DEEP SEQUENTIAL AND STRUCTURAL MODELS OF COMPOSITIONALITY

Ozan Irsoy and Claire Cardie

based on

Opinion Mining with Deep Recurrent Neural Networks (EMLNP 2014)

Deep Recursive Neural Networks for Compositionality in Language (NIPS 2014)

Distributed Meaning Representations

Principle of Compositionality

Meaning of an expression is determined by its parts, and the rules to combine them.

Composing Meaning

How can we utilize these word representations to generate representation for phrases or sentences?

- Orderless (B.O.W-like) composition
  - Elementwise commutative operations
- Sequential (left-to-right) composition
  - Recurrent neural networks
  - Matrix-space models
- Structural (tree-based) composition
  - Recursive neural networks
- Others
  - Convolution based models

Neural Net Based Composition

Neural network based models can exploit the sequential and structural representations that naturally exist in language.

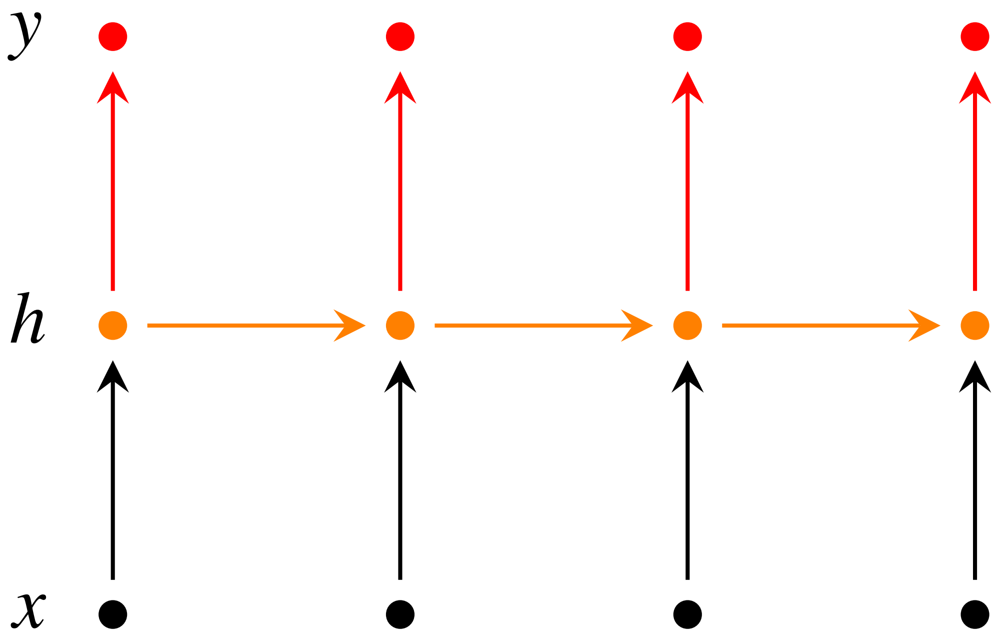
Deep (stacked) versions of these neural networks can utilize the hierarchical data processing capabilities that exist in traditional deep learning approaches (such as in computer vision).

Rough Outline

- Deep recurrent neural networks
  Application to opinion mining

- Deep recursive neural networks
  Application to sentiment analysis

Recurrent Neural Networks



$$h_t = f(Wx_t + Vh_{t-1} + b)$$

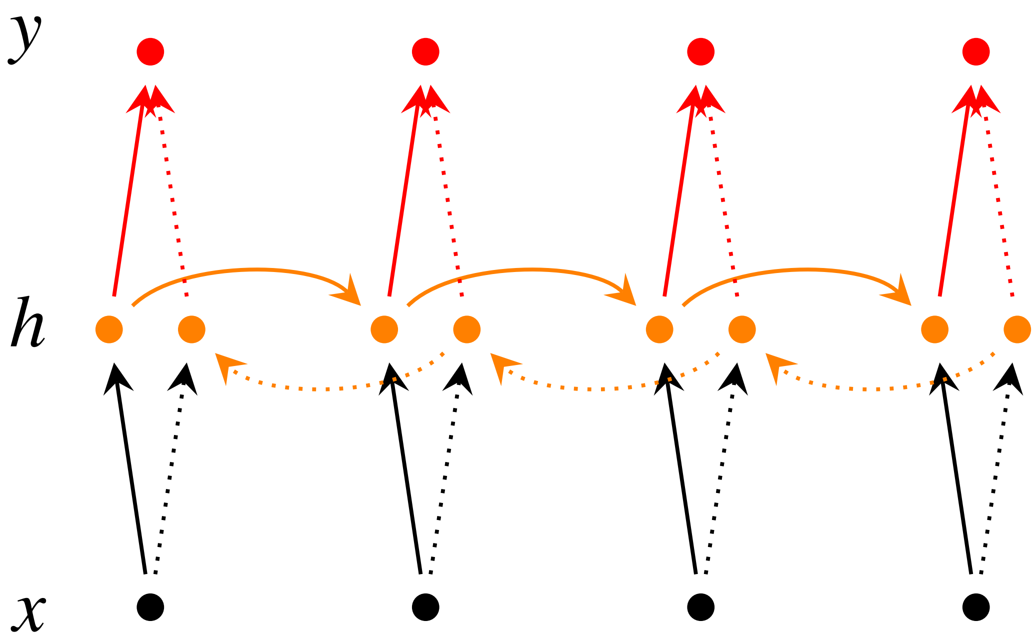
$$y_t = g(Uh_t + c)$$

x represents a token (word) as a vector.

y represents the output label.

h is the memory, computed from the past memory and current word. It summarizes the sentence up to that time.

Bidirectionality



$$\overrightarrow{h}_t = f(\overrightarrow{W}x_t + \overrightarrow{V}\overrightarrow{h}_{t-1} + \overrightarrow{b})$$

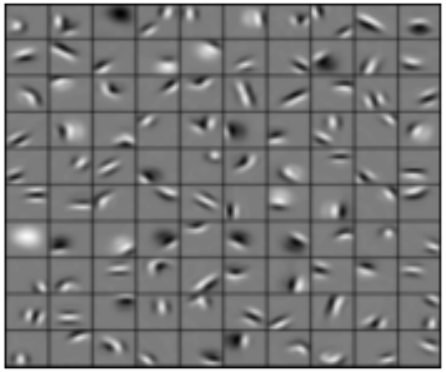
$$\overleftarrow{h}_t = f(\overleftarrow{W}x_t + \overleftarrow{V}\overleftarrow{h}_{t+1} + \overleftarrow{b})$$

$$y_t = g(U[\overrightarrow{h}_t; \overleftarrow{h}_t] + c)$$

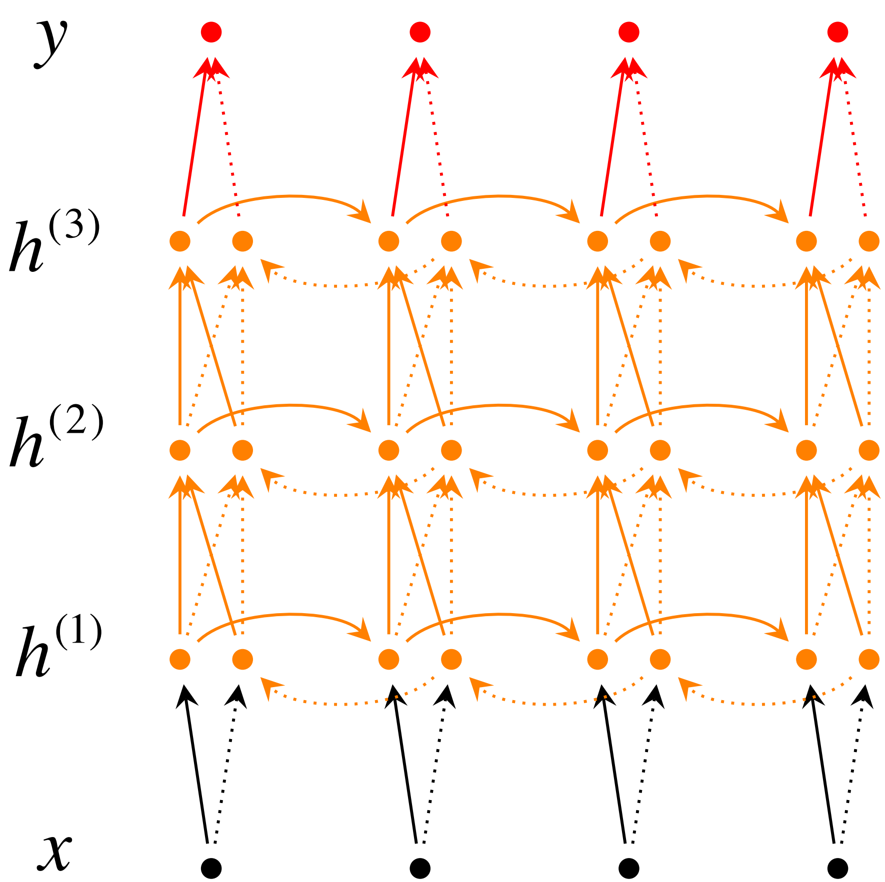
$h = [\overrightarrow{h}; \overleftarrow{h}]$ now represents (summarizes) the past and future around a single token.

Going Deep

Are recurrent networks really *deep*? (e.g. like this)



Going Deep



$$\overrightarrow{h}_t^{(i)} = f(\overrightarrow{W}^{(i)} h_t^{(i-1)} + \overrightarrow{V}^{(i)} \overrightarrow{h}_{t-1}^{(i)} + \overrightarrow{b}^{(i)})$$

$$\overleftarrow{h}_t^{(i)} = f(\overleftarrow{W}^{(i)} h_t^{(i-1)} + \overleftarrow{V}^{(i)} \overleftarrow{h}_{t+1}^{(i)} + \overleftarrow{b}^{(i)})$$

$$y_t = g(U[\overrightarrow{h}_t^{(L)}; \overleftarrow{h}_t^{(L)}] + c)$$

Each memory layer passes an intermediate sequential representation to the next.

Opinion Mining

Fine-grained opinion analysis aims to detect subjectivity (e.g. "hate") and characterize

- Intensity (e.g. strong)
- Sentiment (e.g. negative)
- Opinion holder, target or topic
- …

Important for a variety of NLP tasks such as

- Opinion-oriented question answering
- Opinion summarization

Opinion Mining

In this work, we focus on detecting *direct subjective expressions* (DSEs) and *expressive subjective expressions* (ESEs).

DSE: Explicit mentions of private states or speech events expressing private states

ESE: Expressions that indicate sentiment, emotion, etc. without explicitly conveying them.

Example

The committee, [as usual]$_{ESE}$, [has refused to make any statements]$_{DSE}$.

In BIO notation (where a token is the atomic unit):

The	committee	,	as	usual	,	has
O	O	O	B_ESE	I_ESE	O	B_DSE

refused	to	make	any	statements	.
I_DSE	I_DSE	I_DSE	I_DSE	I_DSE	O

Related Work

Most earlier work formulated as a token-level sequence-labeling problem.

- Conditional Random Field (CRF) approaches (Breck et al. 2007)
- Joint detection of opinion holders with CRFs (Choi et al. 2005)
- Reranking approaches over a sequence labeler (Johansson and Moschitti, 2010 & 2011)
- Semi Markov CRF (semiCRF) based approaches, which operate at the phrase level rather than token level (Yang and Cardie, 2012 & 2013)

Related Work

Success of CRF based approaches hinges critically on access to a good feature set, typically based on

- Constituency and dependency parse trees
- Manually crafted opinion lexicons
- Named entity taggers
- Other preprocessing components

(See Yang and Cardie (2012) for an up-to-date list.)

What about feature learning?

Approach

- We adopt the same sequential prediction approach:
A sentence is a sequence of tokens, each having a BIO based label.
- We use bidirectional shallow and deep Recurrent Neural Networks (RntNN) for sequential prediction.
- RntNNs have access to only a single feature set:
Word vectors (which are trained in an unsupervised fashion).

Data

We use the MPQA 1.2 corpus (Wiebe et al., 2005) which consists of 535 news articles (11,111 sentences) that is manually labeled with DSE and ESEs at the phrase level.

As in previous work, we separate 135 documents as the development set to do model selection, and employ 10-fold cross-validation over the remaining 400 documents.

Performance Metrics

Exact boundaries are difficult, even for human annotators.

Two softer accuracy measures:

- Binary overlap: Every overlapping match between a predicted and true expression is correct.
- Proportional overlap: Every overlapping match is partially correct proportional to the overlapping amount (contribution of each match is in [0, 1]).

Binary and proportional Precision, Recall and F-measure are defined over these accuracy notions.

Network Training

- Softmax and rectifier nonlinearities are used for output and hidden layer activations, respectively.
- Dropout regularization.
- Stochastic gradient descent with Cross-Entropy classification objective.
- Model selection is done via cross-validation over Proportional F1 metric.
- No pre-training, no fine-tuning.
- Two different parameter sizes: ~24k and ~200k. Therefore increasing depth cause a decrease in width.

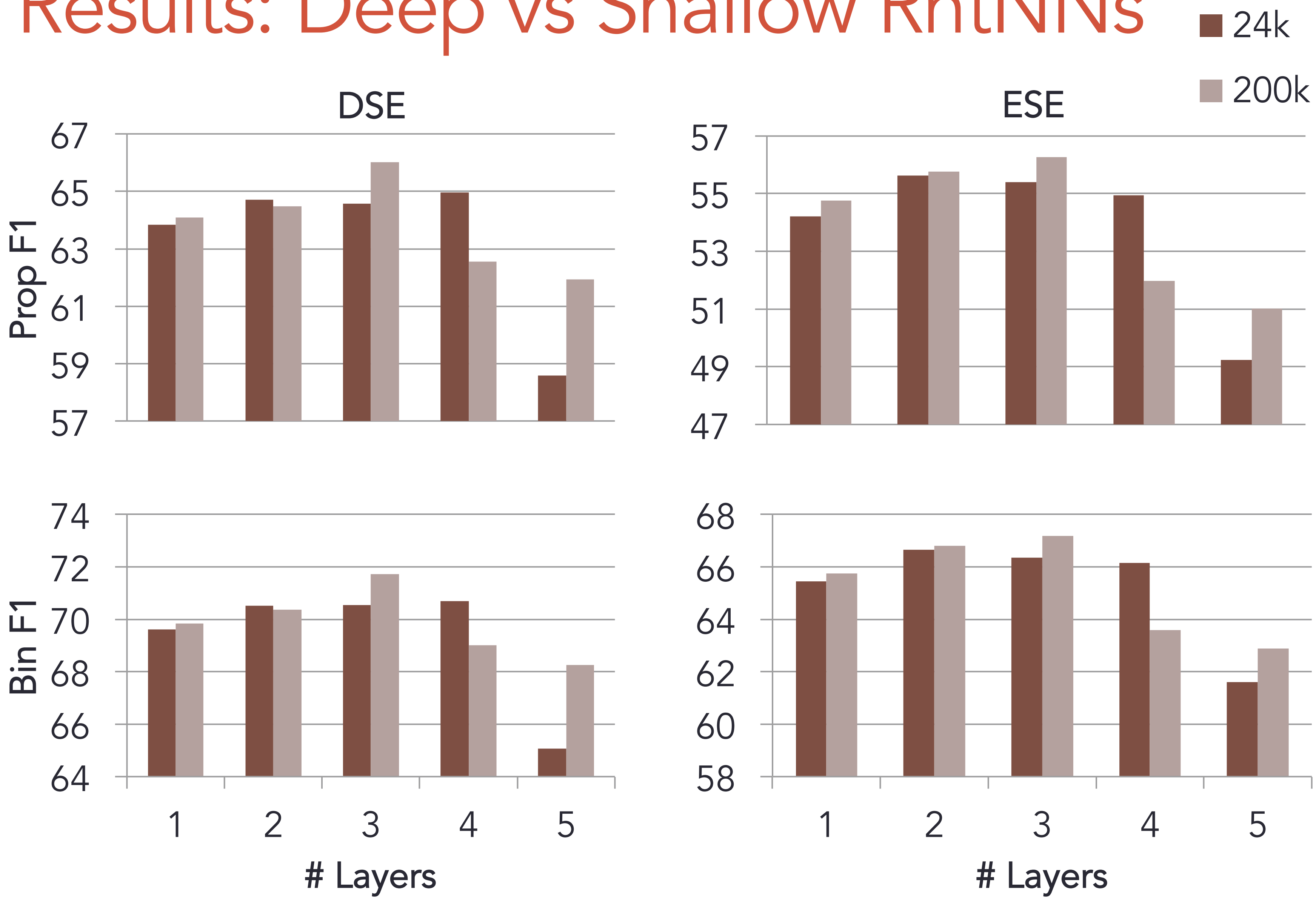
Hypotheses

We expected that deep recurrent nets would improve upon shallow recurrent nets, especially on ESE extraction.

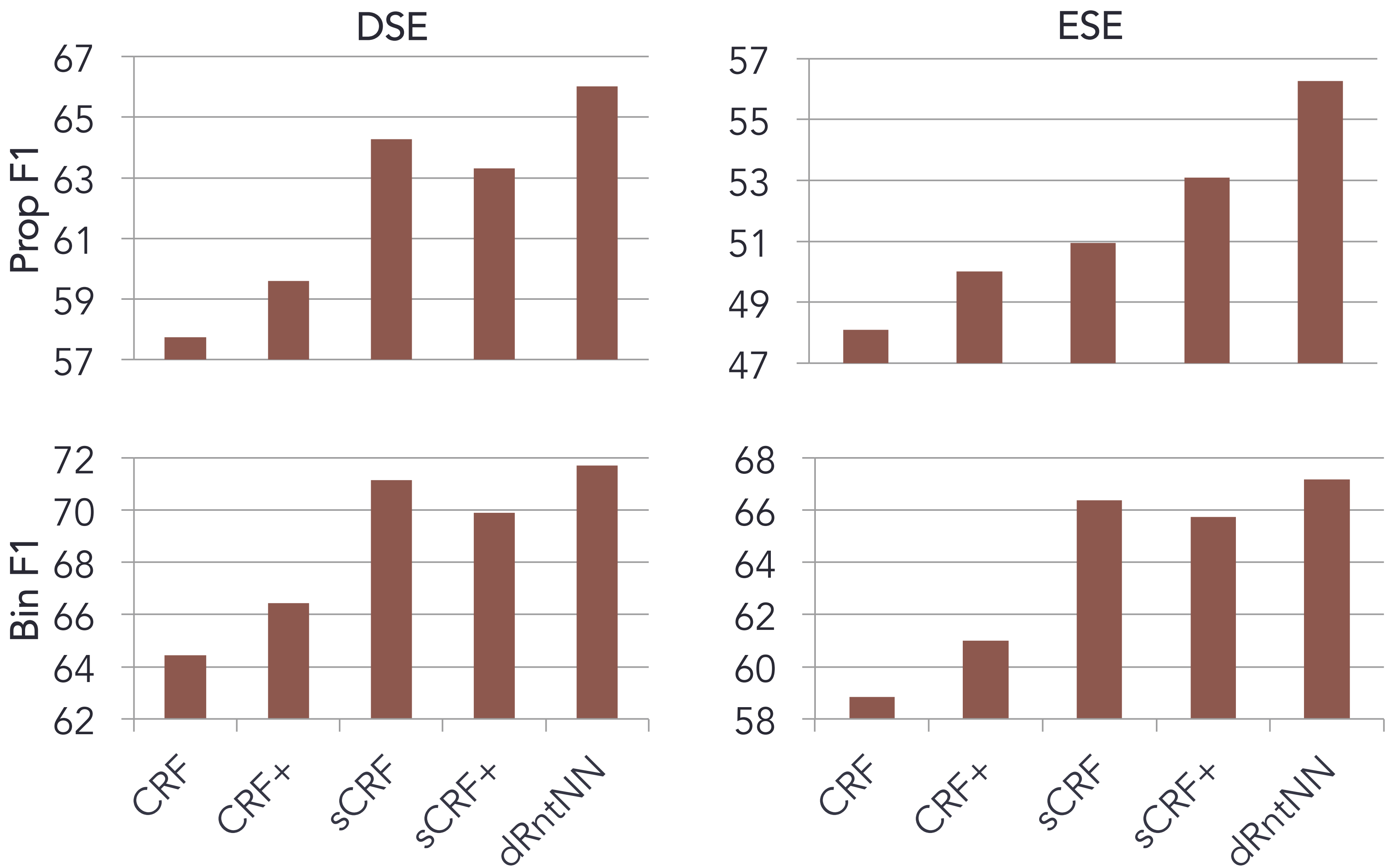
- ESEs are harder to identify: They are variable in length and might involve terms that are neutral in most contexts (e.g. “as usual”, “in fact”).

How the networks would perform against (semi)CRFs was unclear, especially when CRFs are given access to word vectors.

Results: Deep vs Shallow RntNNs



Results: Deep RntNN vs (semi)CRF



Results: Examples

True	The situation obviously remains fluid from hour to hour but it [seems to be] [going in the right direction]
Deep RntNN	The situation [obviously] remains fluid from hour to hour but it [seems to be going in the right] direction
Shallow RntNN	The situation [obviously] remains fluid from hour to hour but it [seems to be going in] the right direction
Semi-CRF	The situation [obviously remains fluid from hour to hour but it seems to be going in the right direction]

Results: Examples

True	have always said this is a multi-faceted campaign [but equally] we have also said any future military action [would have to be based on evidence], …
Deep RntNN	have always said this is a multi-faceted campaign but [equally we] have also said any future military action [would have to be based on evidence], …
Shallow RntNN	have always said this is a multi-faceted [campaign but equally we] have also said any future military action would have to be based on evidence, …
Semi-CRF	have always said this is a multi-faceted campaign but equally we have also said any future military action would have to be based on evidence, …

Results: Examples

True	[In any case], [it is high time] that a social debate be organized …
Deep RntNN	[In any case], it is [high time] that a social debate be organized …
Shallow RntNN	In [any] case, it is high [time] that a social debate be organized …
True	Mr. Stoiber [has come a long way] from his refusal to [sacrifice himself] for the CDU in an election that [once looked impossible to win], …
Deep RntNN	Mr. Stoiber [has come a long way from] his [refusal to sacrifice himself] for the CDU in an election that [once looked impossible to win], …
Shallow RntNN	Mr. Stoiber has come [a long way from] his refusal to sacrifice himself for the CDU in an election that [once looked impossible] to win, …

Conclusion (1)

- Deep recurrent nets perform better than their shallow counterparts of the same size on both DSE and ESE extraction.
- Both shallow and deep RntNNs capture aspects of subjectivity, but deep RntNNs seem to better handle the expression boundaries.
- Deep RntNNs outperforms previous baselines CRF and semi-CRF without having access to the dependency or constituency trees, opinion lexicons or POS tags, even when (semi)CRF has access to word vectors.

Rough Outline

- Deep recurrent neural networks
  Application to opinion mining

- Deep recursive neural networks
  Application to sentiment analysis

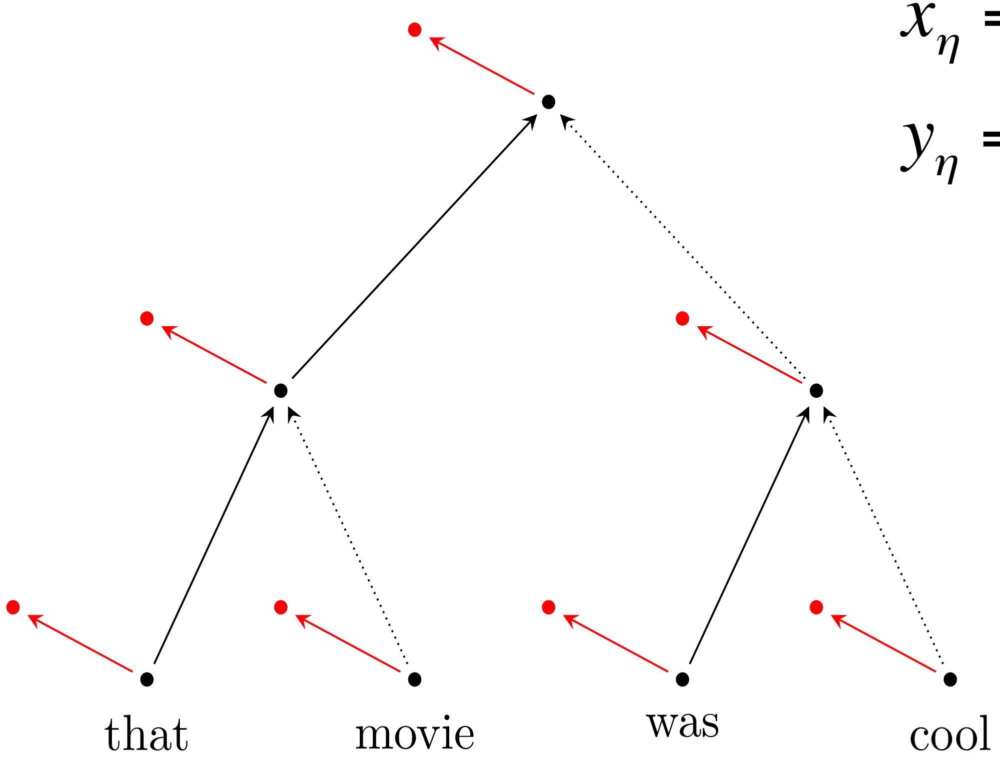
Structures

Sequences are not the only way to represent sentences in language.

Hierarchies such as dependency and constituency parse trees are widely used in the NLP literature, either as the main form of representing sentences, or to generate useful positional or relational features.

Trees provide a natural and intuitive way of composition.

Recursive Neural Networks



$$x_\eta = f(W_L x_{l(\eta)} + W_R x_{r(\eta)} + b)$$

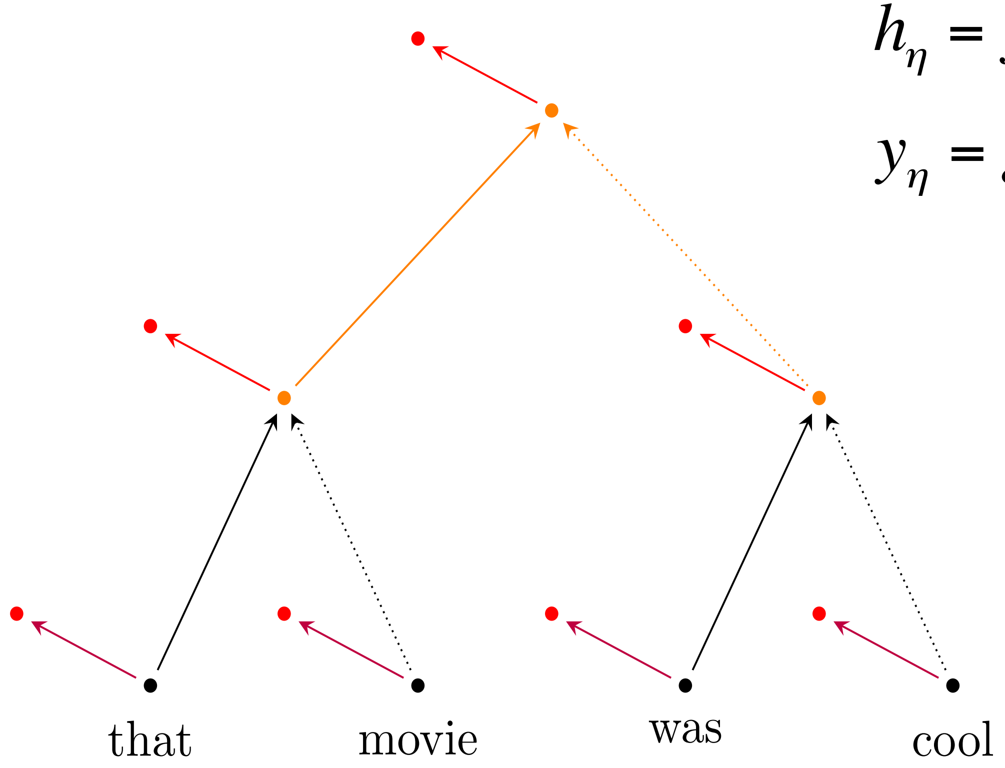
$$y_\eta = g(U x_\eta + c)$$

Representation at each node is a nonlinear transformation of the two children.

Untying Leaves from Internals

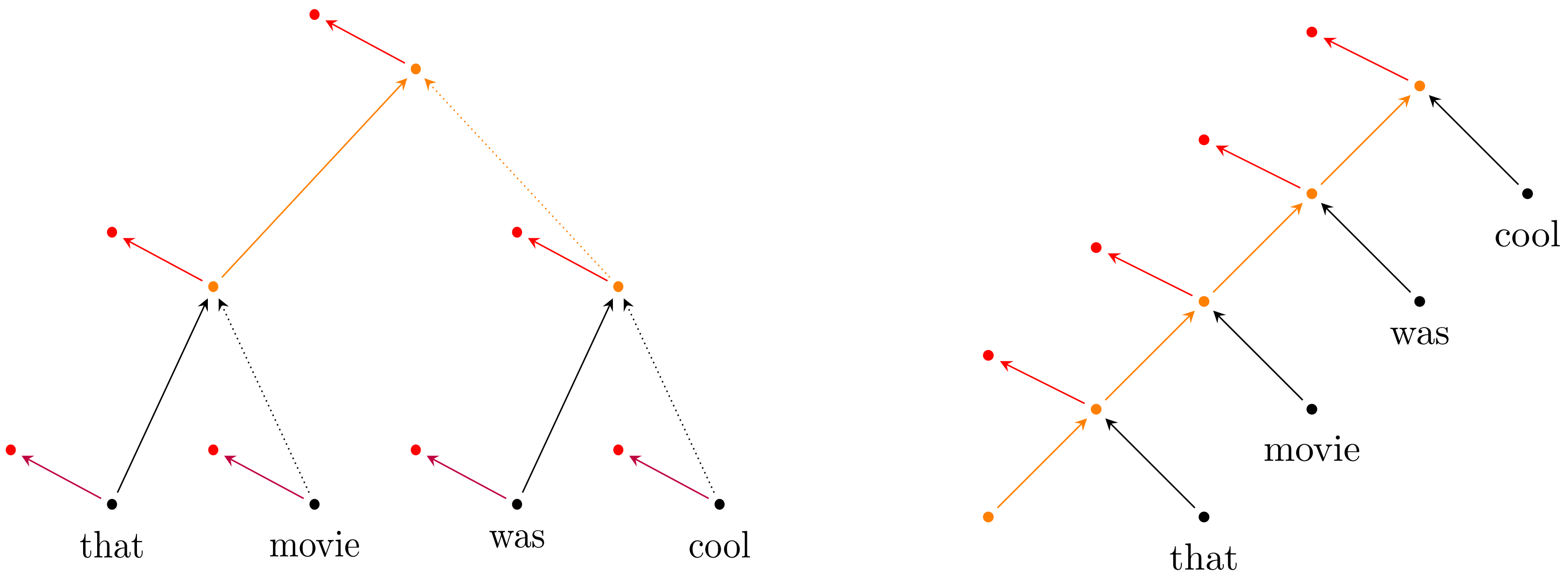
$$h_\eta = f(W_L^{l(\eta)} h_{l(\eta)} + W_R^{r(\eta)} h_{r(\eta)} + b)$$

$$y_\eta = g(U^{(\eta)} h_\eta + c)$$



Recursive connections are parametrized according to whether the incoming edge is from a leaf or an internal.

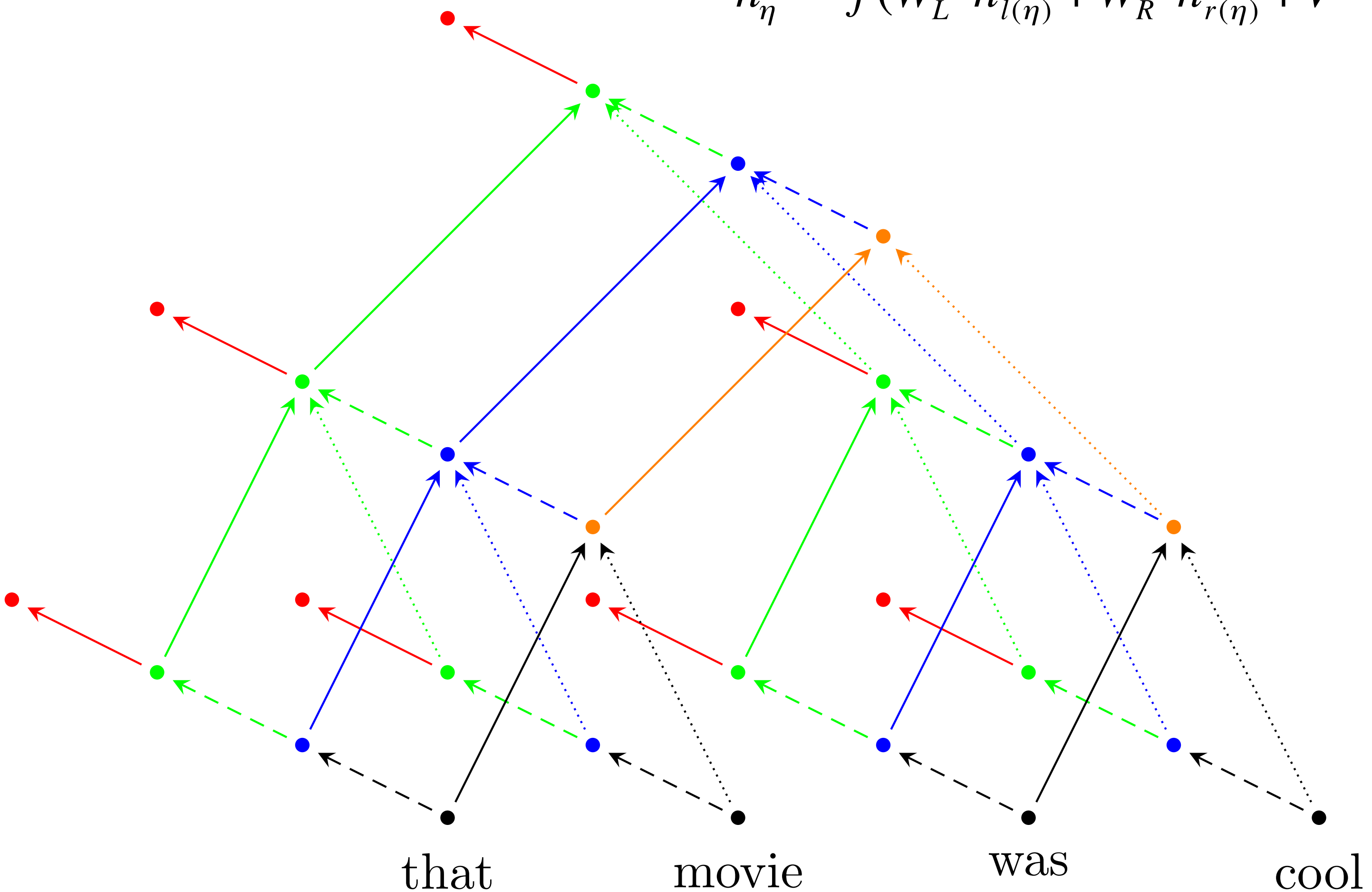
Untying Leaves from Internals



Recurrent neural net is actually a recursive neural net with a left-skewed tree structure.

Going Deep

$$h_\eta^{(i)} = f(W_L^{(i)} h_{l(\eta)}^{(i)} + W_R^{(i)} h_{r(\eta)}^{(i)} + V^{(i)} h_\eta^{(i-1)} + b)$$



Sentiment Analysis

Sentiment analysis aims to categorize contextual polarity of a given text (e.g. positive, negative or neutral).

Fine-grained sentiment analysis additionally aims to detect the intensity of emotions (e.g. positive, very positive). This essentially results in a finer grained sentiment classes.

Data

We use the Stanford Sentiment Treebank (Socher et al, 2013) that includes labels for 215,154 phrases in parse trees of 11,855 sentences. Labels are ordinal sentiment scores in {0, …, 4}.

Training-development-test partitioning of the data from the original work is used to evaluate the models.

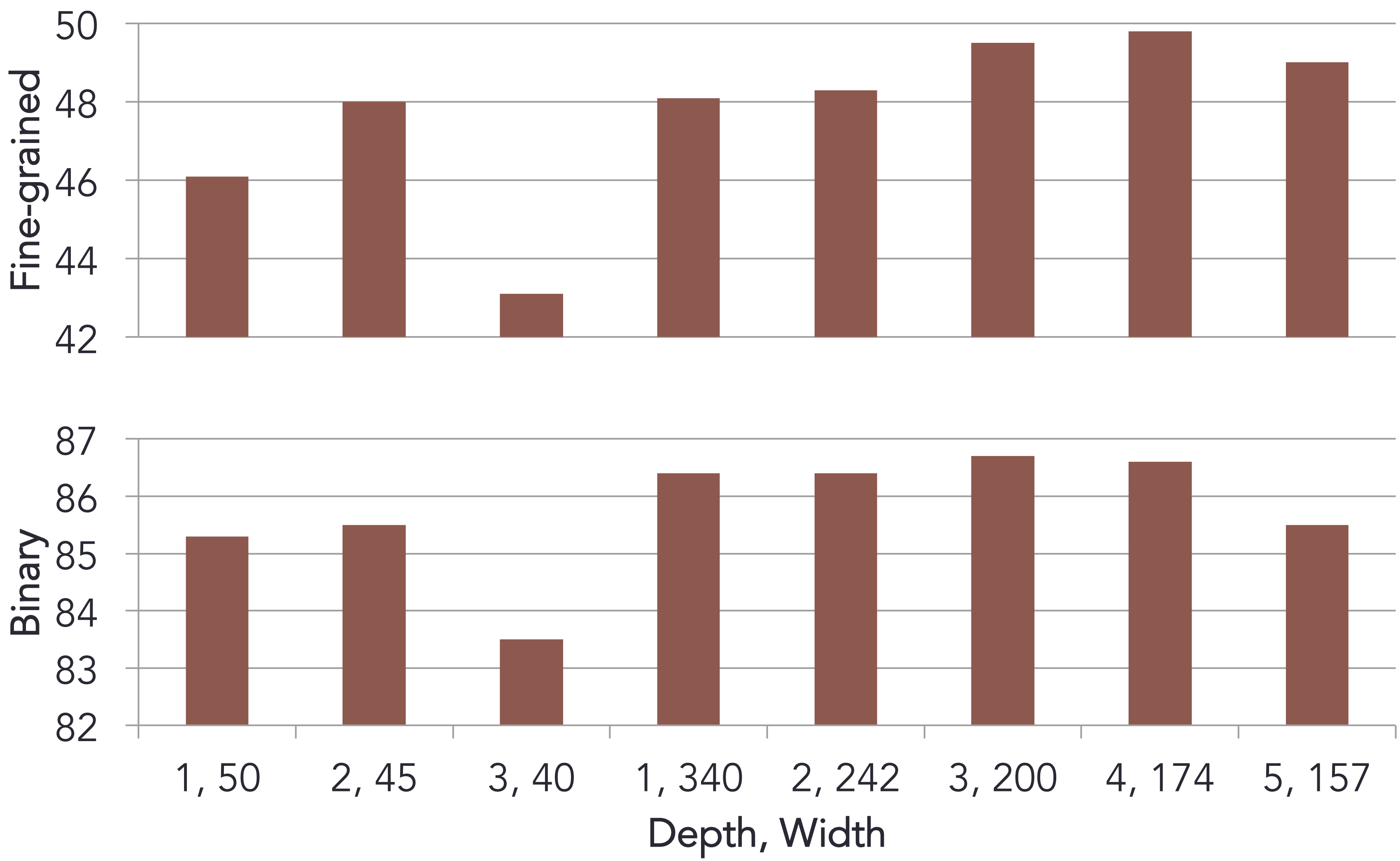
Network Training

- Softmax and rectifier nonlinearities are used for output and hidden layer activations, respectively.
- Dropout regularization.
- Stochastic gradient descent with Cross-Entropy classification objective.
- Model selection (with early stopping) is done over the development set.
- No pre-training, no fine-tuning.

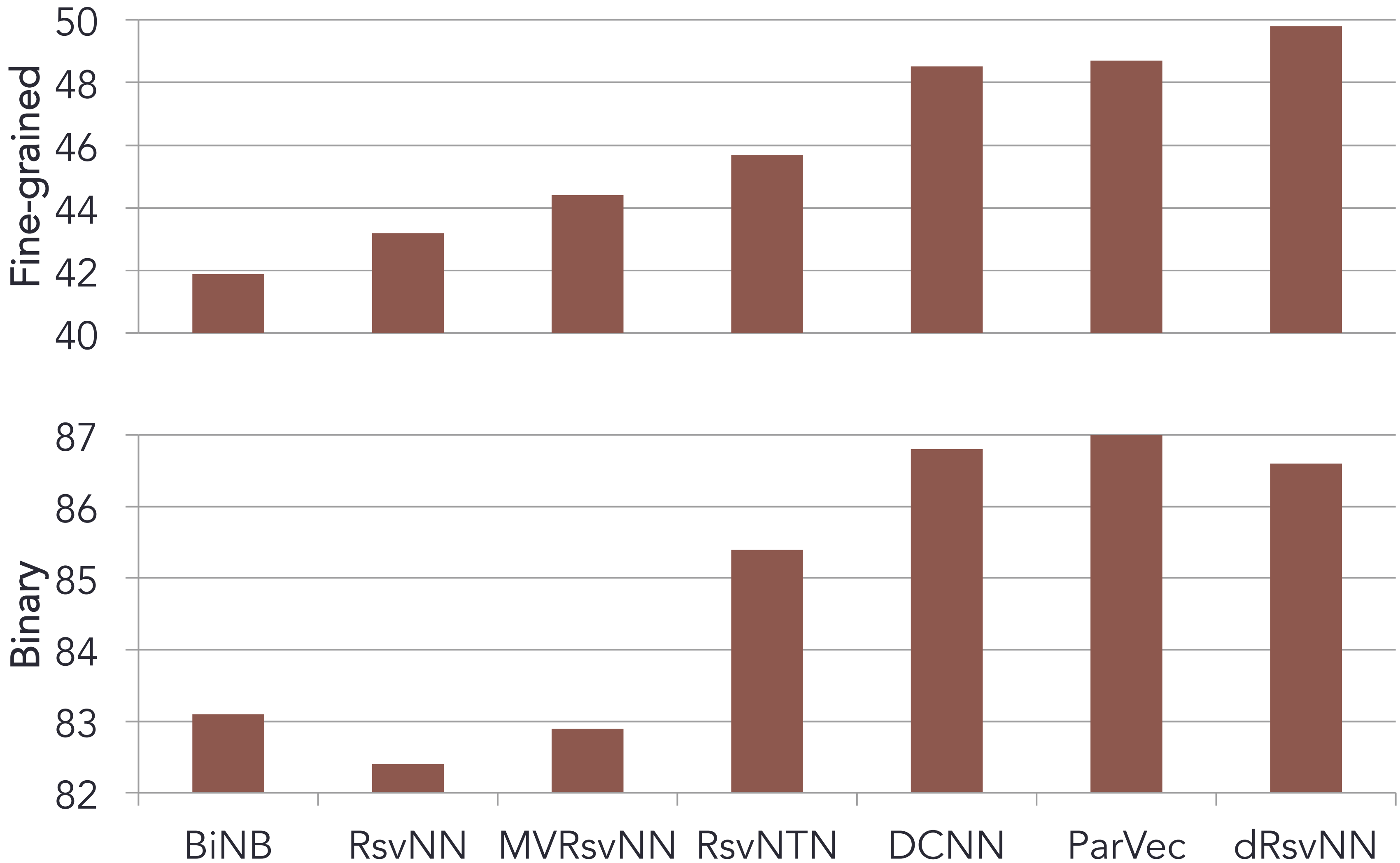
Baselines

- Baselines from (Socher et al, 2013):
  - Bigram Naïve Bayes (BiNB)
  - Recursive Net (RsvNN)
  - Matrix-Vector Recursive Net (MVRsvNN)
- Recursive Neural Tensor Network (RsvNTN) (Socher et al, 2013)
- Dynamic Convolutional Neural Network (DCNN) (Kalchbrenner et al, 2014)
- Paragraph Vectors (ParVec) (Le & Mikolov, 2014)

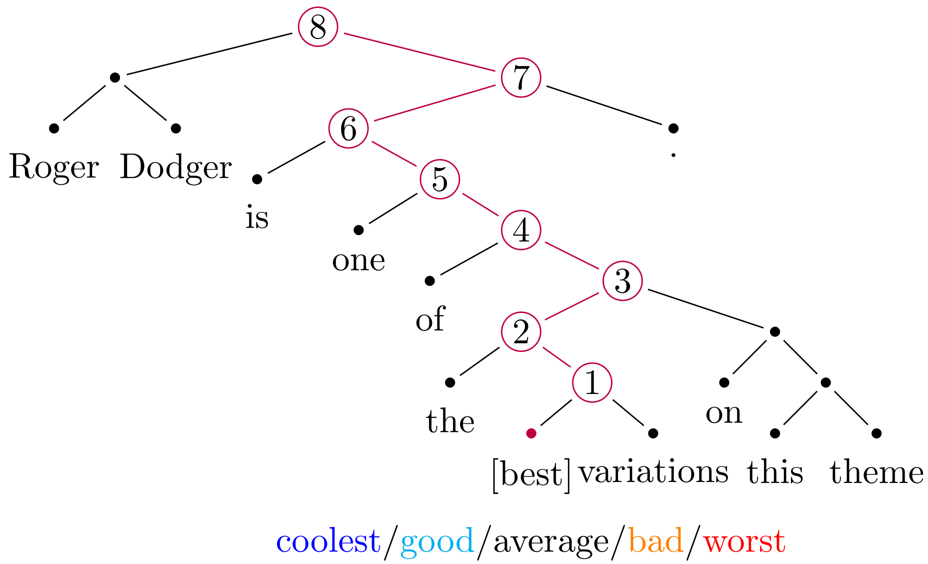
Results: Deep vs Shallow RsvNNs

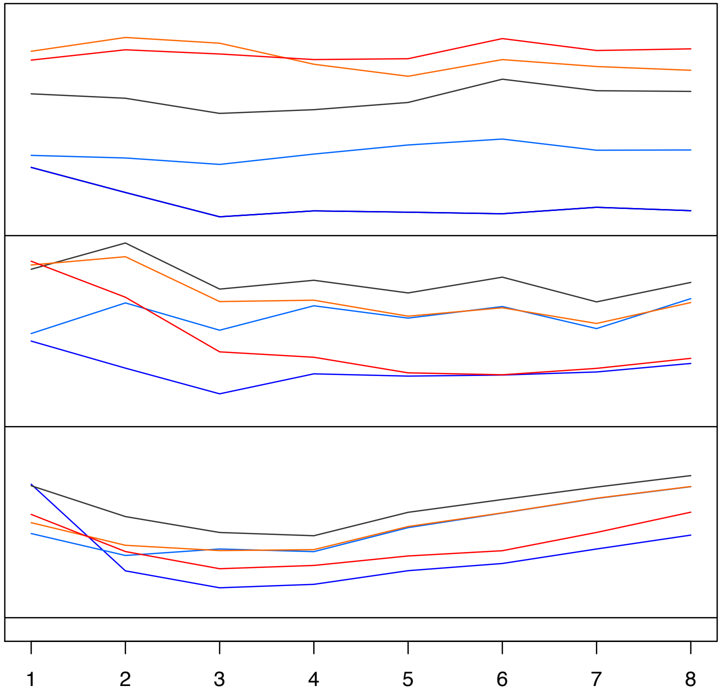


Results: RsvNN vs Baselines



Input Perturbation





Nearest Neighbor Phrases

	charming results	
charming ,	interesting results	charming chemistry
charming and	riveting performances	perfect ingredients
appealingly manic and energetic	gripping performances	brilliantly played
refreshingly adult take on adultery	joyous documentary	perfect medium
unpretentious , sociologically pointed	an amazing slapstick instrument	engaging film

Nearest Neighbor Phrases

	not great	
as great	nothing good	not very informative
a great	not compelling	not really funny
is great	only good	not quite satisfying
Is n't it great	too great	thrashy fun
be great	completely numbing experience	fake fun

Conclusion (2)

- Proposed deep recursive nets perform better than their shallow counterparts in fine-grained sentiment detection
- Additionally, deep recursive nets outperform existing baselines, achieving new state-of-the-art on the Stanford Sentiment Treebank
- Qualitative evaluations show that multiple layers indeed capture different things, they have different notions of similarity.

Future Work

- How does fine-tuning affect the performance?
- How do these models perform on tasks that require reasoning beyond sentences?

- Deeper questions:
  - What is going on behind the curtains of these deep nets in the context of NLP? Can we intuitively explain / visualize how they operate?
  - How do the differences across stacked layers manifest themselves? How is the hierarchy utilized?

Thanks!