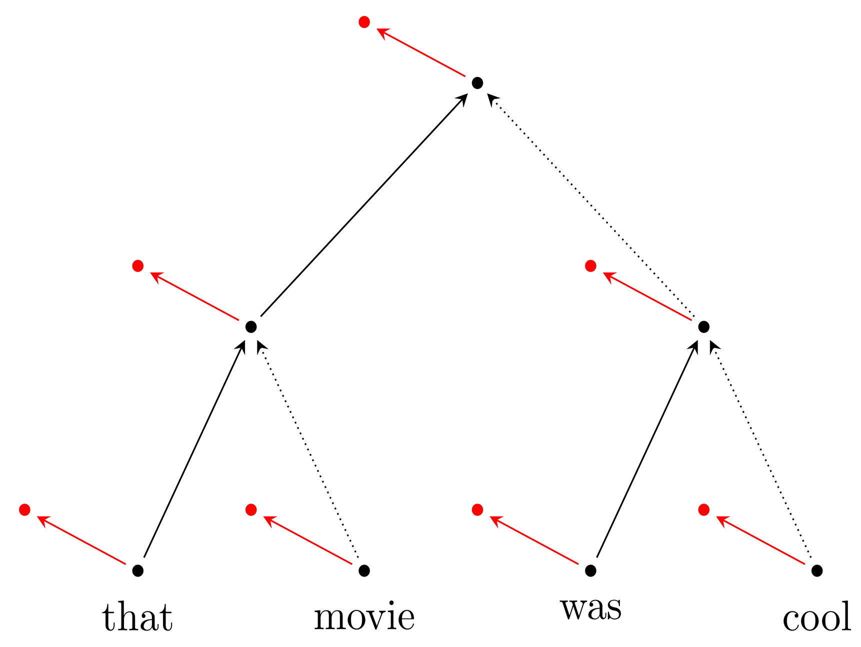
DEEP RECURSIVE NEURAL NETWORKS FOR COMPOSITIONALITY IN LANGUAGE

Ozan İrsoy
Claire Cardie
Cornell University

Recursive neural networks comprise a class of architecture that can operate on structured input. They have been previously successfully applied to model compositionality in natural language using parse-tree-based structural representations. Even though these architectures are deep in structure, they lack the capacity for hierarchical representation that exists in conventional deep feed-forward networks as well as in recently investigated deep recurrent neural networks. In this work we introduce a new architecture – *a deep recursive neural network* (deep RNN) – constructed by stacking multiple recursive layers. We evaluate the proposed model on the task of fine-grained sentiment classification. Our results show that deep RNNs outperform associated shallow counterparts that employ the same number of parameters. Furthermore, our approach outperforms previous baselines on the sentiment analysis task, including a multiplicative RNN variant as well as the recently introduced paragraph vectors, achieving new state-of-the-art results. We provide exploratory analyses of the effect of multiple layers and show that they capture different aspects of compositionality in language.

RECURSIVE NEURAL NETWORK

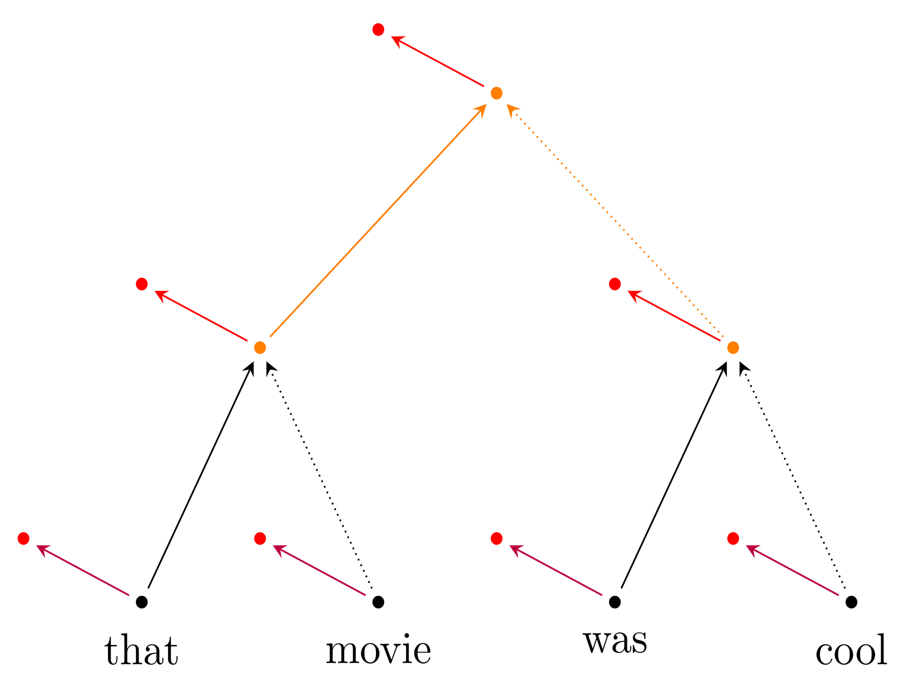


Representation at each node is a nonlinear transformation of the two children:

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

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

UNTIED RECURSIVE NET



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

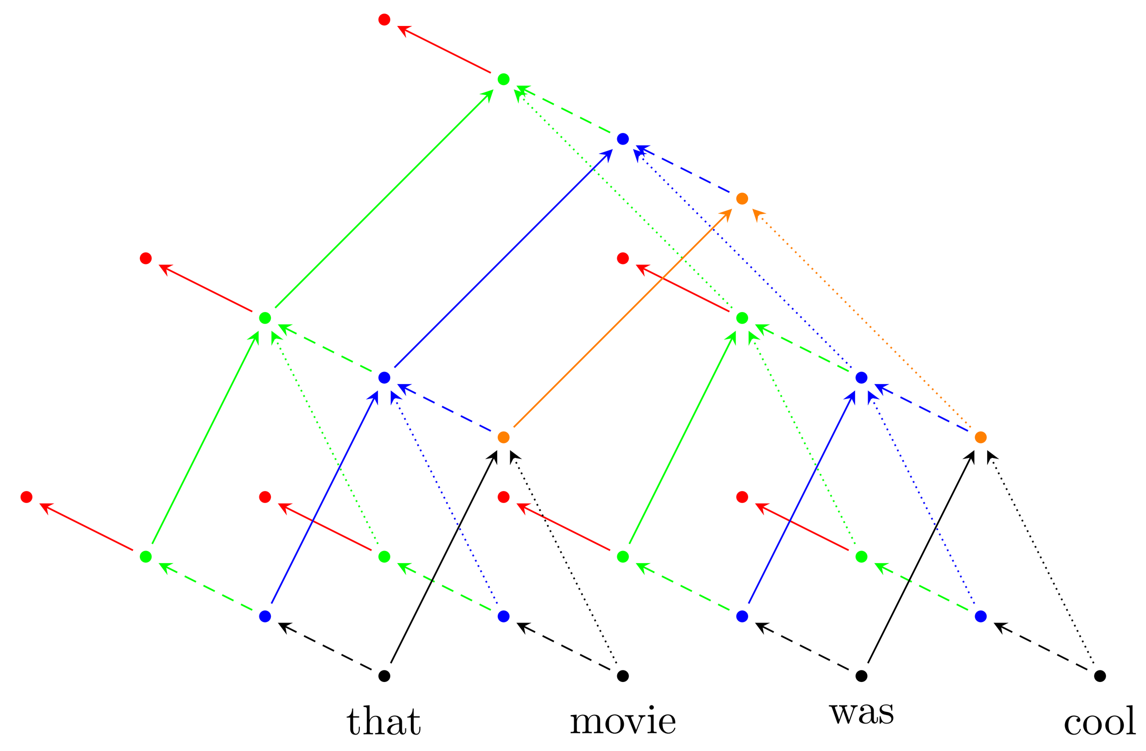
$$h_\eta = f(W_L^{l(\eta)} h_{l(\eta)} + W_R^{r(\eta)} h_{r(\eta)} + b)$$

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

Two advantages of untying:

1. Complexity is linear in *|x|*, not quadratic. Use small models with large word vectors
2. Activation of rectifiers is more natural: Word vectors are dense, hidden layers are sparse.

DEEP RECURSIVE NET



We construct deep recursive nets by stacking multiple recursive layer:

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

Each recursive layer processes a tree and passes it to the next layer.

NETWORK TRAINING

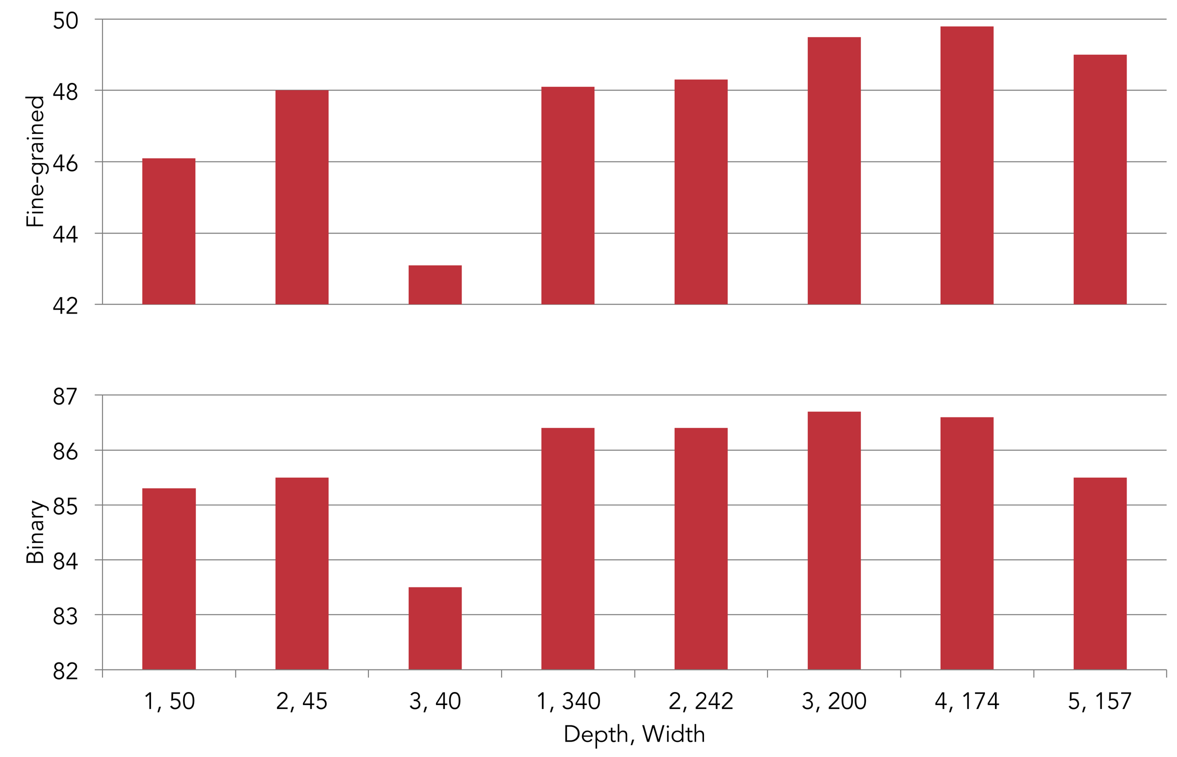
- Softmax and rectifier for output and hidden layer activations, respectively.
- Dropout regularization.
- Stochastic gradient descent with Cross-Entropy classification objective.
- Model selection (with early stopping) is done over the development set.
- No pre-training, no fine-tuning.
- Pretrained word2vec vectors.
- Stanford Sentiment Treebank as the data.

BASELINES

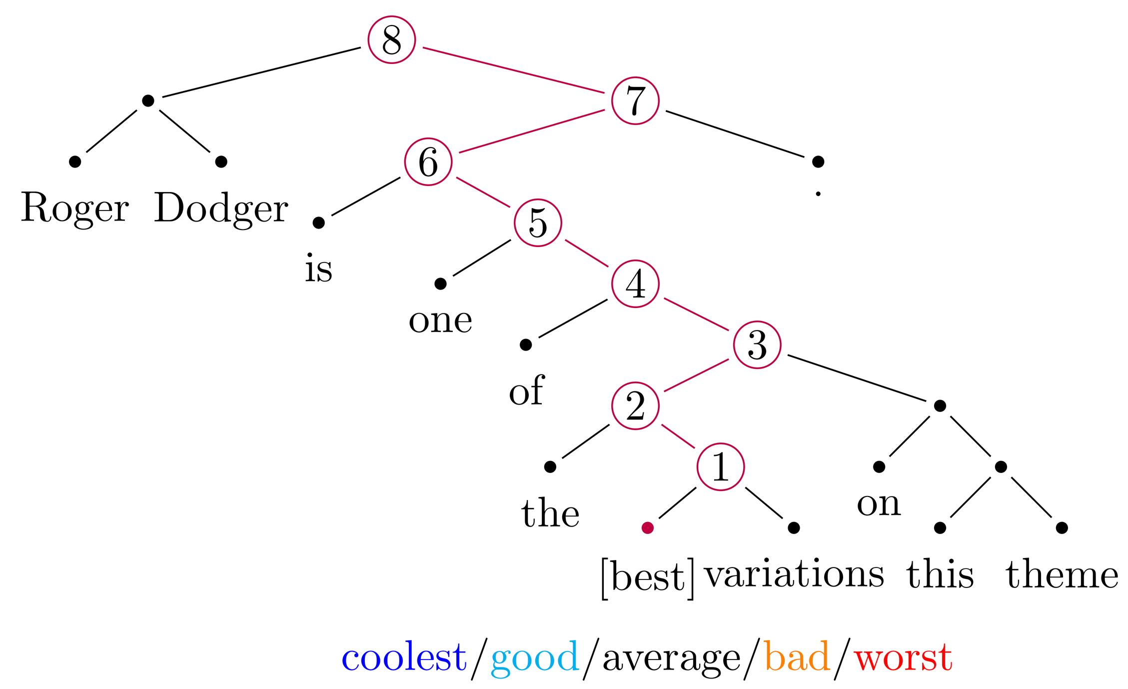
Baselines from (Socher et al, 2013):
- Bigram Naïve Bayes (BiNB),
- Recursive Net (RNN)
- Matrix-Vector Recursive Net (MVRNN)
- Recursive Neural Tensor Net (RNTN)

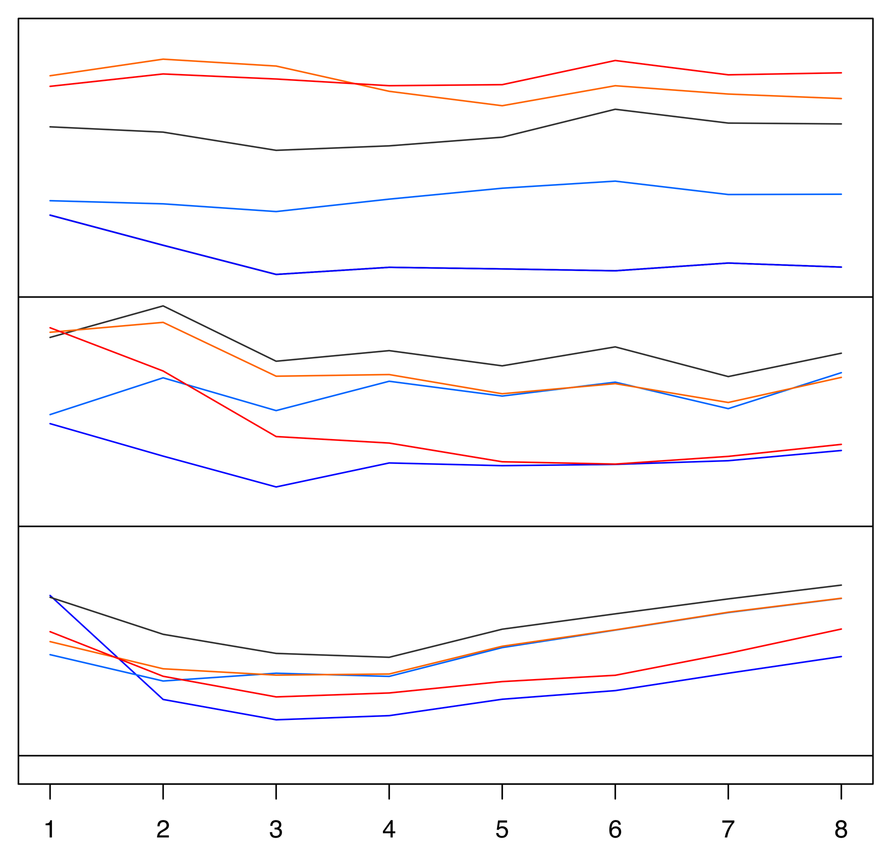
Dynamic Convolutional Net (DCNN) (Kalchbrenner et al, 2014)
ParagraphVectors (ParVec) (Le & Mikolov, 2014)

DEEP VS SHALLOW RNNS



INPUT PERTURBATION



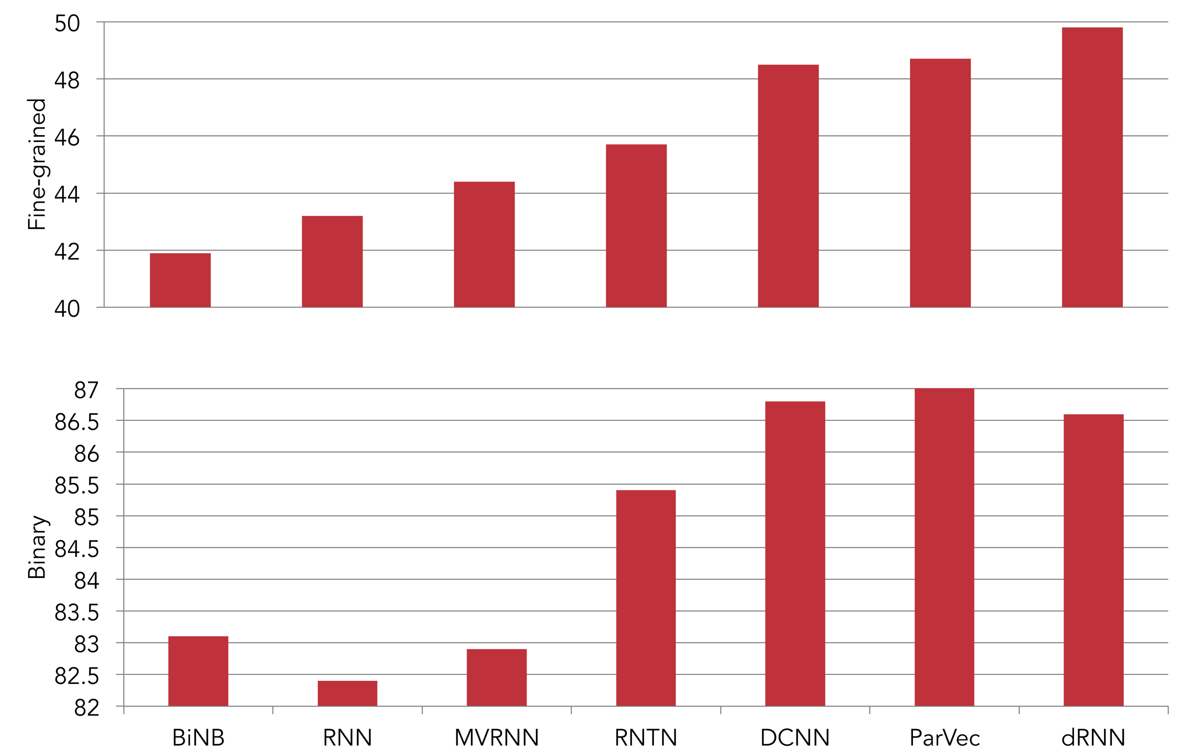


NEAREST PHRASE NEIGHBORS

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

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

DEEP RNN VS BASELINES



CONCLUSIONS

- Untying + pretrained word2vec + rectifier + dropout gives a boost (48.1 vs 43.2)
- Deep recursive nets perform better than their shallow counterparts in fine-grained sentiment detection
- 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.