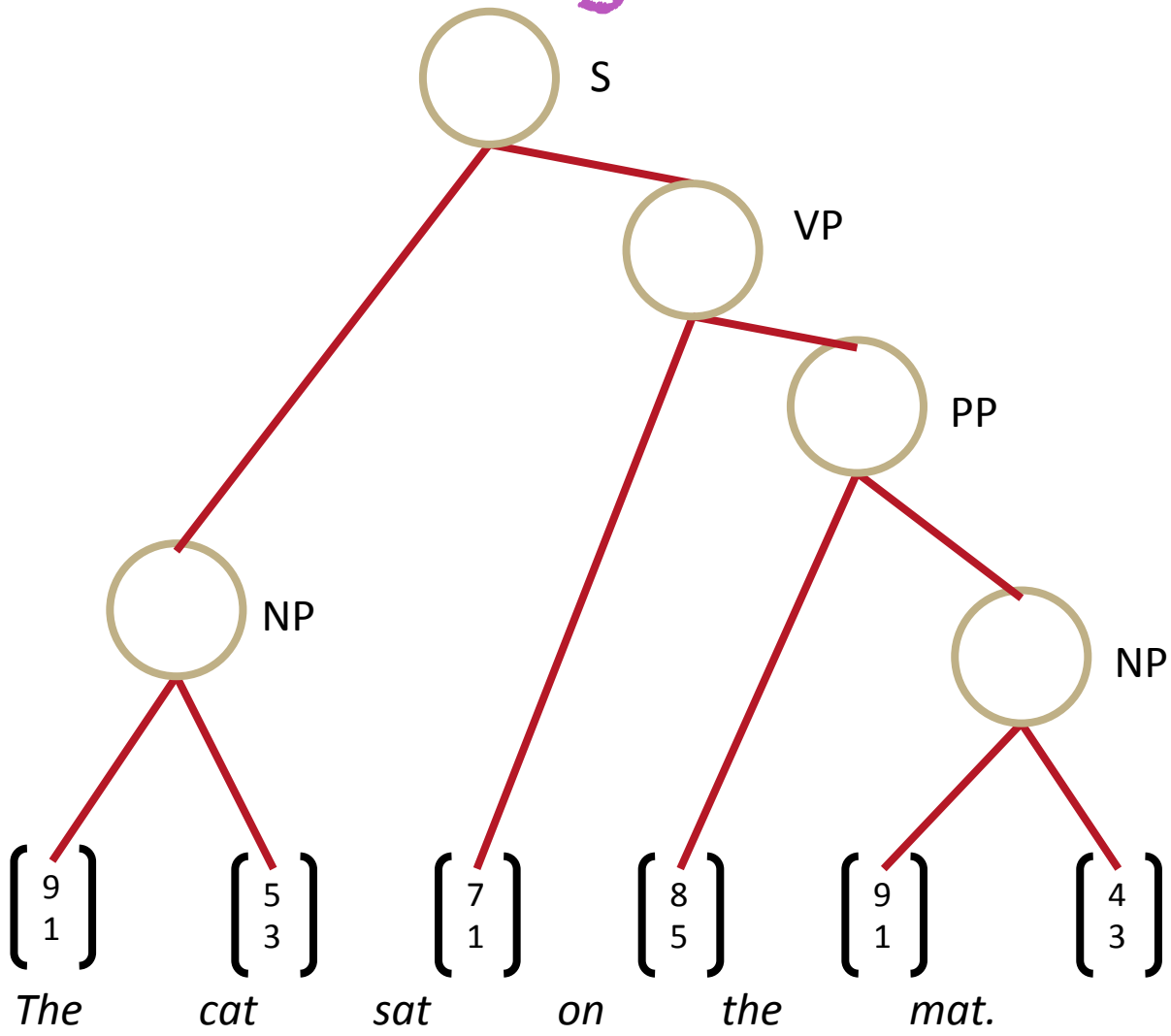
Use principle of compositionality
The meaning (vector) of a sentence is determined by (1) the meanings of its words and (2) the rules that combine them.



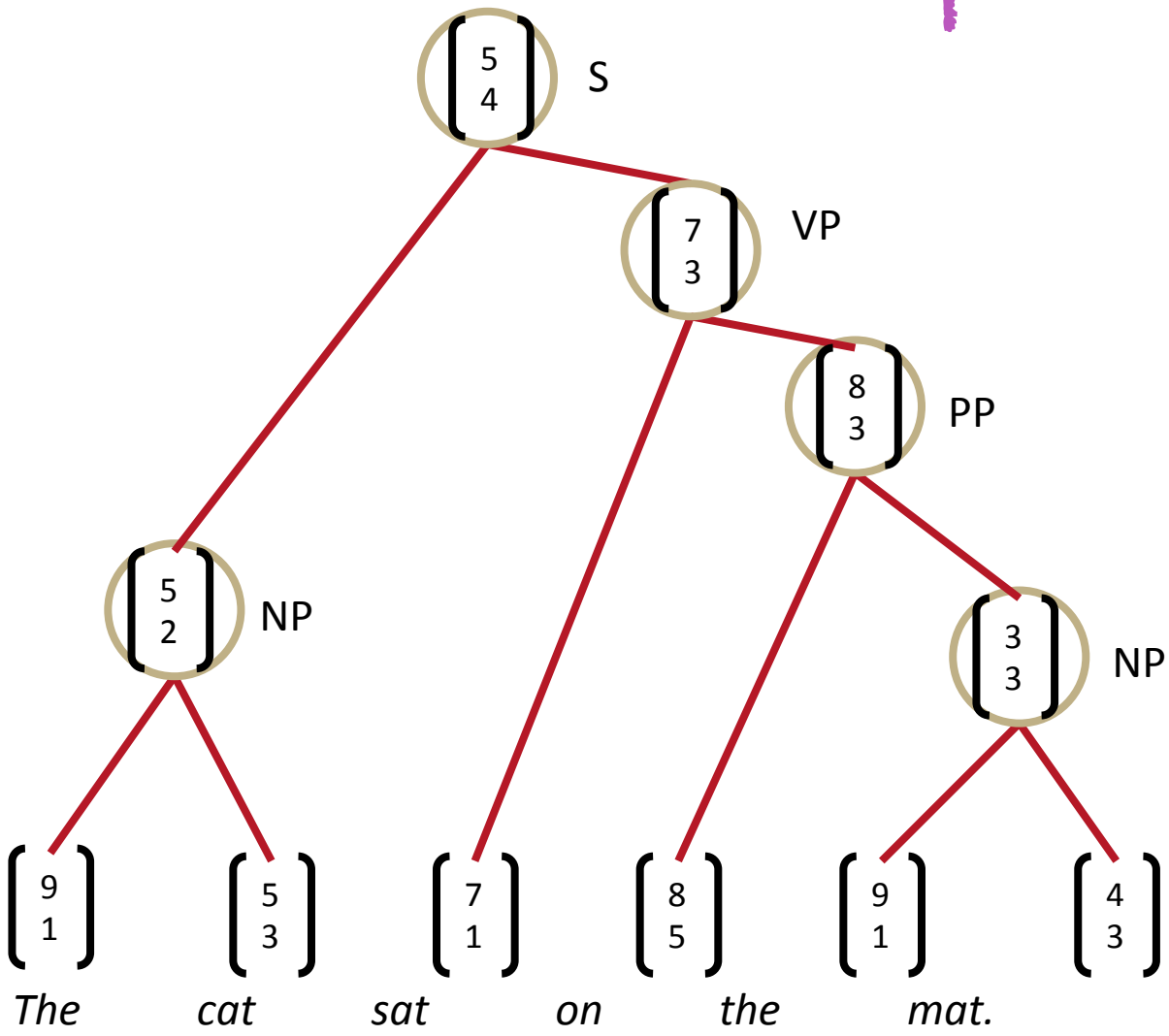


Recursive Neural Nets can jointly learn compositional vector representations and parse trees

Sentence Parsing: What we want



Learn Structure and Representation

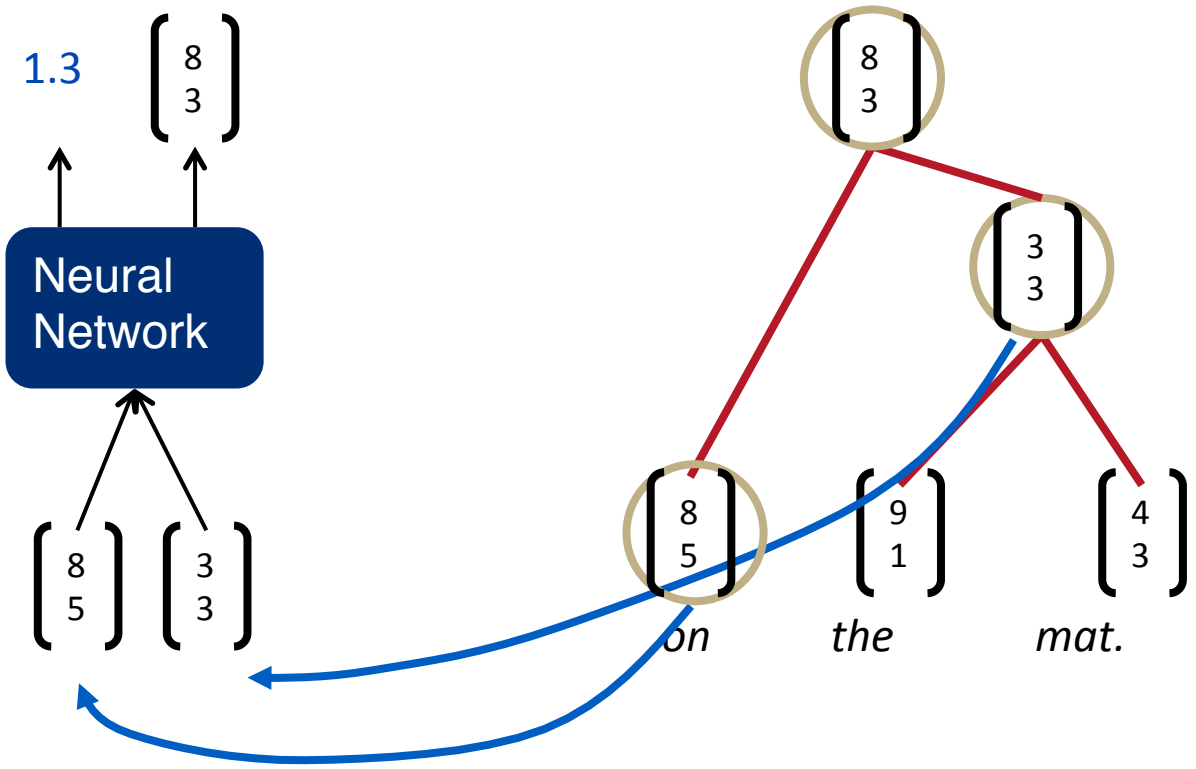


Recursive Neural Networks for Structure Prediction

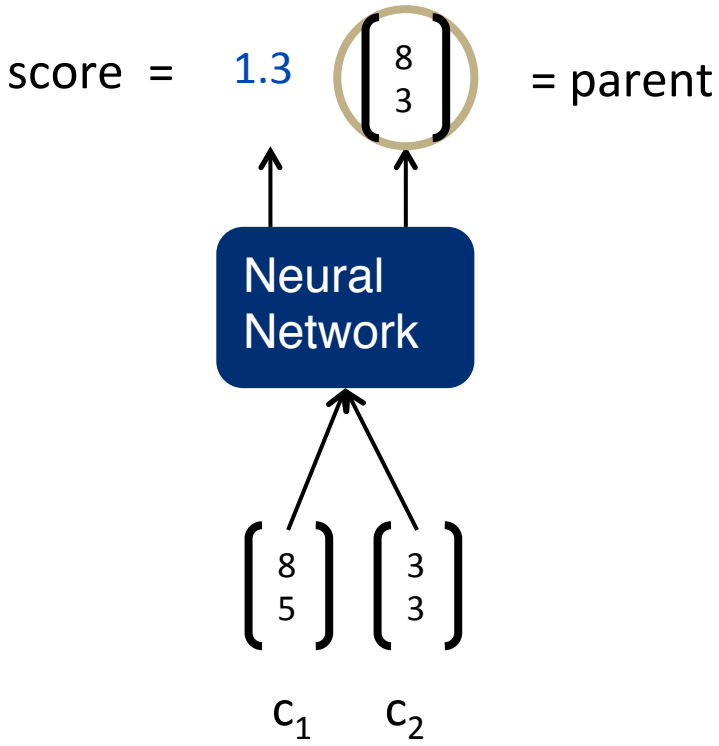
Inputs: two candidate children's representations

Outputs:

1. The semantic representation if the two nodes are merged.
2. Score of how plausible the new node would be.



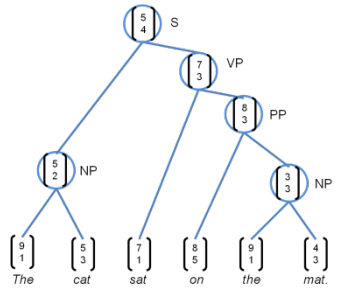
Recursive Neural Network Definition



score = $U^T p$

$$p = \tanh\left(W \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + b\right)$$

Same W parameters at all nodes of the tree

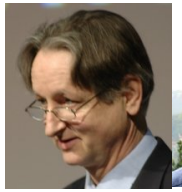


Related Work to Socher et al. (ICML 2011)

- Pollack (1990): Recursive auto-associative memories

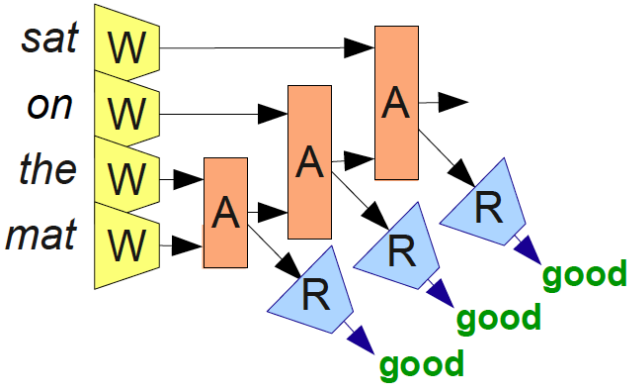

- Previous Recursive Neural Networks work by Goller & Küchler (1996), Costa et al. (2003) assumed fixed tree structure and used one hot vectors.

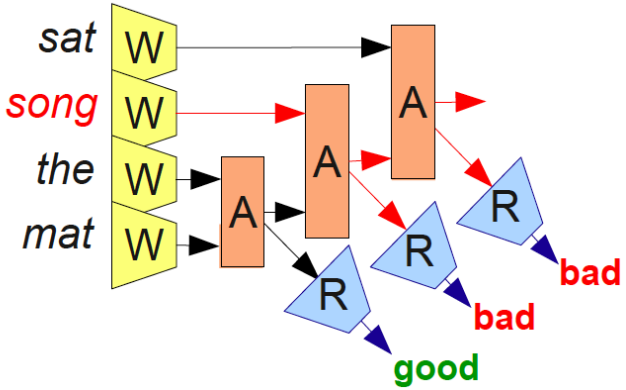
- Hinton (1990) and Bottou (2011): Related ideas about recursive models and recursive operators as smooth versions of logic operations

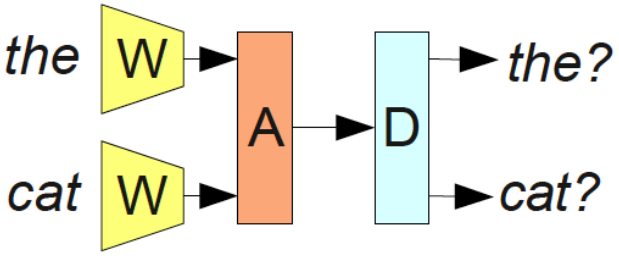


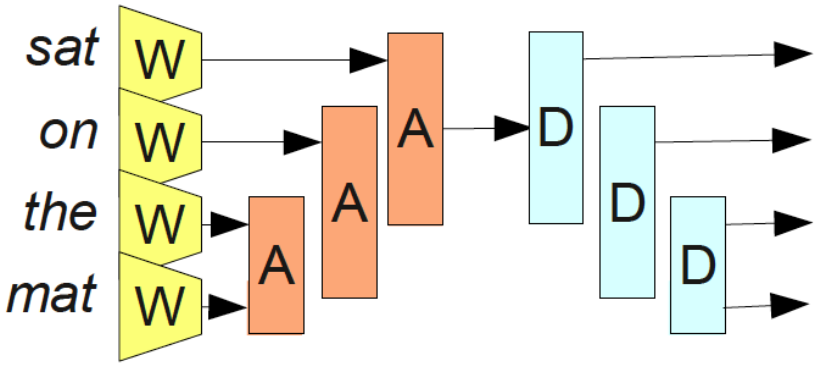

Recursive Application of Relational Operators

Bottou 2011: 'From machine learning to machine reasoning', also Socher ICML2011.

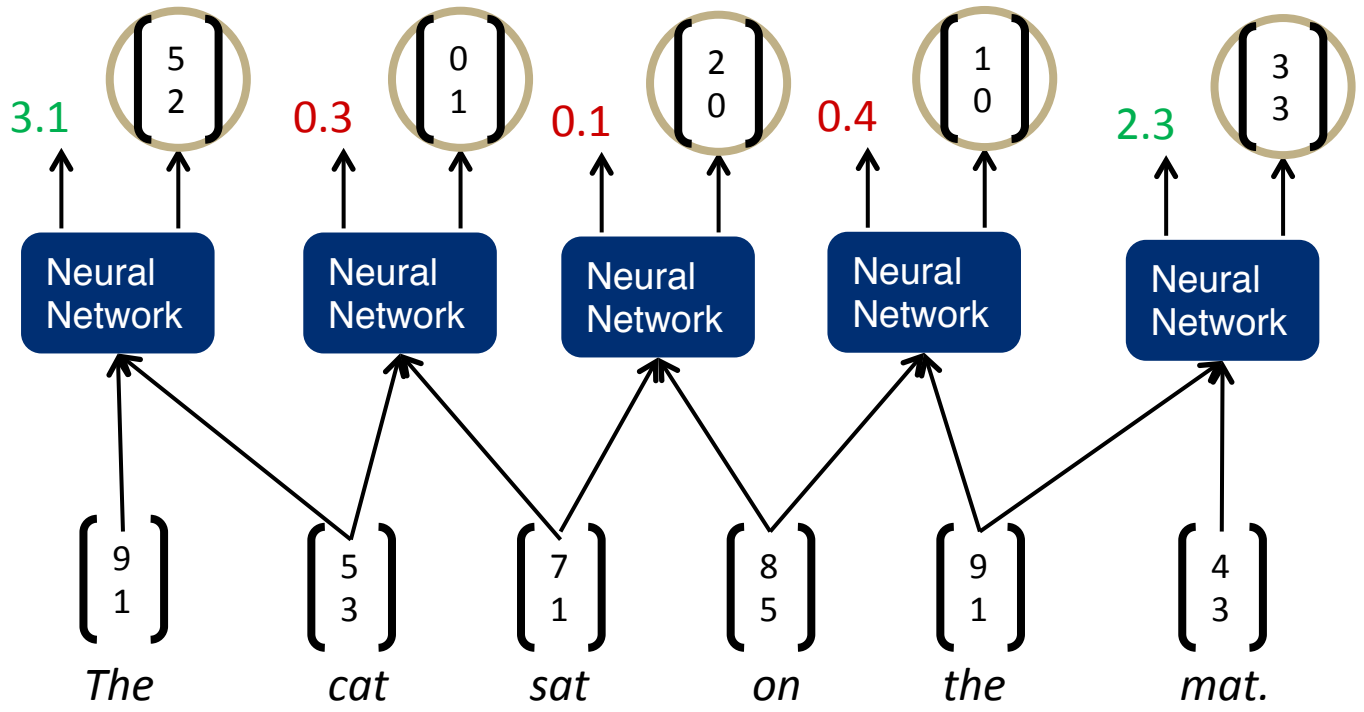




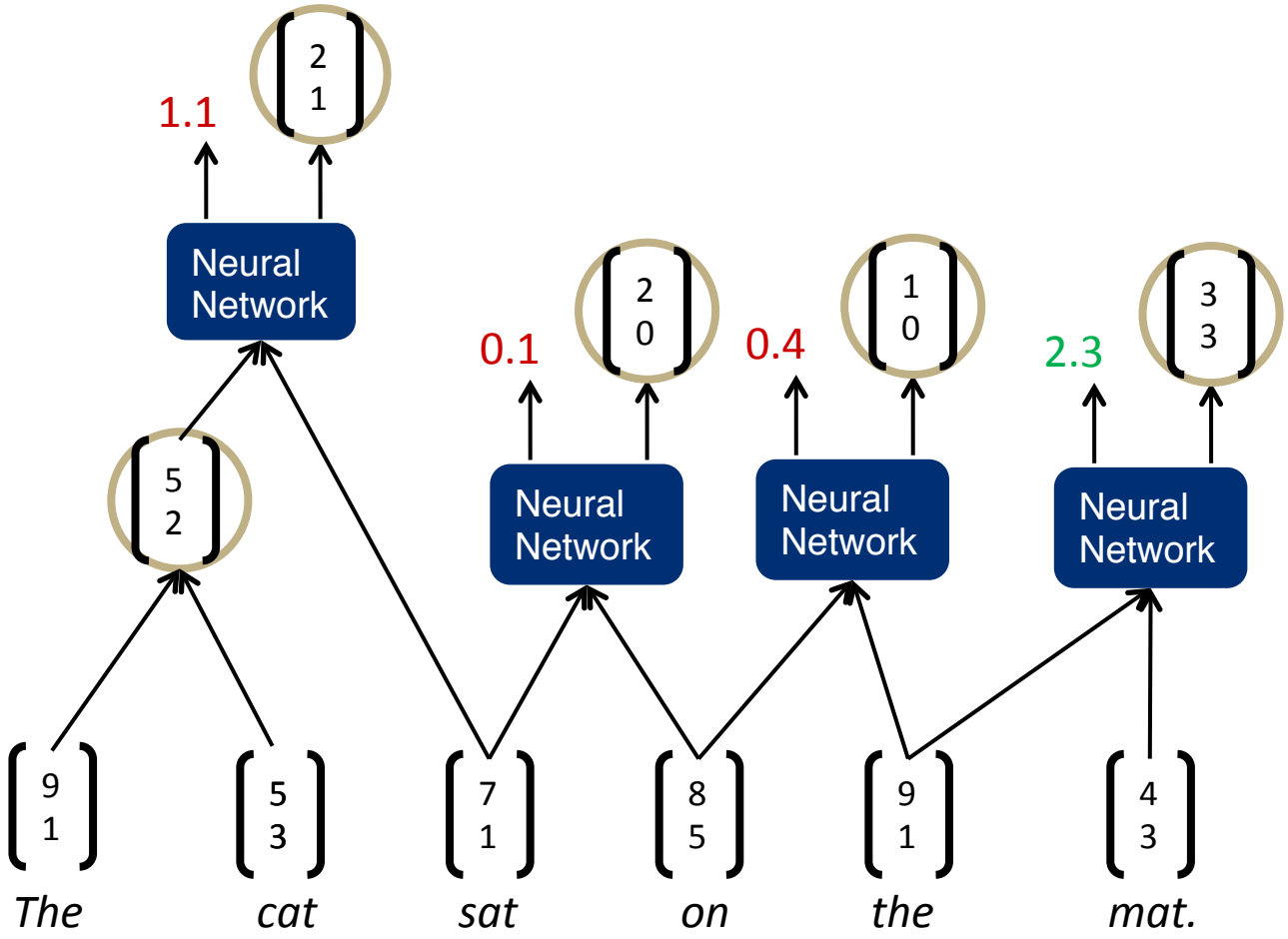




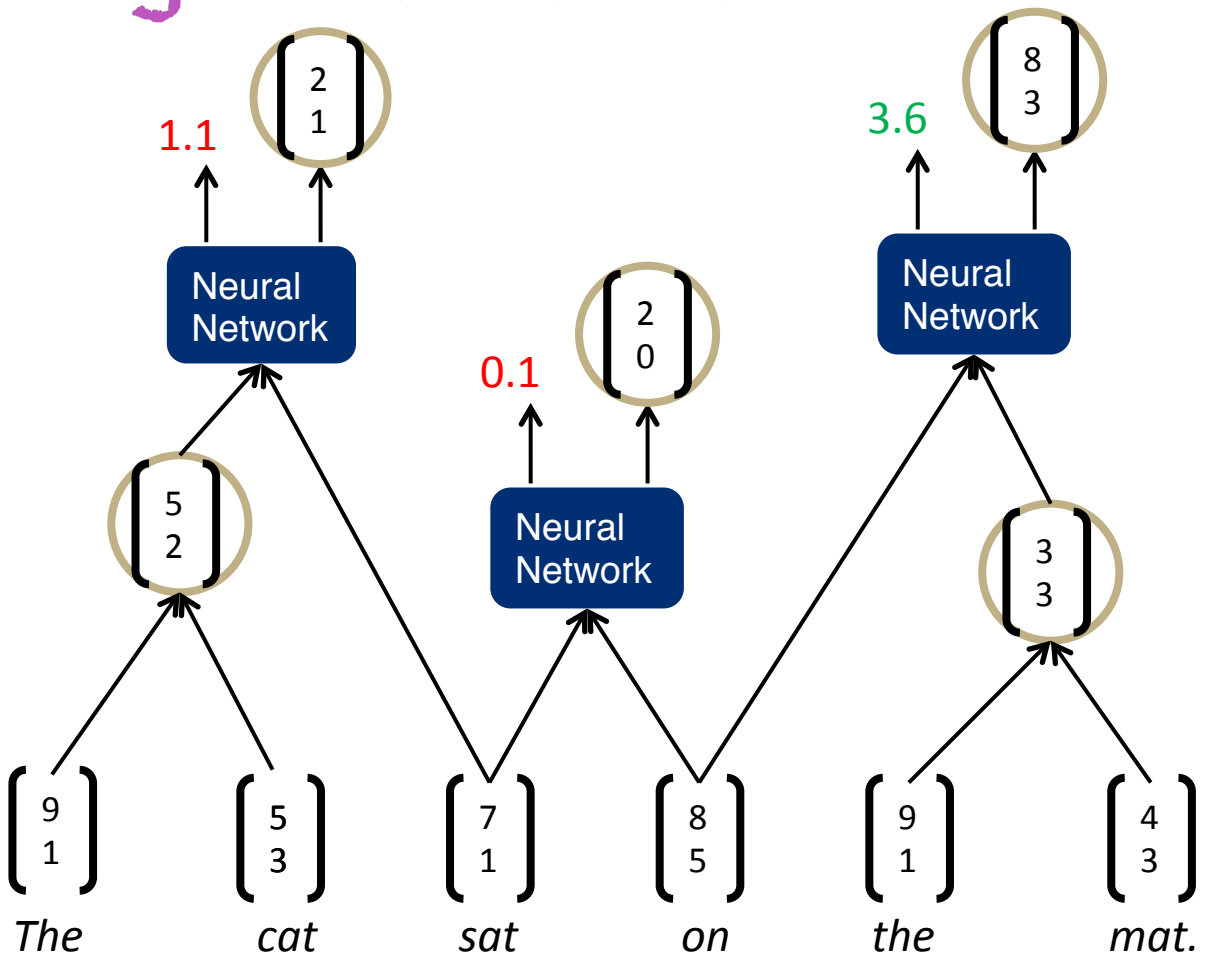
Parsing a sentence with an RNN



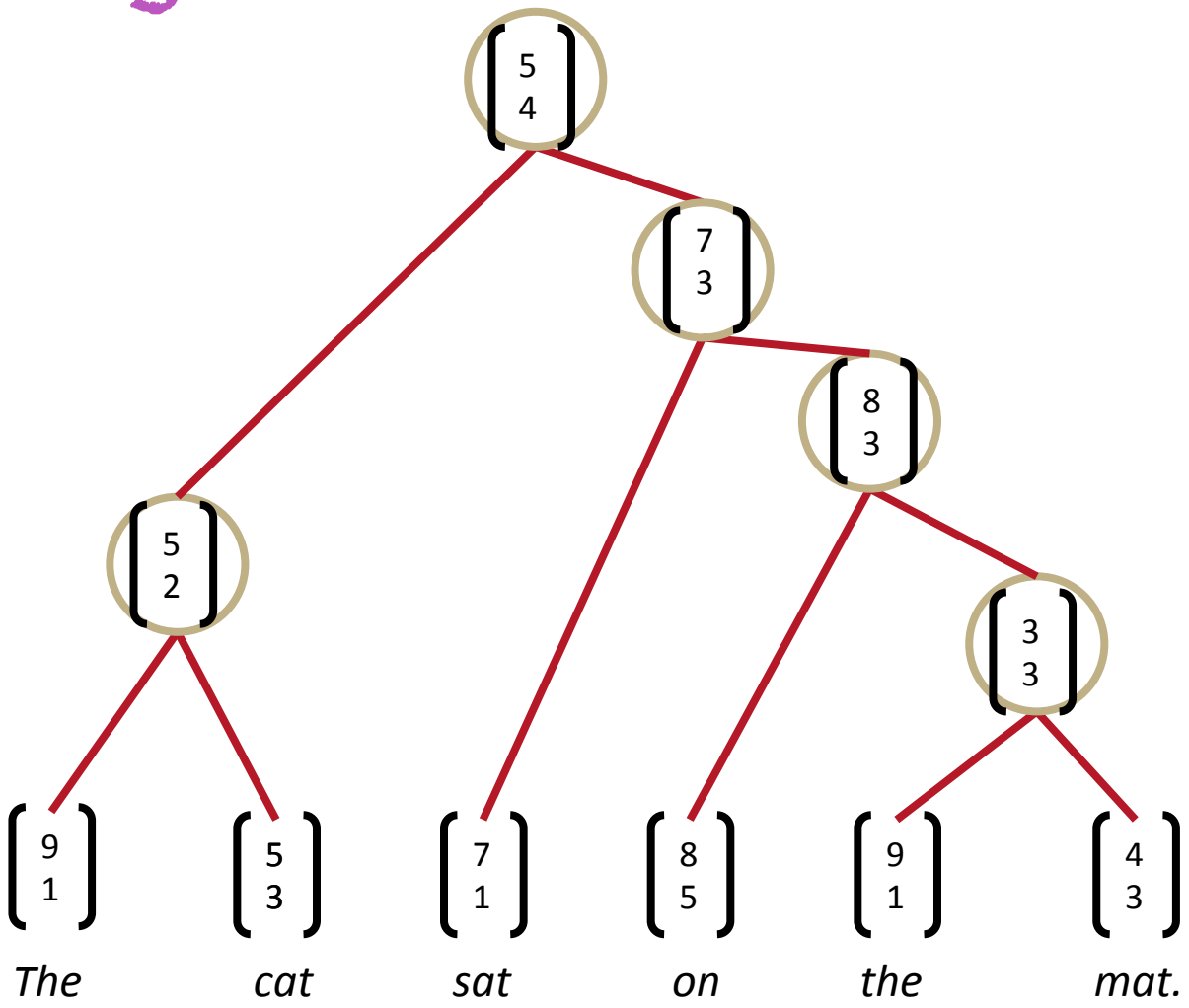
Parsing a sentence



Parsing a sentence



Parsing a sentence



Max-Margin Framework - Details

- The score of a tree is computed by the sum of the parsing decision scores at each node.

- Similar to max-margin parsing (Taskar et al. 2004), a supervised max-margin objective

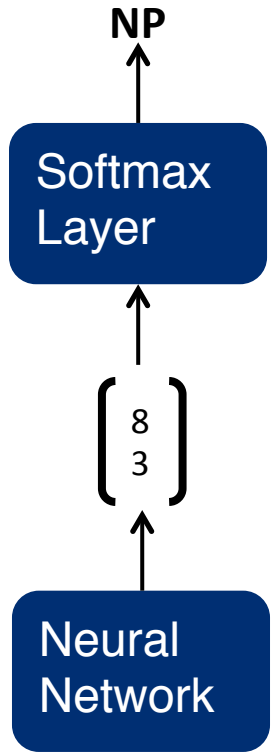
$$J = \sum_i s(x_i, y_i) - \max_{y \in A(x_i)} \left(s(x_i, y) + \Delta(y, y_i) \right)$$

- The loss $\Delta(y, y_i)$ penalizes all incorrect decisions

Labeling in Recursive Neural Networks

- We can use each node's representation as features for a *softmax* classifier:

$$p(c|p) \quad = \quad softmax(Sp)$$



Experiments: Parsing Short Sentences

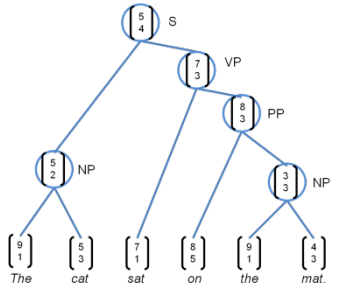
- Standard *WSJ* train/test
- Good results on short sentences
- More work is needed for longer sentences

Model	L15 Dev	L15 Test
Recursive Neural Network	92.1	90.3
Sigmoid NN (Titov & Henderson 2007)	89.5	89.3
Berkeley Parser (Petrov & Klein 2006)	92.1	91.6

All the figures are adjusted for seasonal variations
1. All the numbers are adjusted for seasonal fluctuations
2. All the figures are adjusted to remove usual seasonal patterns

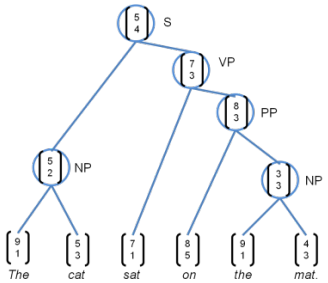
Knight-Ridder wouldn't comment on the offer
1. Harsco declined to say what country placed the order
2. Coastal wouldn't disclose the terms

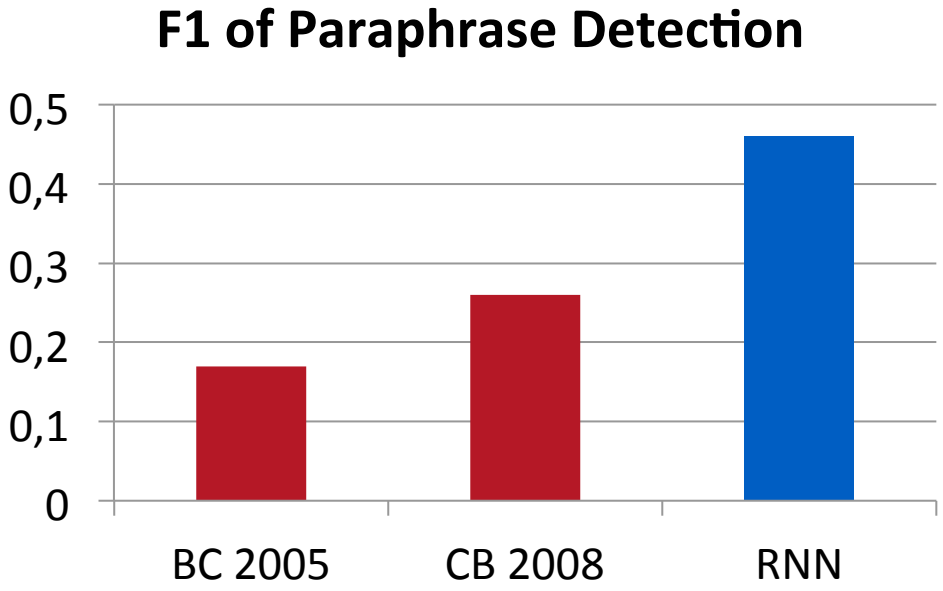
Sales grew almost 7% to $UNK m. from $UNK m.
1. Sales rose more than 7% to $94.9 m. from $88.3 m.
2. Sales surged 40% to UNK b. yen from UNK b.



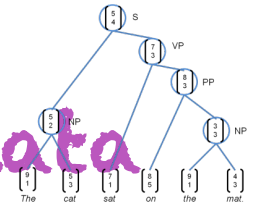
Short Paraphrase Detection

- Goal is to say which of candidate phrases are a good paraphrase of a given phrase
  - Motivated by Machine Translation
  - Initial algorithms: Bannard & Callison-Burch 2005 (BC 2005), Callison-Burch 2008 (CB 2008) exploit bilingual sentence-aligned corpora and hand-built linguistic constraints
  - Re-use system trained on parsing the WSJ





Paraphrase detection task, CCB data

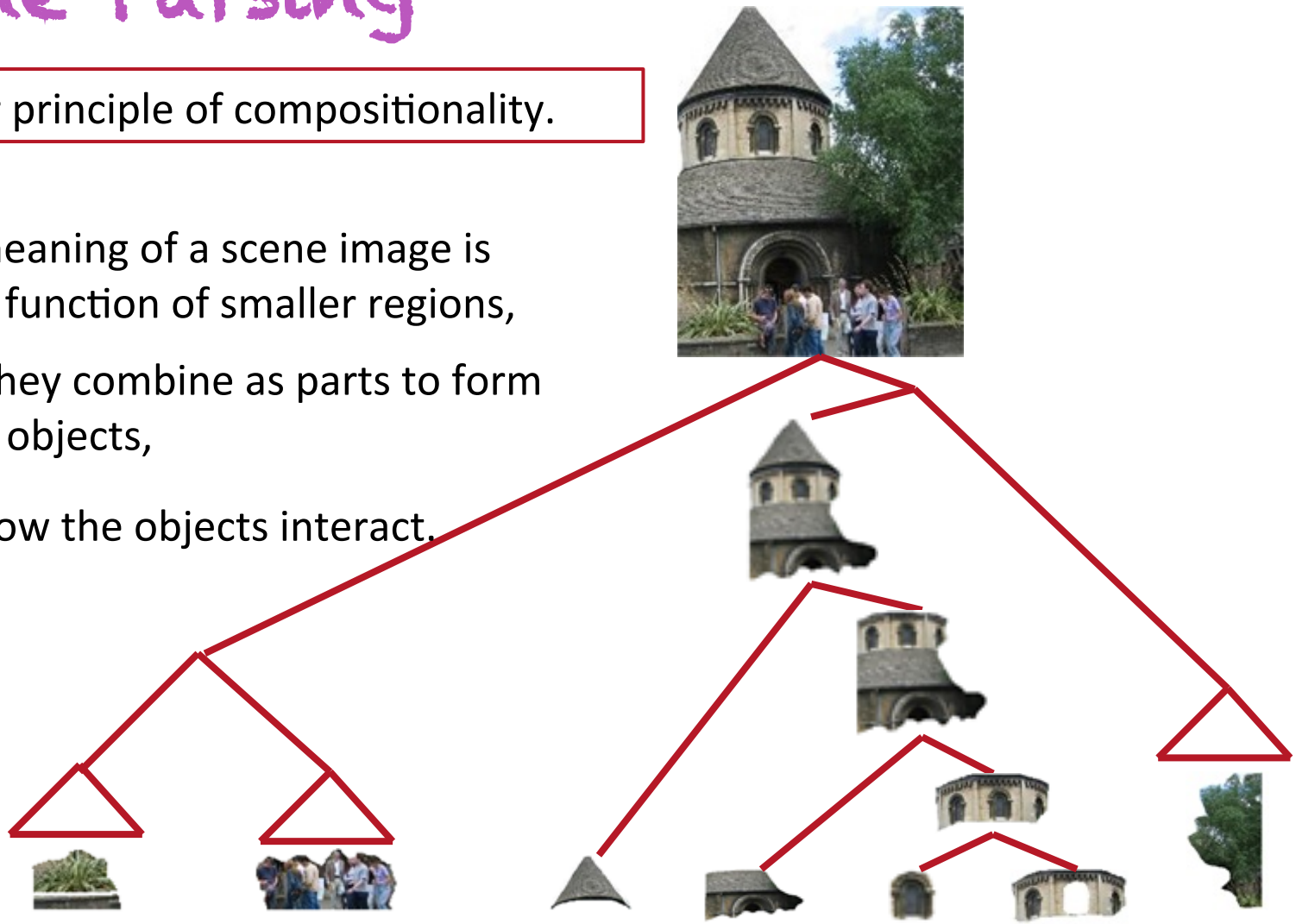


Target	Candidates with human goodness label (1–5) ordered by recursive net
the united states	the usa (5) the us (5) united states (5) north america (4) united (1) the (1) of the united states (3) america (5) nations (2) we (3)
around the world	around the globe(5) throughout the world(5) across the world(5) over the world(2) in the world(5) of the budget(2) of the world(5)
it would be	it would represent (5) there will be (2) that would be (3) it would be ideal (2) it would be appropriate (2) it is (3) it would (2)
of capital punishment	of the death penalty (5) to death (2) the death penalty (2) of (1)
in the long run	in the long term (5) in the short term (2) for the longer term (5) in the future (5) in the end (3) in the long-term (5) in time (5) of the (1)

Scene Parsing

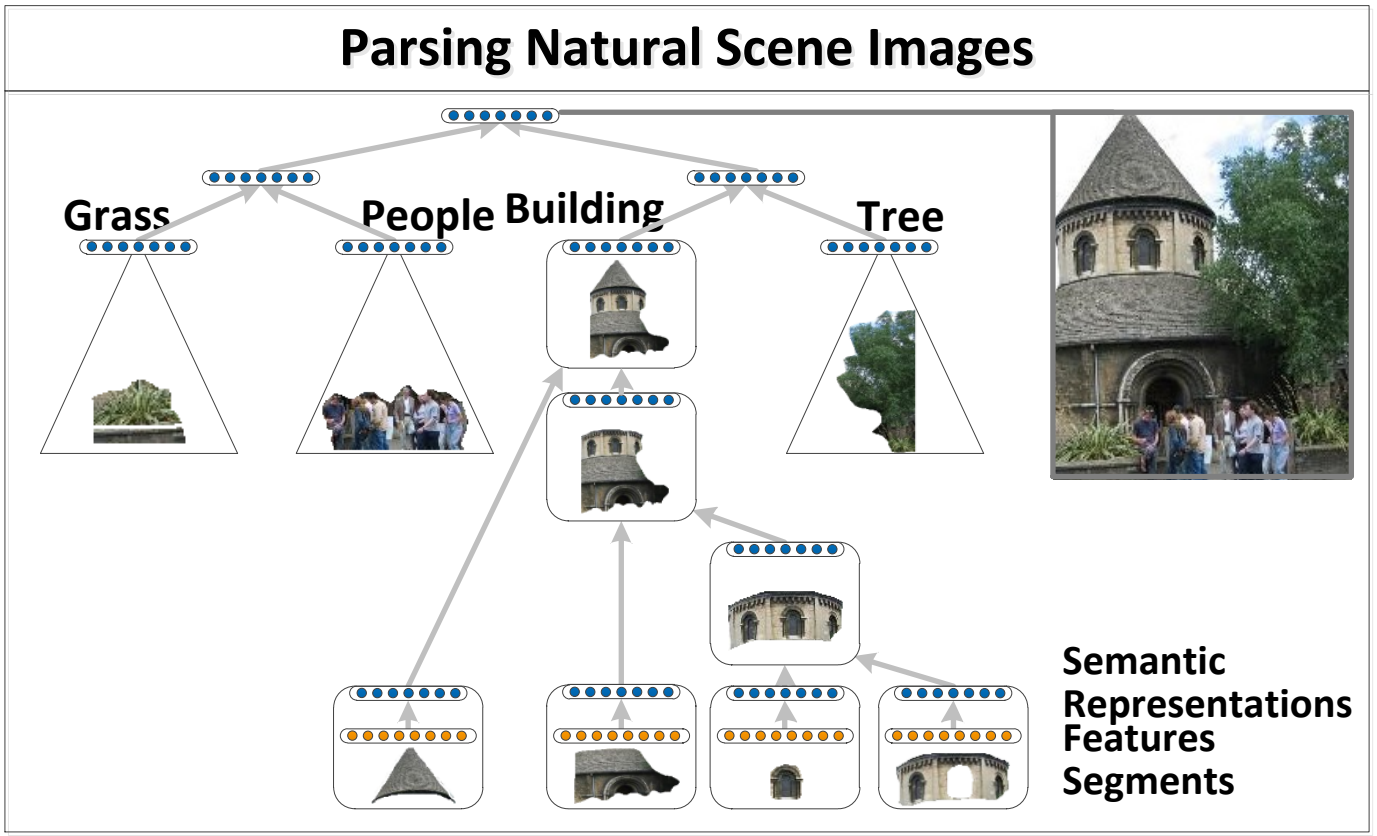
Similar principle of compositionality.

- The meaning of a scene image is also a function of smaller regions,
- how they combine as parts to form larger objects,
- and how the objects interact.



Algorithm for Parsing Images

Same Recursive Neural Network as for natural language parsing!
(Socher et al. ICML 2011)



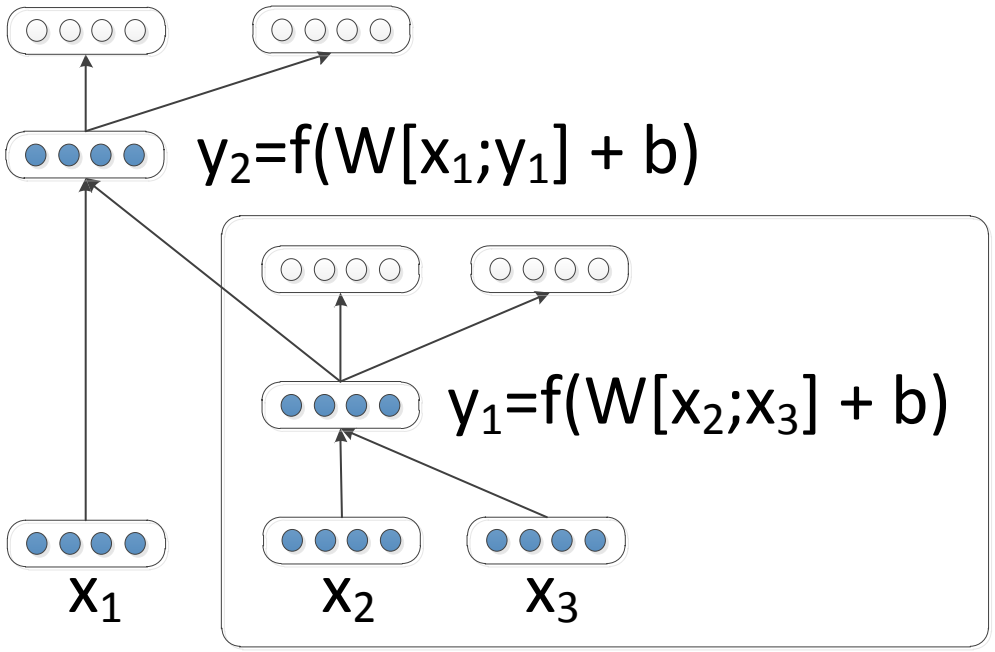
Multi-class segmentation

Method	Accuracy
Pixel CRF (Gould et al., ICCV 2009)	74.3
Classifier on superpixel features	75.9
Region-based energy (Gould et al., ICCV 2009)	76.4
Local labelling (Tighe & Lazebnik, ECCV 2010)	76.9
Superpixel MRF (Tighe & Lazebnik, ECCV 2010)	77.5
Simultaneous MRF (Tighe & Lazebnik, ECCV 2010)	77.5
Recursive Neural Network	**78.1**

Stanford Background Dataset (Gould et al. 2009)

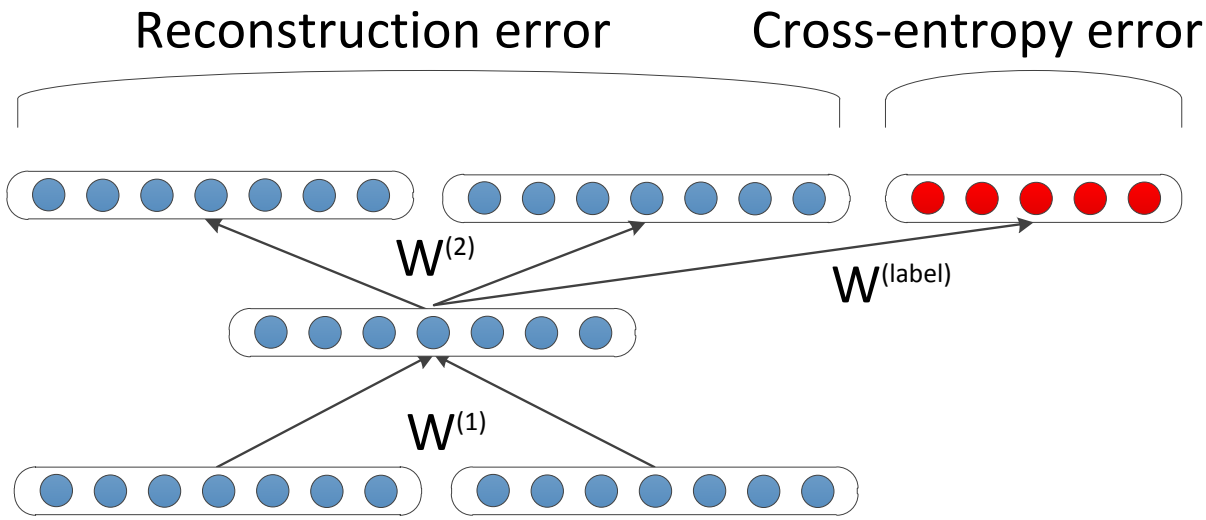
Recursive Autoencoders

- Similar to Recursive Neural Net but instead of a supervised score we compute a reconstruction error at each node. $E_{rec}([c_1; c_2]) = \frac{1}{2} \left|\left|[c_1; c_2] - [c_1'; c_2']\right|\right|^2$



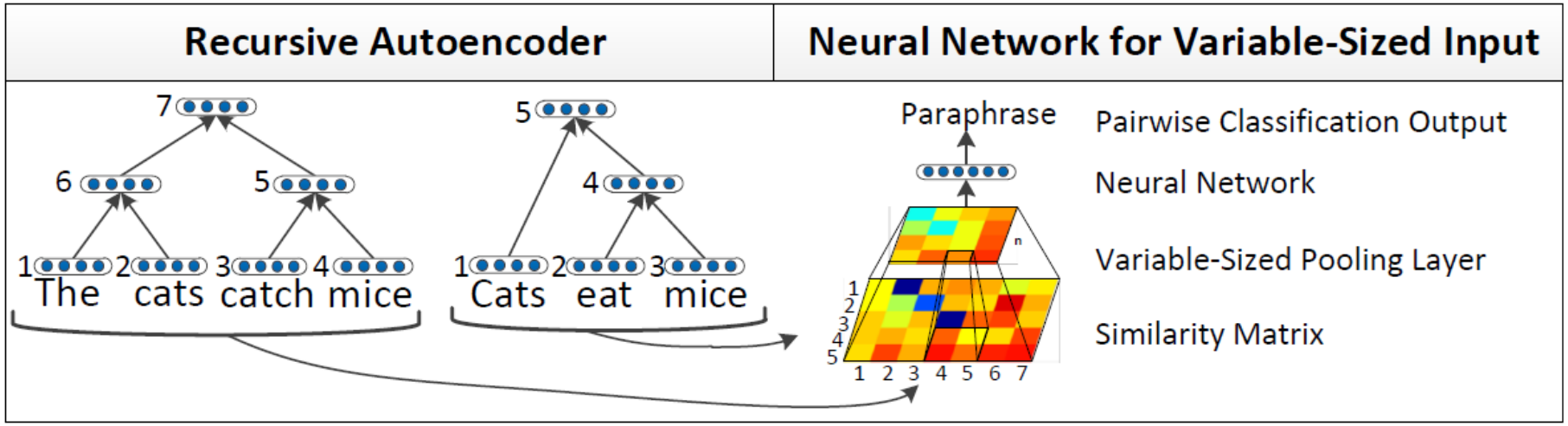
Semi-supervised Recursive Autoencoder

- To capture sentiment and solve antonym problem, add a softmax classifier
- Error is a weighted combination of reconstruction error and cross-entropy
- Socher et al. (EMNLP 2011)



Comparing the meaning of two sentences: Paraphrase Detection

- Unsupervised Unfolding RAE and a pair-wise sentence comparison of nodes in parsed trees
- Socher et al. (NIPS 2011)



Recursive Autoencoders for Full Sentence Paraphrase Detection

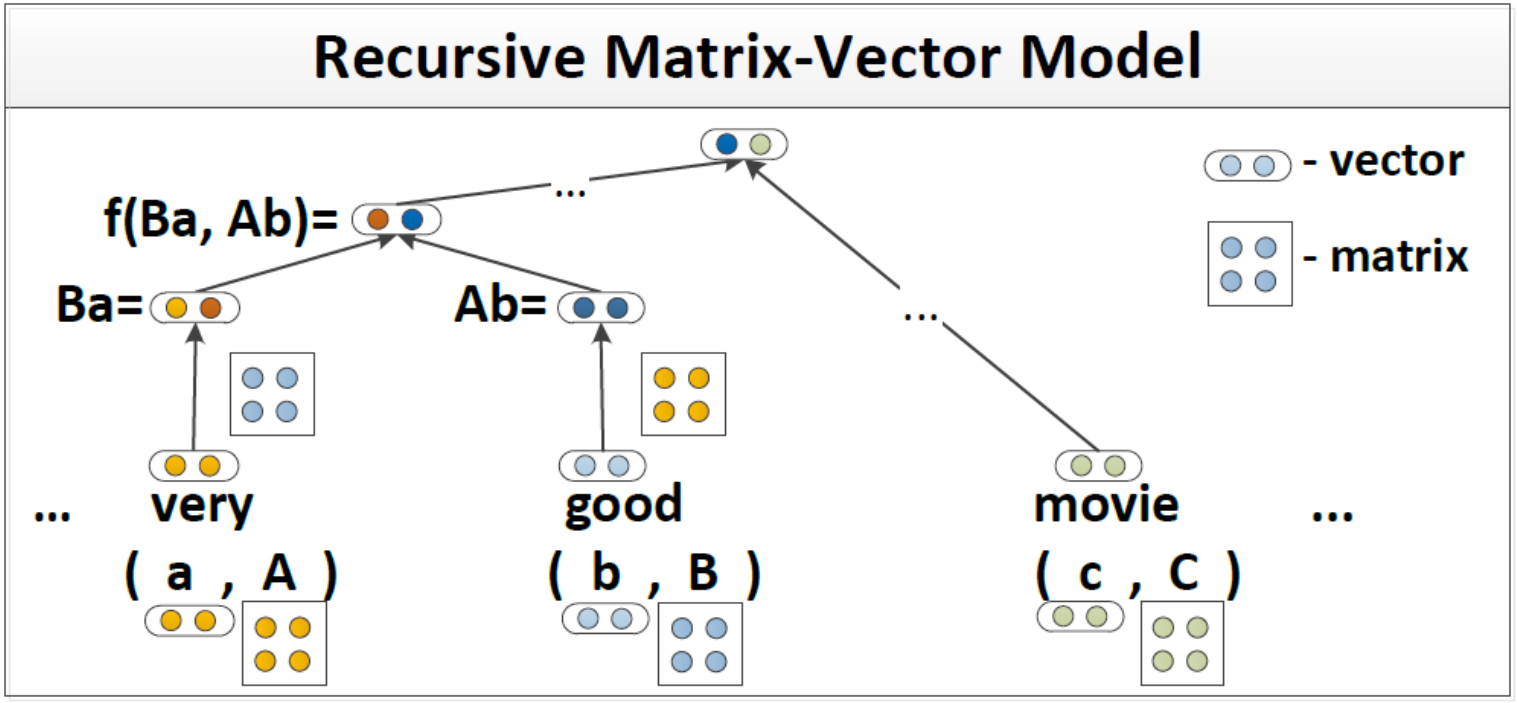
- Experiments on Microsoft Research Paraphrase Corpus
- (Dolan et al. 2004)

Method	Acc.	F1
Rus et al.(2008)	70.6	80.5
Mihalcea et al.(2006)	70.3	81.3
Islam et al.(2007)	72.6	81.3
Qiu et al.(2006)	72.0	81.6
Fernando et al.(2008)	74.1	82.4
Wan et al.(2006)	75.6	83.0
Das and Smith (2009)	73.9	82.3
Das and Smith (2009) + 18 Surface Features	76.1	82.7
F. Bu et al. (ACL 2012): String Re-writing Kernel	76.3	--
Unfolding Recursive Autoencoder (NIPS 2011)	**76.8**	**83.6**

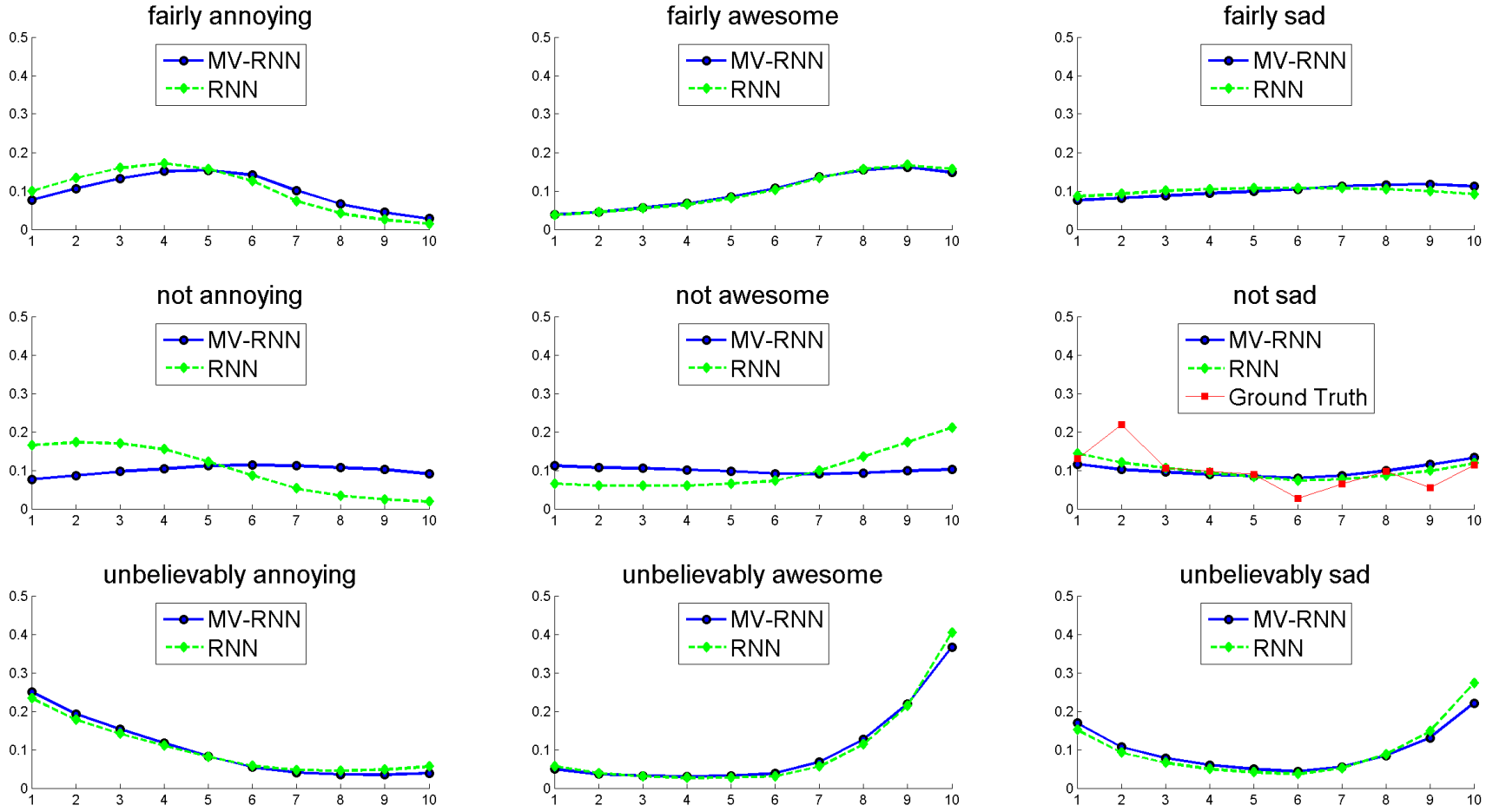
Compositionality Through Recursive Matrix-Vector Recursive Neural Networks

$$p = \tanh\left(W \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + b\right)$$

$$p = \tanh\left(W \begin{bmatrix} C_2 c_1 \\ C_1 c_2 \end{bmatrix} + b\right)$$



Predicting Sentiment Distributions



Discussion

Concerns

- Many algorithms and variants (burgeoning field)

- Hyper-parameters (layer size, regularization, possibly learning rate)
  - Use multi-core machines, clusters and random sampling for cross-validation (Bergstra & Bengio 2012)
  - Pretty common for powerful methods, e.g. BM25

Concerns

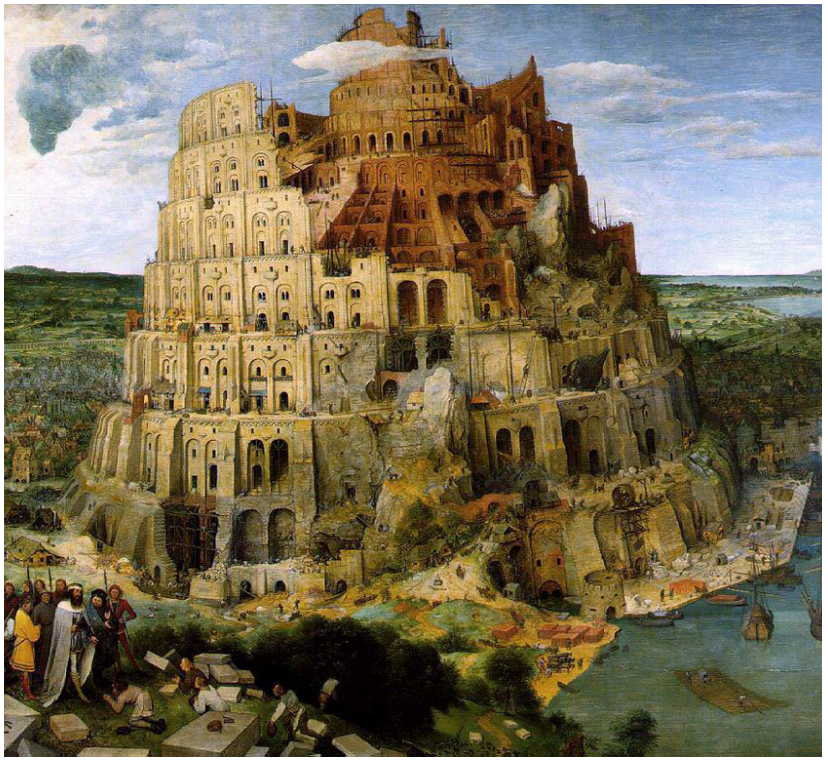
- Slower to train than linear models
  - Only by a small constant factor, and much more compact than non-parametric (e.g. n-gram models)
  - Very fast during inference/test time (feed-forward pass is just a few matrix multiplies)
- Need more training data?
  - Can *handle and benefit from* more training data (esp. unlabeled), suitable for age of Big Data (Google trains neural nets with a billion connections, [Le et al, ICML 2012])
  - Need less ***labeled*** data

Concern: non-convex optimization

- Can initialize system with convex learner
  - Convex SVM
  - Fixed feature space

- Then optimize non-convex variant (add and tune learned features), can't be worse than convex learner

Transfer Learning



- Application of deep learning could be in areas where there are not enough labeled data but a transfer is possible
- Domain adaptation already showed that effect, thanks to unsupervised feature learning
- **Two transfer learning competitions won in 2011**
- Transfer to resource-poor languages would be a great application [Gouws, PhD proposal 2012]

Learning Multiple Levels of Abstraction

- The big payoff of deep learning is to allow learning higher levels of abstraction
- Higher-level abstractions disentangle the factors of variation, which allows much easier generalization and transfer
- More abstract representations

→Successful transfer (domains, languages)