NN architectures for sentiment analysis

... and other NLP tasks.

Christian Weilbach

November 7, 2016

Universität Heidelberg

Overview

Motivation

Motivation

Problem: Difficult movie review classification as basis. (Pang and Lee 2004a)

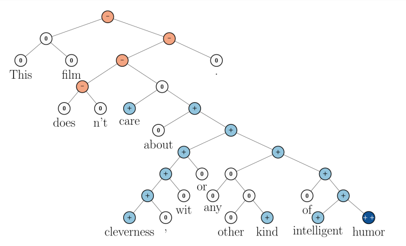


Figure 1: Parse tree with sentiment labels, Socher et al. 2013a.

- **Linguistic**: Take syntactic structure into account, compared to *Bag-of-Words* attempts or shallow syntactic features.
- $\Rightarrow$ take *semantic compositionality* into account

Demo

Related Works

Compositionality

- Some prior work on compositionality in Vector Space , i.e. by Matrices (Yessenalina and Cardie 2011), for instance as grammatical reductions in *Lambek pregroup grammar* (Grefenstette and Sadrzadeh 2011)
- **Logical Forms**: *but* cannot themselves capture sentiment

Sentence:	what states border texas
Logical Form:	$\lambda x.state(x) \wedge borders(x, texas)$

Figure 2: Mapping a sentence to its logical form.

NN for NLP

Neural Networks...

Figure 3: CC-BY-SA, author Alan Chi

NN Background

- Architecture
- Inference
- Learning

Architecture

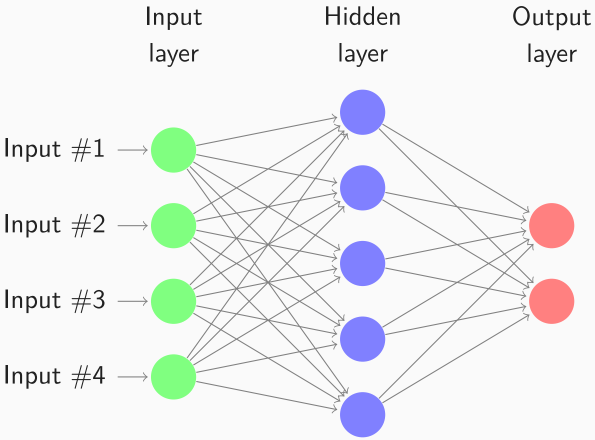


Figure 4: A small feedforward neural network.

Inference

Explanation of non-linear transformation in terms of topology: `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`

Learning - Backpropagation

- Essentially application of *chain rule*.
- but *very efficient*
- can be calculcated automatically if derivatives for each composed function are known

Pros & Cons

- Simple primitives: linear algebra $+$ point-wise non-linearities
- Function composition: flexibility
- end-to-end training of joint objective possible (less feature engineering)
- Toolboxes with efficient optimization available

- Arbitrarily designed architecture
- Non-Convex optimization
  - dependent on initialization!
  - finds local optimum
  - computationally expensive optimization
  - *a lot* of data needed

  $\Rightarrow$ cannot be used as blackbox (like e.g. SVM)
- No direct probabilistic/statistical interpretation

Word embedding

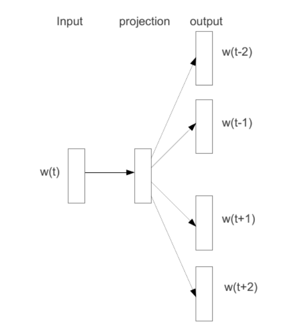


Figure 5: Skip-gram word embedding, Mikolov et al. 2013.

- every word gets *random* initialized vector
- trained *jointly* with the neural models
- can be understood as factorization of PMI matrix (Levy and Goldberg 2014)
- but in the following: *learned* **sentiment vectors** (not pretrained w2v)

Recurrent NNs

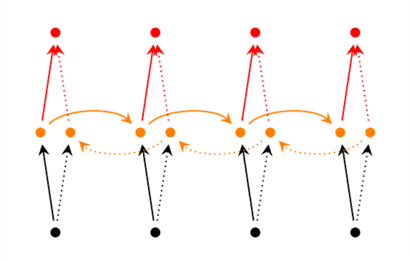


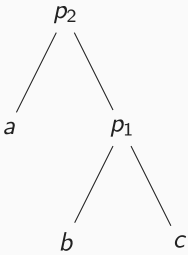
Figure 6: Bidirectional RNN, Irsoy and Cardie 2014.

- Outperforms previous CRF-based approaches **without!** lexicons and syntactic features on MPQA Wiebe corpus for DSE, ESE classification

Recursive NNs

- ***Recursive*** **NN** $\neq$ ***Recurrent*** **NN** (see Bengio and Courville 2016 for detailed discussion)
- Idea: *recursively* merge the parse tree

$$
\begin{aligned}
p_1 &= f(\hat{p}_1), && (1)\\
\hat{p}_1 &= W \begin{bmatrix} b \\ c \end{bmatrix} && (2)\\
p_2 &= f(\hat{p}_2), && (3)\\
\hat{p}_2 &= W \begin{bmatrix} a \\ p_1 \end{bmatrix} && (4)\\
f(x) &= \tanh(x) \text{ (choice)} && (5)\\
W &\in \mathbb{R}^{d \times 2d} && (6)
\end{aligned}
$$



Recursive Neural Network models for sentiment analysis

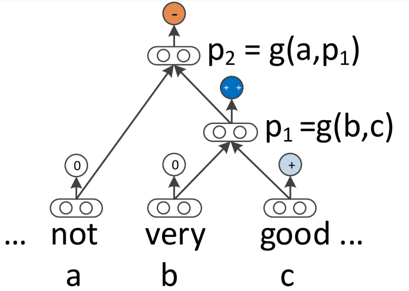


Figure 7: Sentiment labels in recursive architecture, Socher et al. 2013a.

Sentiment (++,+,o, -,–) at each node is extracted with a softmax function.

$$y_a = \text{softmax}(W_s a), \tag{7}$$

$$W_s \in \mathbb{R}^{5xd} \tag{8}$$

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{5} \exp(x_j)} \tag{9}$$

$$\delta^{y^a} = ((y^a - t^a)^T W_s) \otimes f'(x^a) \tag{10}$$

Recursive backpropagation through structure

$$\delta^{p2,com(bined)} = \underbrace{\delta^{y^{p2}}}_{\text{from classifier for } p_2} \tag{11}$$

$$(\delta^{p2,down})^T = (\delta^{p2,com})^T W \quad \underbrace{\otimes f'(\hat{p}_2)}_{\text{elem. deriv. for } f} \quad , \tag{12}$$

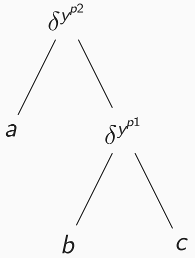
$$\delta^{p1,com} = \delta^{p2,down}_{[d+1:2d]} + \delta^{y^{p1}} \tag{13}$$

$$(\delta^{p1,down})^T = (\delta^{p1,com})^T W \otimes f'(\hat{p}_1)\,, \tag{14}$$

$$\Rightarrow \delta W = \underbrace{\delta^{y^{p2}} \begin{bmatrix} a & p_1 \end{bmatrix}^T + \delta^{p1,com} \begin{bmatrix} b & c \end{bmatrix}^T}_{\text{total derivative on } all \text{ nodes}} \tag{15}$$

$$\delta^b = \delta^{p1,down}_{[1:d]} + \delta^{y^b} \tag{16}$$

$$\delta^{y^{p1},y^{p2},y^b} \in \mathbb{R}^d \tag{17}$$



M(atrix)V(ector)-RNN

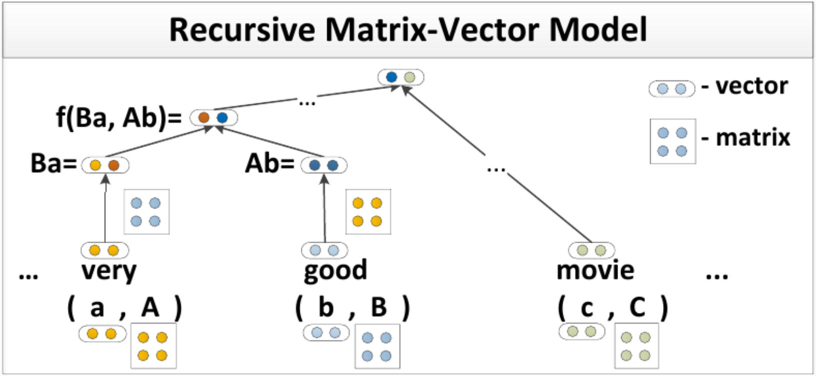


Figure 8: Example MV-RNN architecture, Socher, Huval, et al. 2012.

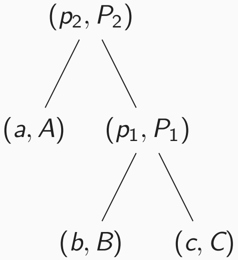
MV-RNN

- Track additional matrix for each word
- Intuition: Learn context matrix for other words

$$p_1 = f\left(W \begin{bmatrix} Cb \\ Bc \end{bmatrix}\right), \tag{18}$$

$$P_1 = f\left(W_M \begin{bmatrix} C \\ B \end{bmatrix}\right), \tag{19}$$

$$W, W_M \in \mathbb{R}^{d \times 2d} \tag{20}$$



MV-RNN Summary

- in RNN input vectors only *implicitly(?)* interact
- powerful composition function with fixed number of parameters desired
- (Matrix, Vector) tuple for every word $\Rightarrow$ very large parameter space $\Rightarrow$ uses low-rank matrix approximation $A = UV + \mathrm{diag}(a)$, $\mathrm{rank}(UV) = 3$
- builds on idea of compositionality with matrix
- operator words (e.g. *extremely*): full matrix, vector $= 0$ *vs* non-operator words: identity matrix, feature-rich vector
- *outperformed* previous models in 2011
- can learn **propositional logic** as composition Socher, Huval, et al. 2012

Recursive Neural Tensor Network

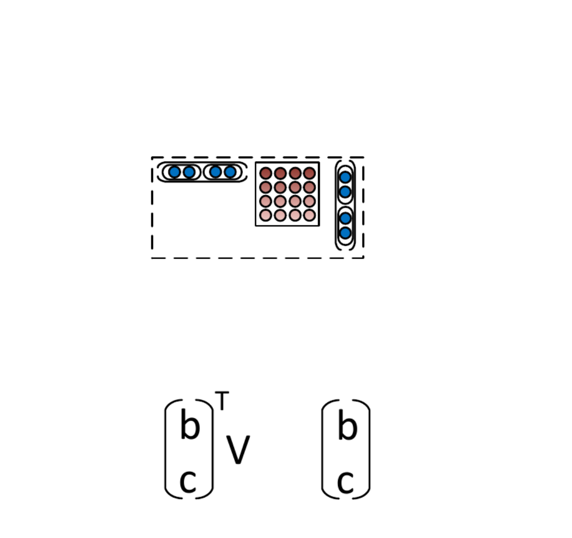


Figure 9: Tensor interaction in an RNTN, Socher et al. 2013a.

RNTN

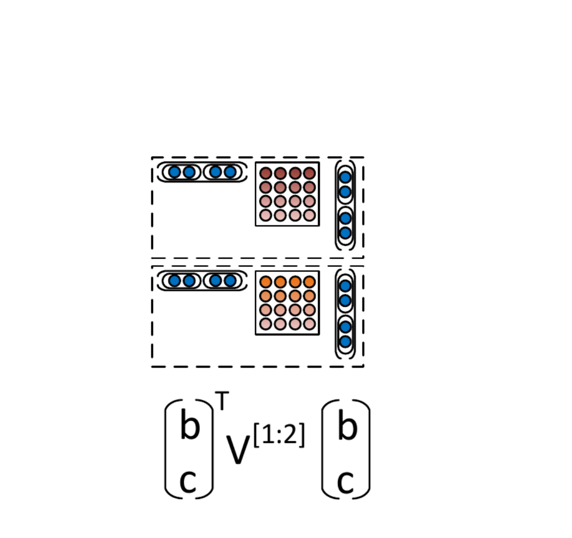


Figure 10: Tensor interaction in an RNTN, Socher et al. 2013a.

RNTN

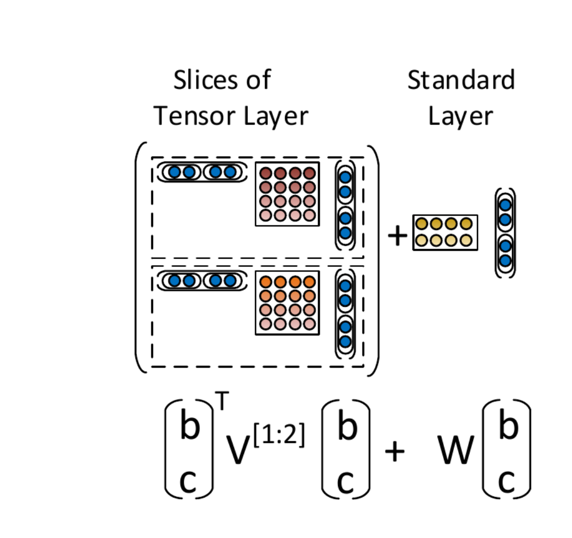


Figure 11: Tensor interaction in an RNTN, Socher et al. 2013a.

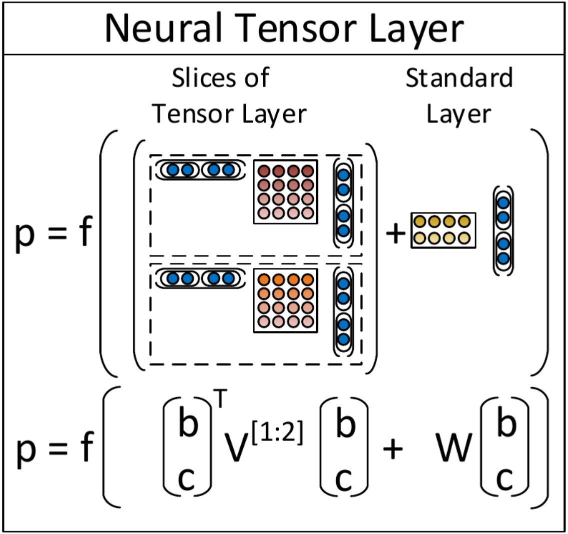


Figure 12: Tensor interaction in an RNTN, Socher et al. 2013a.

RNTN

$$p_1 = f\left(\begin{bmatrix} b \\ c \end{bmatrix}^T V^{1:d} \begin{bmatrix} b \\ c \end{bmatrix} + W \begin{bmatrix} b \\ c \end{bmatrix}\right), \tag{21}$$

$$V \in \mathbb{R}^{2d \times 2d \times d}, W \in \mathbb{R}^{d \times 2d} \tag{22}$$

RNTN Loss function

$$E(\theta) = \sum_{i \in \text{data}} \sum_{j \in \text{labels}} t_j^i \log y_j^i + \lambda \|\theta\|^2 \tag{23}$$

- categorical *cross-entropy* as cost function
- **minimizes** Kullback-Leibler divergence
  $\Rightarrow$ distance between label distribution and prediction
- $\lambda \|\theta\|^2$ is regularizer over the weights (prior)

RNTN learning (backpropagation through structure)

Gradient for tensor product V needed, $\frac{d}{dW}E$ comes from standard NN. Recall that we only need the derivatives of the inner product for each tensor node to backpropagate errors:

$$\frac{\partial E}{\partial V^{[k]}} = \delta_k^{p_2,com} \begin{bmatrix} a \\ p_1 \end{bmatrix} \begin{bmatrix} a \\ p_1 \end{bmatrix}^T \tag{24}$$

$$\delta^{p_2,down} = (\delta^{p_2,com} W + S) \otimes f' \left(\begin{bmatrix} \hat{a} \\ \hat{p}_1 \end{bmatrix} \right) \tag{25}$$

$$S = \sum_{k=1}^{d} \delta_k^{p_2,com} \left(V^{[k]} + (V^{[k]})^T \right) \begin{bmatrix} a \\ p_1 \end{bmatrix} \tag{26}$$

Sentiment Treebank

Annotation

Figure 13: Labeling interface.

- based on *rottentomatoes* corpus (Pang and Lee 2004b) (problems explained in Socher et al. 2013b
- parsed with binarized **PCFG Stanford Parser** (Klein and Manning 2003)
- $11,855$ sentences, $215,154$ phrases
- Annotated by 3 human judges
- 25 different sentiment values

Metrics

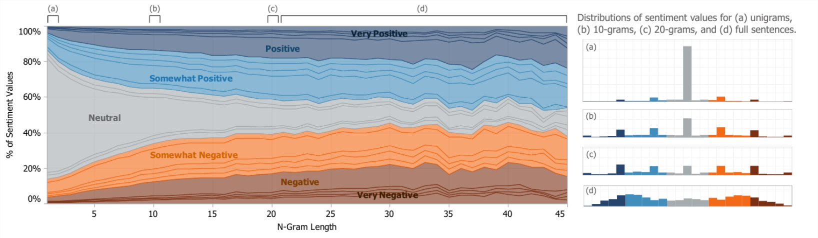


Figure 14: Sentiment annotations at each *n*-gram length.

- Few annotators used extreme values
- Even 5 classes are sufficient

Experiments

Fine-grained Sentiment

Model	Fine-grained		Positive/Negative	
	All	Root	All	Root
NB	67.2	41.0	82.6	81.8
SVM	64.3	40.7	84.6	79.4
BiNB	71.0	41.9	82.7	83.1
VecAvg	73.3	32.7	85.1	80.1
RNN	79.0	43.2	86.1	82.4
MV-RNN	78.7	44.4	86.8	82.9
RNTN	**80.7**	**45.7**	**87.6**	**85.4**

Figure 15: Accuracy for fine grained (5-class) and binary predictions at the sentence level (root) for all nodes, Socher et al. 2013a.

Fine-grained Sentiment

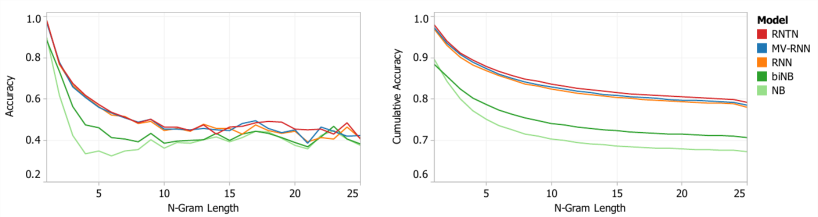


Figure 16: Accuracy curves for fine grained sentiment, left for each *n*-gram, right accumulated by $\leq$ *n*-grams, Socher et al. 2013a.

Full Sentence Binary Sentiment

- sentiment treebank improves baseline methods as well
- with coarse sentence level lables complex phenomena cannot be labeled
- sentiment treebank together with RNTN pushes accuracy on binary classification for short phrases from $< 80\%$ up to 85.4%

Contrastive Conjunction

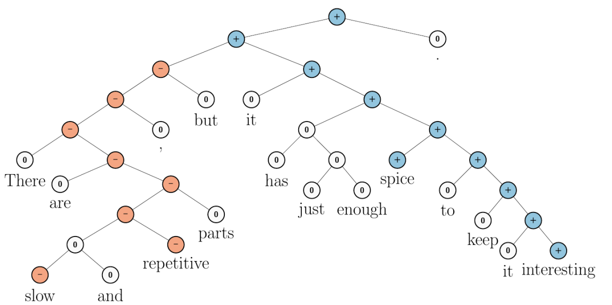


Figure 17: *X but Y* contrastive conjunction, Socher et al. 2013a.

- *binary* sentiment classification, 131 sentences in dataset
- **conditions:**
  1. subexpressions *X* and *Y* must be inverse
  2. correctly classified
  3. and node spanning *Y* and dominating *but*
- accuracy: RNTN 41%, MV-RNN 37%, RNN 36%, BiNB 27%

High Level Negation

- Two types of negation:
- **Negating Positive Sentences**
- **Negating Negative Sentences**
- Dataset of 21 positive and 21 negative sentences

Set 1: Negating Positive Sentences

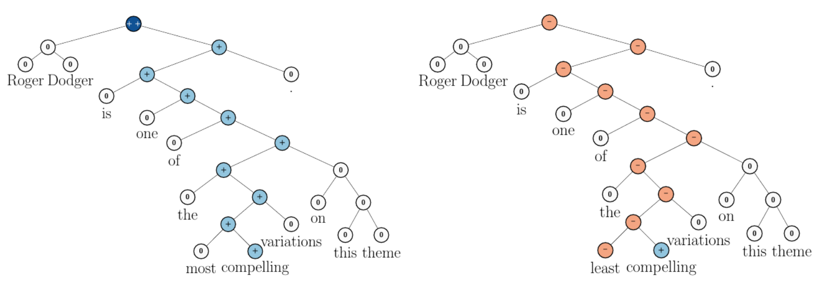


Figure 18: Negating positive sentence, *least* causes a subtle negation, Socher et al. 2013a.

Set 1: Negating Positive Sentences

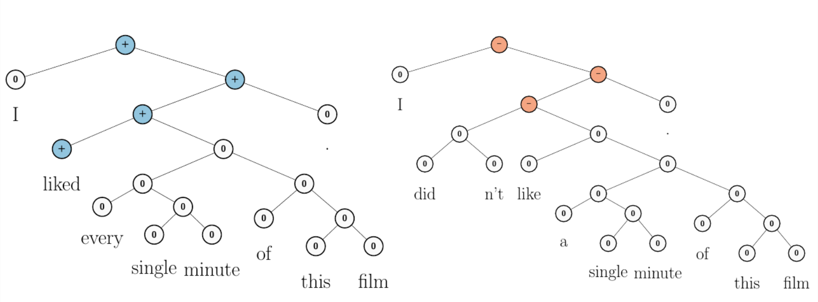


Figure 19: Negating positive sentence by *not.*, Socher et al. 2013a.

Set 2: Negating Negative Sentences

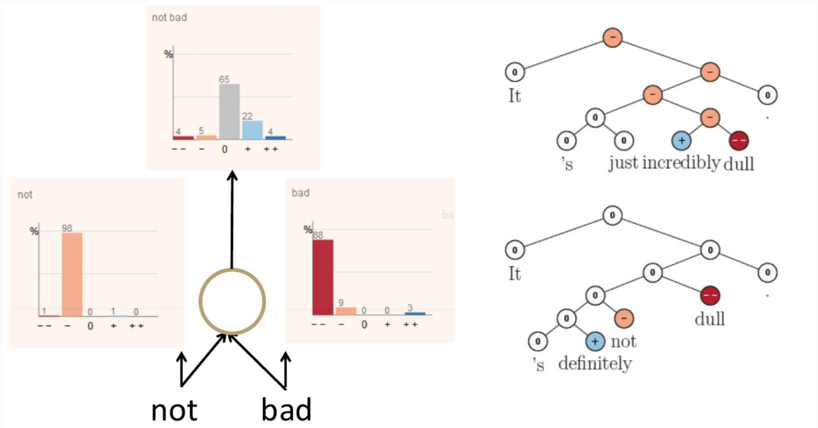


Figure 20: Negating *dull* turns sentence neutral, Socher et al. 2013a, `http://cs224d.stanford.edu/lectures/CS224d-Lecture11.pdf`.

Negation accuracy

Model	Accuracy	
	Negated Positive	Negated Negative
biNB	19.0	27.3
RNN	33.3	45.5
MV-RNN	52.4	54.6
RNTN	**71.4**	**81.8**

Figure 21: RNTN achieves best performance on negation, Socher et al. 2013a.

Change in activation

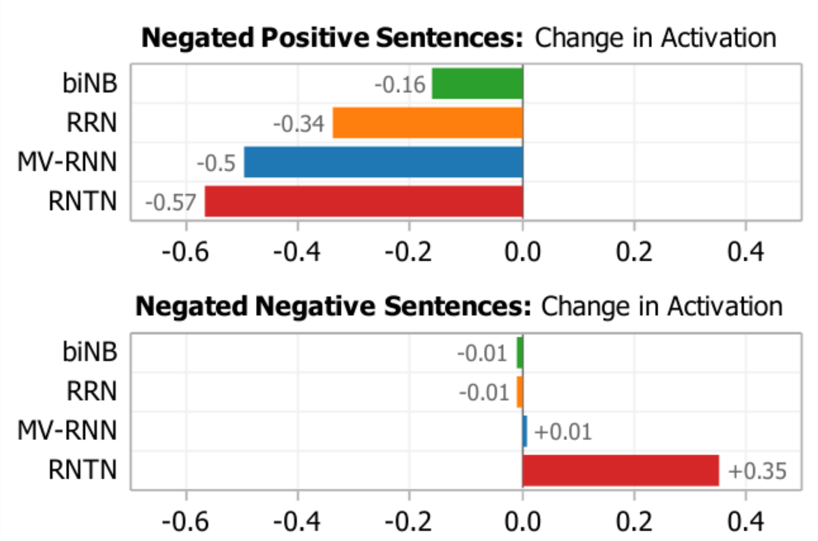


Figure 22: Only RNTN caputres negation of both types, Socher et al. 2013a.

Most Positive and Negative Phrases

n	Most positive n-grams	Most negative n-grams
1	engaging; best; powerful; love; beautiful	bad; dull; boring; fails; worst; stupid; painfully
2	excellent performances; A masterpiece; masterful film; wonderful movie; marvelous performances	worst movie; very bad; shapeless mess; worst thing; instantly forgettable; complete failure
3	an amazing performance; wonderful all-ages triumph; a wonderful movie; most visually stunning	for worst movie; A lousy movie; a complete failure; most painfully marginal; very bad sign
5	nicely acted and beautifully shot; gorgeous imagery, effective performances; the best of the year; a terrific American sports movie; refreshingly honest and ultimately touching	silliest and most incoherent movie; completely crass and forgettable movie; just another bad movie. A cumbersome and cliche-ridden movie; a humorless, disjointed mess
8	one of the best films of the year; A love for films shines through each frame; created a masterful piece of artistry right here; A masterful film from a master filmmaker,	A trashy, exploitative, thoroughly unpleasant experience ; this sloppy drama is an empty vessel.; quickly drags on becoming boring and predictable.; be the worst special-effects creation of the year

Figure 23: Examples of most positive and most negative phrases classified by RNTN, Socher et al. 2013a.

Outlook

DCNN

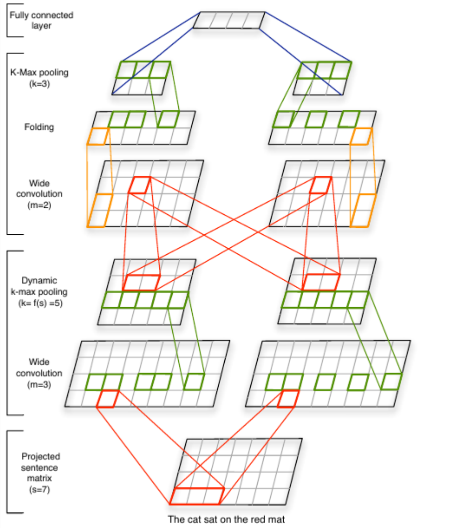


Figure 24: State-of-the-art convolutional neural network. Kalchbrenner, Grefenstette, and Blunsom 2014

Recent results

Method	Fine-grained	Binary
RAE (Socher et al., 2013)	43.2	82.4
MV-RNN (Socher et al., 2013)	44.4	82.9
RNTN (Socher et al., 2013)	45.7	85.4
DCNN (Blunsom et al., 2014)	48.5	86.8
Paragraph-Vec (Le and Mikolov, 2014)	48.7	87.8
CNN-non-static (Kim, 2014)	48.0	87.2
CNN-multichannel (Kim, 2014)	47.4	**88.1**
DRNN (Irsoy and Cardie, 2014)	49.8	86.6
LSTM	45.8	86.7
Bidirectional LSTM	49.1	86.8
2-layer LSTM	47.5	85.5
2-layer Bidirectional LSTM	46.2	84.8
Constituency Tree LSTM (no tuning)	46.7	86.6
Constituency Tree LSTM	**50.6**	86.9

Figure 25: Recent results on Stanford Sentiment Treebank.
`http://cs224d.stanford.edu/lectures/CS224d-Lecture11.pdf`

Conclusion

Conclusion for RNTN paper

Positive

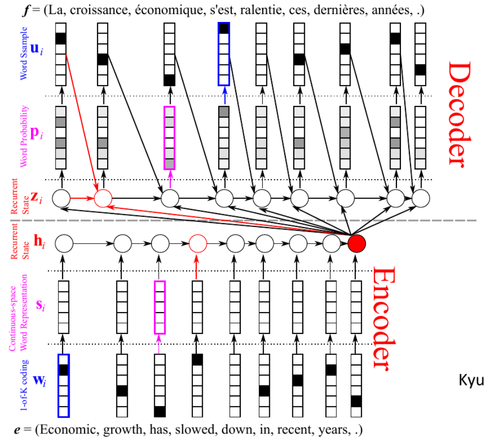
- RNTN is very *flexible*
- *no* feature engineering
- linguistically finer/better built corpus
- reference implementation available in `http://stanfordnlp.github.io/CoreNLP/`

Negative

- interaction of tensor slices not analyzed (compared to MV-RNN)
- unique **parse tree needed** & prevents *batch-learning* (parallelization)
- tensor shape somewhat *arbitrarily fixed* to vector length
- variety of NN models, what aspects are important (recursive, recurrent, convolutional)?

Questions?

Backup slides



Kyunghyun Cho et al. 2014

8/12/15

Recursive Autoencoder

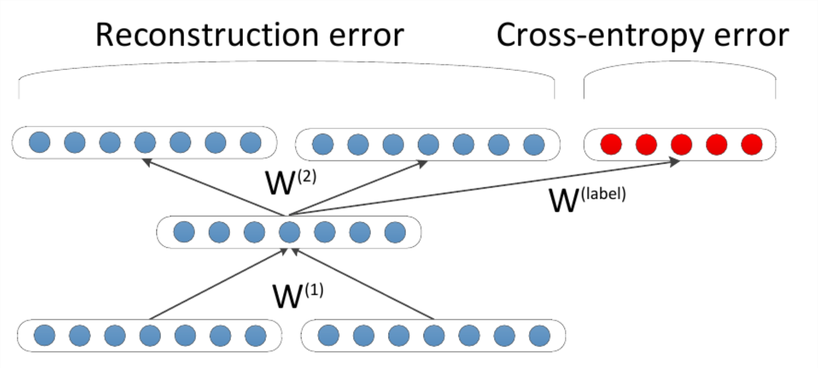


Figure 26: Sketch of an autoencoder (unsupervised). (Socher, Pennington, et al. 2011).

Recursive Autoencoder

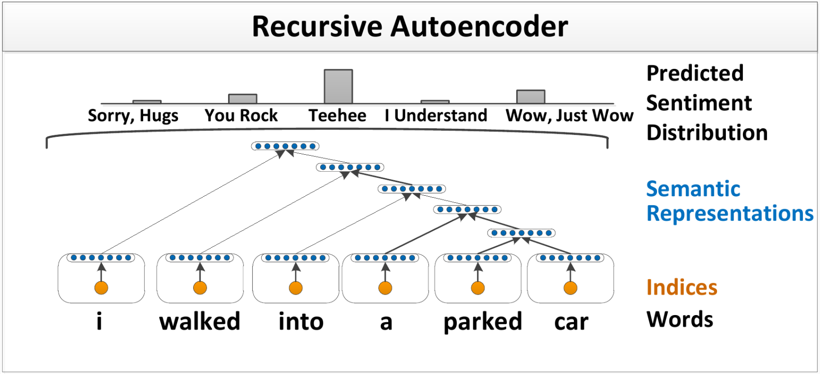


Figure 27: Recursive autoencoder architecture. Word indices (orange) are mapped into a semantic vector space (blue) and at each node a sentiment prediction is possible. (Socher, Pennington, et al. 2011).

Recursive Autoencoder

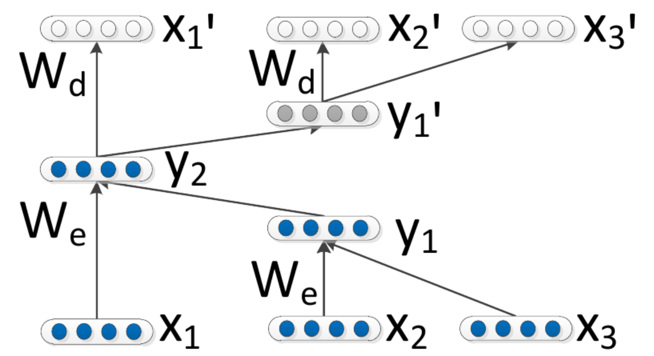


Figure 28: Unfolding Recursive Autoencoder to capture meaning of phrases (Richard Socher and others 2011).

- Used in paraphrase detection

MV-RNN

Method	Acc.
Tree-CRF (Nakagawa et al., 2010)	77.3
RAE (Socher et al., 2011c)	77.7
Linear MVR	77.1
MV-RNN	**79.0**

Figure 29: Performance on full length movie review polarity, Socher, Huval, et al. 2012.

Most Positive and Negative Phrases

n	**Most positive n-grams**	**Most negative n-grams**
1	engaging ; best ; powerful ; love ; beautiful ; entertaining ; clever ; terrific ; excellent ; great ;	bad ; dull ; boring ; fails ; worst ; stupid ; painfully ; cheap ; forgettable ; disaster ;
2	excellent performances ; amazing performance ; terrific performances ; A masterpiece ; masterful film ; wonderful film ; terrific performance ; masterful piece ; wonderful movie ; marvelous performances ;	worst movie ; bad movie ; very bad ; shapeless mess ; worst thing ; tepid waste ; instantly forgettable ; bad film ; extremely bad ; complete failure ;
3	an amazing performance ; a terrific performance ; a wonderful film ; wonderful all-ages triumph ; A masterful film ; a wonderful movie ; a tremendous performance ; drawn excellent performances ; most visually stunning ; A stunning piece ;	for worst movie ; A lousy movie ; most joyless movie ; a complete failure ; another bad movie ; fairly terrible movie ; a bad movie ; extremely unfunny film ; most painfully marginal ; very bad sign ;
5	nicely acted and beautifully shot ; gorgeous imagery , effective performances ; the best of the year ; a terrific American sports movie ; very solid , very watchable ; a fine documentary does best ; refreshingly honest and ultimately touching ;	silliest and most incoherent movie ; completely crass and forgettable movie ; just another bad movie . ; drowns out the lousy dialogue ; a fairly terrible movie ... ; A cumbersome and cliche-ridden movie ; a humorless , disjointed mess ;
8	one of the best films of the year ; simply the best family film of the year ; the best film of the year so far ; A love for films shines through each frame ; created a masterful piece of artistry right here ; A masterful film from a master filmmaker , ; 's easily his finest American film ... comes ;	A trashy , exploitative , thoroughly unpleasant experience ; this sloppy drama is an empty vessel . ; a meandering , inarticulate and ultimately disappointing film ; an unimaginative , nasty , glibly cynical piece ; bad , he 's really bad , and ; quickly drags on becoming boring and predictable . ; be the worst special-effects creation of the year ;

Figure 30: RNTN selects more strongly positive phrases, Socher et al. 2013a.

Metrics

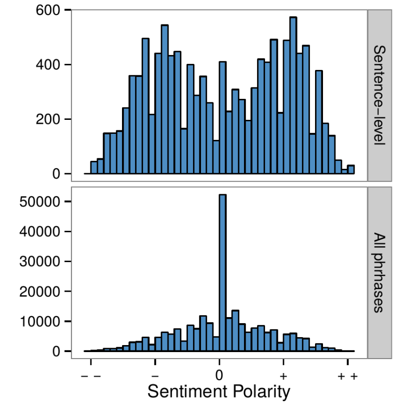


Figure 31: Top: Bimodal distribution over sentence sentiment. Bottom: Large percentage of phrases is neutral. Socher et al. 2013b

- Few annotators used extreme values
- Even 5 classes are sufficient

References

Ian Goodfellow Yoshua Bengio and Aaron Courville. "Deep Learning". Book in preparation for MIT Press. 2016. URL: http://www.deeplearningbook.org.

Edward Grefenstette and Mehrnoosh Sadrzadeh. "Experimental Support for a Categorical Compositional Distributional Model of Meaning". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 1394–1404. ISBN: 978-1-937284-11-4. URL: http://dl.acm.org/citation.cfm?id=2145432.2145580.

References II

Ozan Irsoy and Claire Cardie. “Opinion Mining with Deep Recurrent Neural Networks”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 720–728. URL: http://www.aclweb.org/anthology/D14-1080.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. “A Convolutional Neural Network for Modelling Sentences”. In: *CoRR* abs/1404.2188 (2014). URL: http://arxiv.org/abs/1404.2188.

References III

Dan Klein and Christopher D. Manning. “Accurate Unlexicalized Parsing”. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. ACL '03. Sapporo, Japan: Association for Computational Linguistics, 2003, pp. 423–430. DOI: 10.3115/1075096.1075150. URL: http://dx.doi.org/10.3115/1075096.1075150.

Omer Levy and Yoav Goldberg. “Neural Word Embedding as Implicit Matrix Factorization”. In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2177–2185. URL: http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization.pdf.

Tomas Mikolov et al. “Distributed Representations of Words and Phrases and their Compositionality”. In: *CoRR* abs/1310.4546 (2013). URL: http://arxiv.org/abs/1310.4546.

Bo Pang and Lillian Lee. “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts”. In: *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*. ACL ’04. Barcelona, Spain: Association for Computational Linguistics, 2004. DOI: 10.3115/1218955.1218990. URL: http://dx.doi.org/10.3115/1218955.1218990.

References V

Bo Pang and Lillian Lee. “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”. In: *In Proceedings of the ACL*. 2004, pp. 271–278.

Richard Socher and others. “Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection”. In: *Advances in Neural Information Processing Systems 24*. 2011.

References VI

Richard Socher, Jeffrey Pennington, et al. “Semi-supervised Recursive Autoencoders for Predicting Sentiment Distributions”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP ’11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 151–161. ISBN: 978-1-937284-11-4. URL: http://dl.acm.org/citation.cfm?id=2145432.2145450.

References VII

Richard Socher, Brody Huval, et al. "Semantic Compositionality Through Recursive Matrix-vector Spaces". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL '12. Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 1201–1211. URL: `http://dl.acm.org/citation.cfm?id=2390948.2391084`.

Richard Socher et al. "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, WA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642.

Richard Socher et al. “Supplementary Material: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, WA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642.

Ainur Yessenalina and Claire Cardie. “Compositional Matrix-space Models for Sentiment Analysis”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP ’11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 172–182. ISBN: 978-1-937284-11-4. URL: `http://dl.acm.org/citation.cfm?id=2145432.2145452`.