Discriminative Neural Sentence Modeling by Tree-Based Convolution

Lili Mou,[1] Hao Peng,[1] Ge Li, Yan Xu, Lu Zhang, Zhi Jin

Software Institute, Peking University, P. R. China

EMNLP, Lisbon, Portugal
September, 2015

Outline

Outline

1. Introduction & Related Work
2. Tree-Based Convolution
   - c-TBCNN
   - d-TBCNN
3. Experimental Results
   - Experiment I: Sentiment Analysis
   - Experiment II: Question Classification
   - Model Analysis
4. Conclusion

Sentence Modeling

Sentence modeling

- To capture the meaning of a sentence
- Related to various tasks in NLP [Kalchbrenner et al., 2014]
  - Sentiment analysis
  - Paraphrase detection
  - Language-image matching

Our focus: *discriminative* sentence modeling

- Classify a sentence according to a certain criterion

An Example

Sentiment analysis

A movie review

An idealistic love story that brings out the latent 15-year-old romantic in everyone.

The sentiment?

☺☺ ☺ ⊖ ☹ ☹☹

Positive Neutral Negative

Feature Engineering

- Bag-of-words
- n-gram
- More dedicated ones, e.g.,[Silva et al., 2011]...

Problem: Sentence modeling is usually NON-TRIVIAL

Example [Socher et al., 2011]

`white blood cells destroying an infection`
`an infection destroying white blood cells`

Kernel Machines, e.g., SVM

+ Circumvent explicit feature representation
− Crucial to design the kernel function, which summarizes all data information

Neural networks

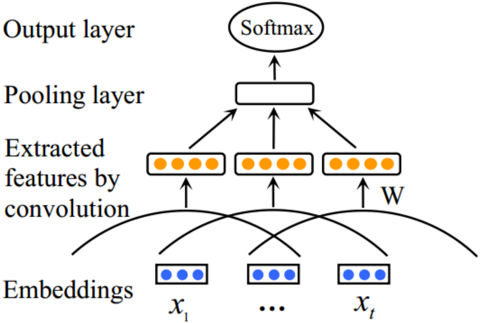
Automatic feature learning

- Word embeddings [Mikolov et al., 2013]
- Paragraph vectors [Le and Mikolov, 2014]

Prevailing neural sentence models

- Convolutional neural networks (CNNs) [Collobert and Weston, 2008]
- Recursive neural networks (RNNs) [Socher et al., 2011]
  - ✄ A variant: Recurrent neural networks

Convolutional Neural Networks (CNNs)



- Effective feature learning
- Unable to capture tree structural information

"Are tree structures necessary for deep learning of representations?"

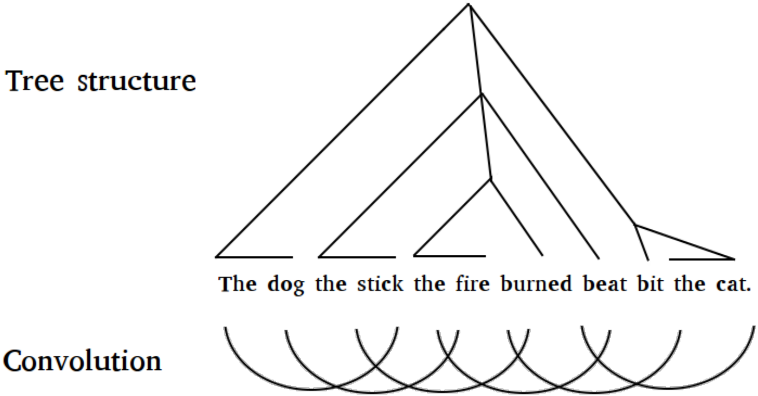
Example [Pinker, 1994]

The dog the stick the fire burned beat bit the cat.

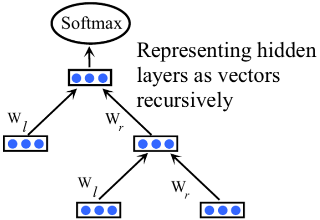
If if if it rains it pours I get depressed I should get help.

That that that he left is apparent is clear is obvious.

CNNs versus Sentence Structures

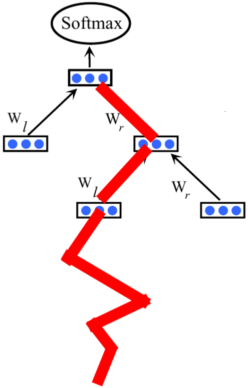


Recursive Neural Networks (RNNs)



+ Structure-sensitive
− Long propagation path

Long Propagation Path



- ☹ Burying illuminating information under complicated structure
- ☹ Gradient blowup or vanishing

Our Intuition

Can we combine the merits of CNNs and RNNs

- Having short propagation path like CNNs
- Capturing structure info like RNNs

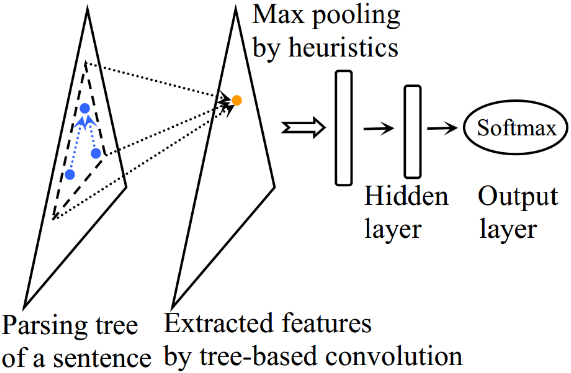
Our solution:

Tree-Based Convolutional Neural Network (TBCNN)

Outline

1. Introduction & Related Work
2. Tree-Based Convolution
   - c-TBCNN
   - d-TBCNN
3. Experimental Results
   - Experiment I: Sentiment Analysis
   - Experiment II: Question Classification
   - Model Analysis
4. Conclusion

Architecture of TBCNN



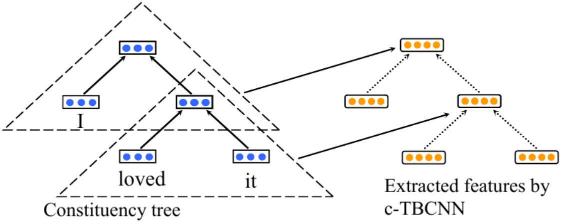
Technical Points

How to Represent nodes as vectors in consistency trees?

How to Handle nodes with different numbers of children in dependency trees?

How to Pool over varying sized and shaped structures?

c-TBCNN



- Pretrain an RNN and fix
- Perform convolution

E.g., A convolutional window of depth 2
i.e., a parent p with children l and r

$$\boldsymbol{y} = f\left(W_p^{(c)}\boldsymbol{p} + W_l^{(c)}\boldsymbol{c}_l + W_r^{(c)}\boldsymbol{c}_r + \boldsymbol{b}^{(c)}\right)$$

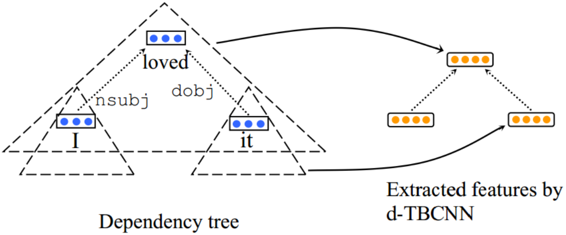
Remark on Complexity

- Exponential to the window depth
- Linear to the number of nodes

☑ Tree-based convolution does not add to complexity,

☐ But is less flexible than “flat” CNNs.

d-TBCNN



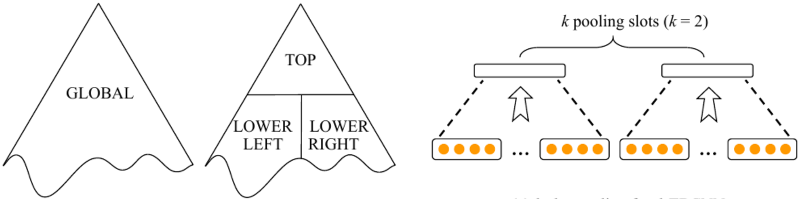
Associate weights with dependency types (e.g., `nsubj`, `dobj`) rather than positions

$$\boldsymbol{y} = f\left(W_p^{(d)}\boldsymbol{p} + \sum_{i=1}^{n} W_{r[c_i]}^{(d)}\boldsymbol{c}_i + \boldsymbol{b}^{(d)}\right)$$

$r[c_i]$: relation of between p and c_i

Pooling Heuristics

- Global pooling
- 3-slot pooling for c-TBCNN
- k-slot pooling for d-TBCNN



(a) Global pooling (b) 3-slot pooling for c-TBCNN (c) k-slot pooling for d-TBCNN

Outline

1. Introduction & Related Work
2. Tree-Based Convolution
   - c-TBCNN
   - d-TBCNN
3. Experimental Results
   - Experiment I: Sentiment Analysis
   - Experiment II: Question Classification
   - Model Analysis
4. Conclusion

Sentiment Analysis

Dataset

- Stanford sentiment tree bank
- 5 labels: $++/+/0/-/--$
- 8544/1101/2210 sentences, ~150k phrases

Our settings

- 5-way classification + binary classification
- Training: sentences + phrases
- Testing: sentences only

Data samples	Label
Offers that rare combination of entertainment and education.	$++$
An idealistic love story that brings out the latent 15-year-old romantic in everyone.	$+$
Its mysteries are transparently obvious, and it's too slowly paced to be a thriller.	$-$

Group	Method	5-class accuracy	2-class accuracy
Baseline	SVM	40.7	79.4
	Naïve Bayes	41.0	81.8
CNNs	1-layer convolution	37.4	77.1
	Deep CNN	48.5	86.8
	Non-static	48.0	87.2
	Multichannel	47.4	**88.1**
RNNs	Basic	43.2	82.4
	Matrix-vector	44.4	82.9
	Tensor	45.7	85.4
	Tree LSTM	51.0	88.0
	Deep RNN	49.8	86.6†
Recurrent	LSTM	45.8	86.7
	bi-LSTM	49.1	86.8
Vector	Word vector avg.	32.7	80.1
	Paragraph vector	48.7	87.8
TBCNNs	c-TBCNN	50.4	86.8†
	d-TBCNN	**51.4**	87.9†

Group	Method	5-class accuracy	2-class accuracy
Baseline	SVM	40.7	79.4
	Naïve Bayes	41.0	81.8
CNNs	1-layer convolution	37.4	77.1
	Deep CNN	48.5	86.8
	Non-static	48.0	87.2
	Multichannel	47.4	**88.1**
RNNs	Basic	43.2	82.4
	Matrix-vector	44.4	82.9
	Tensor	45.7	85.4
	Tree LSTM	51.0	88.0
	Deep RNN	49.8	86.6†
Recurrent	LSTM	45.8	86.7
	bi-LSTM	49.1	86.8
Vector	Word vector avg.	32.7	80.1
	Paragraph vector	48.7	87.8
TBCNNs	c-TBCNN	50.4	86.8†
	d-TBCNN	**51.4**	87.9†

Question Classification

Dataset

- 5452 training + 500 test
- Labels
    - abbreviation
    - entity
    - description
    - human
    - location
    - numeric

Data samples	Label
What is the temperature at the center of the earth?	number
What state did the Battle of Bighorn take place in?	location

Results

Method	Acc. (%)	Reported in
SVM 10k features + 60 rules	95.0	[Silva et al., 2011]
CNN-non-static	93.6	[Kim, 2014]
CNN-mutlichannel	92.2	[Kim, 2014]
RNN	90.2	[Zhao et al., 2015]
Deep-CNN	93.0	[Kalchbrenner et al., 2014]
Ada-CNN	92.4	[Zhao et al., 2015]
c-TBCNN	94.8	Our implementation
d-TBCNN	**96.0**	Our implementation

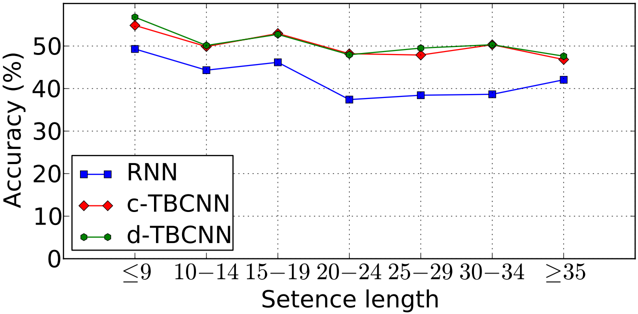
Model Analysis: Pooling Methods

Model	Pooling method	5-class accuracy (%)
c-TBCNN	Global	48.48 ± 0.54
	3-slot	48.69 ± 0.40
d-TBCNN	Global	49.39 ± 0.24
	2-slot	49.94 ± 0.63

Remarks

- Averaged over 5 random initializations
- Hyperparameters predefined, less optimal

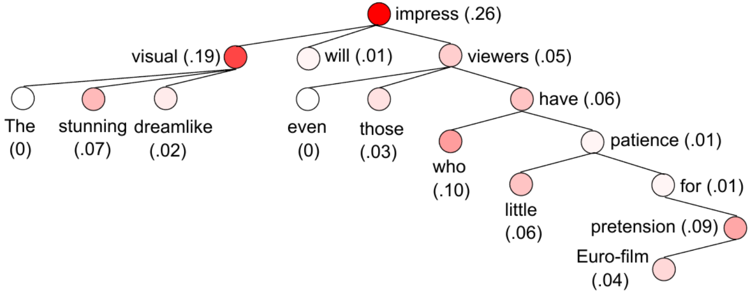
Model Analysis: Sentence Length



Reimplemented RNN: 42.7% accuracy, slightly lower than 43.2% reported in [Socher et al., 2011]

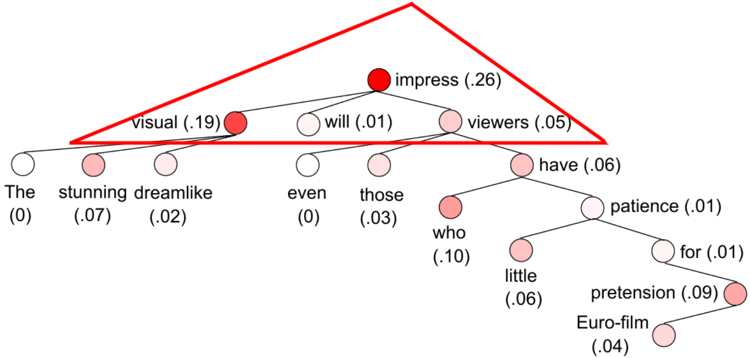
Visualization

"The stunning dreamlike visual will impress even those who have little patience for Euro-film pretension."



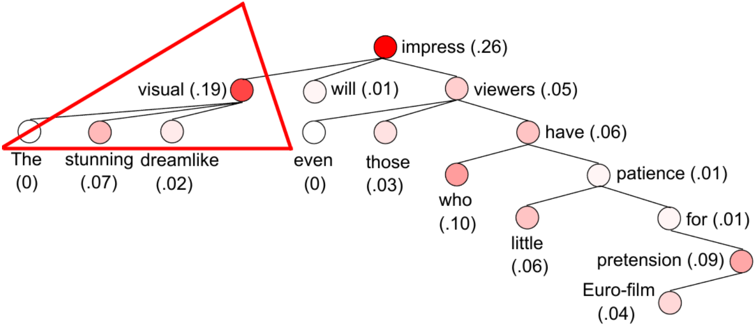
Visualization

"The stunning dreamlike visual will impress even those who have little patience for Euro-film pretension."



Visualization

"The stunning dreamlike visual will impress even those who have little patience for Euro-film pretension."



Outline

Conclusion

		Way of information propagation	
		Iterative	Sliding
Structure	Flat	Recurrent	Convolution
	Tree	Recursive	Tree-based convolution

Thank you for listening!

Q & A

References

Collobert, R. and Weston, J. (2008).
A unified architecture for natural language processing: Deep neural networks with multitask learning.
In *Proceedings of the 25th International Conference on Machine learning*.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014).
A convolutional neural network for modelling sentences.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Kim, Y. (2014).
Convolutional neural networks for sentence classification.

Le, Q. and Mikolov, T. (2014).
Distributed representations of sentences and documents.
In *Proceedings of the 31st International Conference on Machine Learning*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
In *Advances in Neural Information Processing Systems*.

Pinker, S. (1994).
The Language Instinct: The New Science of Language and Mind.
Pengiun Press.

Silva, J., Coheur, L., Mendes, A., and Wichert, A. (2011).
From symbolic to sub-symbolic information in question classification.
Artificial Intelligence Review, 35(2):137–154.

Socher, R., Pennington, J., Huang, E., Ng, A., and Manning, C. (2011).
Semi-supervised recursive autoencoders for predicting sentiment distributions.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Zhao, H., Lu, Z., and Poupart, P. (2015).
Self-adaptive hierarchical sentence model.
arXiv preprint arXiv:1504.05070, to appear in Proceedints of Intenational Joint Conference in Artificial Intelligence.