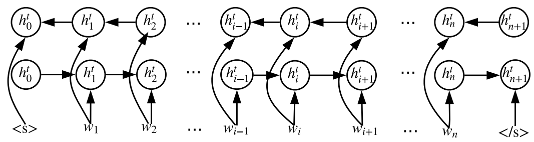
Sentence-State LSTM for Text Representation

Yue Zhang[1], **Qi Liu**[1] and **Linfeng Song**[2]

[1]Singapore University of Technology and Design, [2]University of Rochester

Introduction

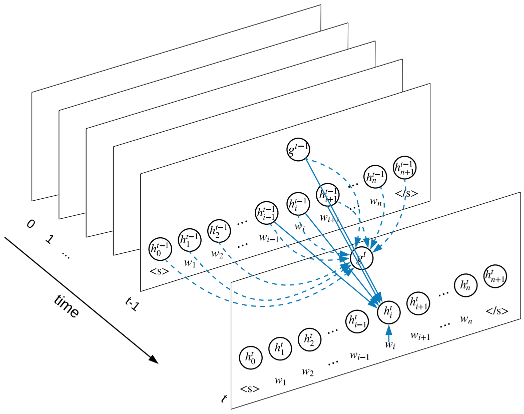
1. Bi-directional LSTM



2. Disadvantages:

1. BiLSTM is slow, due to its non-parallelism caused by its sequential nature (Vaswani et al., 2017).
2. Lack of balance between local n-gram and global sequence information (Wang et al., 2016).
3. Less effective in capturing long range dependencies (Koehn and Knowles, 2017).

Method



Word level nodes:

$$\boldsymbol{\xi}_i^t = [\boldsymbol{h}_{i-1}^{t-1}, \boldsymbol{h}_i^{t-1}, \boldsymbol{h}_{i+1}^{t-1}]$$

$$\hat{\boldsymbol{i}}_i^t = \sigma(\boldsymbol{W}_i\boldsymbol{\xi}_i^t + \boldsymbol{U}_i\boldsymbol{x}_i + \boldsymbol{V}_i\boldsymbol{g}^{t-1} + \boldsymbol{b}_i)$$

$$\hat{\boldsymbol{l}}_i^t = \sigma(\boldsymbol{W}_l\boldsymbol{\xi}_i^t + \boldsymbol{U}_l\boldsymbol{x}_i + \boldsymbol{V}_l\boldsymbol{g}^{t-1} + \boldsymbol{b}_l)$$

$$\hat{\boldsymbol{r}}_i^t = \sigma(\boldsymbol{W}_r\boldsymbol{\xi}_i^t + \boldsymbol{U}_r\boldsymbol{x}_i + \boldsymbol{V}_r\boldsymbol{g}^{t-1} + \boldsymbol{b}_r)$$

$$\hat{\boldsymbol{f}}_i^t = \sigma(\boldsymbol{W}_f\boldsymbol{\xi}_i^t + \boldsymbol{U}_f\boldsymbol{x}_i + \boldsymbol{V}_f\boldsymbol{g}^{t-1} + \boldsymbol{b}_f)$$

$$\hat{\boldsymbol{s}}_i^t = \sigma(\boldsymbol{W}_s\boldsymbol{\xi}_i^t + \boldsymbol{U}_s\boldsymbol{x}_i + \boldsymbol{V}_s\boldsymbol{g}^{t-1} + \boldsymbol{b}_s)$$

$$\boldsymbol{o}_i^t = \sigma(\boldsymbol{W}_o\boldsymbol{\xi}_i^t + \boldsymbol{U}_o\boldsymbol{x}_i + \boldsymbol{V}_o\boldsymbol{g}^{t-1} + \boldsymbol{b}_o)$$

$$\boldsymbol{u}_i^t = tanh(\boldsymbol{W}_u\boldsymbol{\xi}_i^t + \boldsymbol{U}_u\boldsymbol{x}_i + \boldsymbol{V}_u\boldsymbol{g}^{t-1} + \boldsymbol{b}_u)$$

$$\boldsymbol{i}_i^t, \boldsymbol{l}_i^t, \boldsymbol{r}_i^t, \boldsymbol{f}_i^t, \boldsymbol{s}_i^t = softmax(\hat{\boldsymbol{i}}_i^t, \hat{\boldsymbol{l}}_i^t, \hat{\boldsymbol{r}}_i^t, \hat{\boldsymbol{f}}_i^t, \hat{\boldsymbol{s}}_i^t)$$

$$\boldsymbol{c}_i^t = \boldsymbol{l}_i^t \odot \boldsymbol{c}_{i-1}^{t-1} + \boldsymbol{f}_i^t \odot \boldsymbol{c}_i^{t-1} + \boldsymbol{r}_i^t \odot \boldsymbol{c}_{i+1}^{t-1} + \boldsymbol{s}_i^t \odot \boldsymbol{c}_g^{t-1} + \boldsymbol{i}_i^t \odot \boldsymbol{u}_i^t$$

$$\boldsymbol{h}_i^t = \boldsymbol{o}_i^t \odot tanh(\boldsymbol{c}_i^t)$$

Sentence level node:

$$\bar{\boldsymbol{h}} = avg(\boldsymbol{h}_0^{t-1}, \boldsymbol{h}_1^{t-1}, \dots, \boldsymbol{h}_{n+1}^{t-1})$$

$$\hat{\boldsymbol{f}}_g^t = \sigma(\boldsymbol{W}_g\boldsymbol{g}^{t-1} + \boldsymbol{U}_g\bar{\boldsymbol{h}} + \boldsymbol{b}_g)$$

$$\hat{\boldsymbol{f}}_i^t = \sigma(\boldsymbol{W}_f\boldsymbol{g}^{t-1} + \boldsymbol{U}_f\boldsymbol{h}_i^{t-1} + \boldsymbol{b}_f)$$

$$\boldsymbol{o}^t = \sigma(\boldsymbol{W}_o\boldsymbol{g}^{t-1} + \boldsymbol{U}_o\bar{\boldsymbol{h}} + \boldsymbol{b}_o)$$

$$\boldsymbol{f}_0^t, \dots, \boldsymbol{f}_{n+1}^t, \boldsymbol{f}_g^t = softmax(\hat{\boldsymbol{f}}_0^t, \dots, \hat{\boldsymbol{f}}_{n+1}^t, \hat{\boldsymbol{f}}_g^t)$$

$$\boldsymbol{c}_g^t = \boldsymbol{f}_g^t \odot \boldsymbol{c}_g^{t-1} + \sum_i \boldsymbol{f}_i^t \odot \boldsymbol{c}_i^{t-1}$$

$$\boldsymbol{g}^t = \boldsymbol{o}^t \odot tanh(\boldsymbol{c}_g^t)$$

Tasks

1. Classification (vanilla attention):

$$\boldsymbol{y} = softmax(\boldsymbol{W}_c\boldsymbol{g} + \boldsymbol{b}_c)$$

$$\boldsymbol{g} = \sum_t \alpha_t \boldsymbol{h}_t$$

2. Sequence Labeling (vanilla CRF):

$$\boldsymbol{y}_i = softmax(\boldsymbol{W}_s\boldsymbol{h}_i + \boldsymbol{b}_s)$$

$$P(\boldsymbol{Y}_1^n|\boldsymbol{h}, \boldsymbol{W}_s, \boldsymbol{b}_s) = \frac{\prod_{i=1}^n \psi_i(y_{i-1}, y_i, \boldsymbol{h})}{\sum_{\boldsymbol{Y}_1^{n'}} \prod_{i=1}^n \psi_i(y'_{i-1}, y'_i, \boldsymbol{h})}$$

$$\psi_i(y_{i-1}, y_i, \boldsymbol{h}) = exp(\boldsymbol{W}_s^{y_{i-1}, y_i} h_i + b_s^{y_{i-1}, y_i})$$

Contrast with existing work

Model	Simultaneous	N-gram	Global	Recurrent
Bi-LSTM	×	×	sequential	√
CNN	√	√	pooling	×
SAN	√	×	attention	×
S-LSTM	√	√	gates	√

Experiments

1. Data

1) Classification:

Movie review (Pang and Lee (2008)), 16 datasets (Liu et al. (2017))

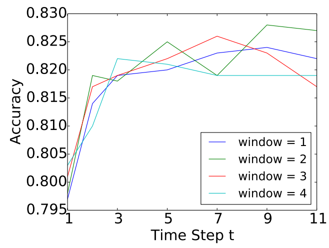
2) Sequence Labeling

NER: CoNLL (Sang et al., 2003)

POS tagging: PTB (Marcus et al., 1993)

2. Development

Model	Time (s)	Acc	# Param
LSTM	67	80.72	5,977K
BiLSTM	106	81.73	7,059K
2 stacked BiLSTM	207	81.97	9,221K
3 stacked BiLSTM	310	81.53	11,383K
4 stacked BiLSTM	411	81.37	13,546K
S-LSTM	65	82.64*	8,768K
CNN	34	80.35	5,637K
2 stacked CNN	40	80.97	5,717K
3 stacked CNN	47	81.46	5,808K
4 stacked CNN	51	81.39	5,855K
Transformer (N=6)	138	81.03	7,234K
Transformer (N=8)	174	81.86	7,615K
Transformer (N=10)	214	81.63	8,004K
BiLSTM+Attention	126	82.37	7,419K
S-LSTM+Attention	87	83.07*	8,858K



Iteration number

3. Classification

Model	Accuracy	Train (s)	Test (s)
Socher et al. (2011)	77.70	–	–
Socher et al. (2012)	79.00	–	–
Kim (2014)	81.50	–	–
Qian et al. (2016)	81.50	–	–
BiLSTM	81.61	51	1.62
2 stacked BiLSTM	81.94	98	3.18
3 stacked BiLSTM	81.71	137	4.67
3 stacked CNN	81.59	31	1.04
Transformer (N=8)	81.97	89	2.75
S-LSTM	**82.45***	41	1.53

Movie review

Dataset	SLSTM	Time (s)	BiLSTM	Time (s)	2 BiLSTM	Time (s)
Camera	**90.02***	50 (2.85)	87.05	115 (8.37)	88.07	221 (16.1)
Video	**86.75***	55 (3.95)	84.73	140 (12.59)	85.23	268 (25.86)
Health	**86.5**	37 (2.17)	85.52	118 (6.38)	85.89	227 (11.16)
Music	**82.04***	52 (3.44)	78.74	185 (12.27)	80.45	268 (23.46)
Kitchen	**84.54***	40 (2.50)	82.22	118 (10.18)	83.77	225 (19.77)
DVD	**85.52***	63 (5.29)	83.71	166 (15.42)	84.77	217 (28.31)
Toys	85.25	39 (2.42)	85.72	119 (7.58)	**85.82**	231 (14.83)
Baby	**86.25***	40 (2.63)	84.51	125 (8.50)	85.45	238 (17.73)
Books	**83.44***	64 (3.64)	82.12	240 (13.59)	82.77	458 (28.82)
IMDB	**87.15***	67 (3.69)	86.02	248 (13.33)	86.55	486 (26.22)
MR	**76.2**	27 (1.25)	75.73	39 (2.27)	75.98	72 (4.63)
Appeal	85.75	35 (2.83)	86.05	119 (11.98)	**86.35***	229 (22.76)
Magazines	**93.75***	51 (2.93)	92.52	214 (11.06)	92.89	417 (22.77)
Electronics	**83.25***	47 (2.55)	82.51	195 (10.14)	82.33	356 (19.77)
Sports	**85.75***	44 (2.64)	84.04	172 (8.64)	84.78	328 (16.34)
Software	**87.75***	54 (2.98)	86.73	245 (12.38)	86.97	459 (24.68)
Average	**85.38***	47.30 (2.98)	84.01	153.48 (10.29)	84.64	282.24 (20.2)

16 sets for classification

4. Sequential labeling

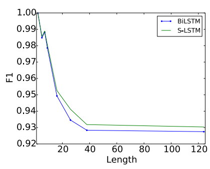
Model	F1	Train (s)	Test (s)
Collobert et al. (2011)	89.59	–	–
Passos et al. (2014)	90.90	–	–
Luo et al. (2015)	91.20	–	–
Huang et al. (2015)	90.10	–	–
Lample et al. (2016)	90.94	–	–
Ma and Hovy (2016)	91.21	–	–
Yang et al. (2017)	91.26	–	–
Rei (2017)	86.26	–	–
Peters et al. (2017)	**91.93**	–	–
BiLSTM	90.96	82	9.89
2 stacked BiLSTM	91.02	159	18.88
3 stacked BiLSTM	91.06	235	30.97
S-LSTM	**91.57***	79	9.78

Named entity recognition

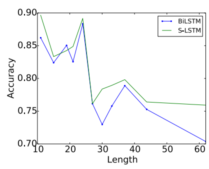
Model	Accuracy	Train (s)	Test (s)
Manning (2011)	97.28	–	–
Collobert et al. (2011)	97.29	–	–
Sun (2014)	97.36	–	–
søgaard (2011)	97.50	–	–
Huang et al. (2015)	**97.55**	–	–
Ma and Hovy (2016)	**97.55**	–	–
Yang et al. (2017)	**97.55**	–	–
BiLSTM	97.35	254	22.50
2 stacked BiLSTM	97.41	501	43.99
3 stacked BiLSTM	97.40	746	64.96
S-LSTM	**97.55**	237	22.16

POS tagging

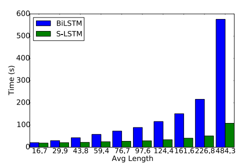
5. Contrast with Bi-LSTM



Classification



Sequence labeling



Classification

References

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NIPS. pages 6000-6010.

2. Xingyou Wang, Weijie Jiang, and Zhiyong Luo. 2016. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In Proceedings of COLING 2016. pages 2428-2437.

3. Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation. Vancouver, pages 28-39.

4. Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis.Foundations and Trends in Information Retrieval 2(1-2):1-135.

5. Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In Proceedings of ACL 2017. Vancouver, Canada, pages 1-10.

6. Tjong Kim Sang, Erik F, and De Meulder Fien. 2003. Introduction to the conll2003 shared task: Languageindependent named entity recognition. In Proceedings of HLTNAACL 2003-Volume 4. pages 142147.

7. Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. Computational linguistics 19(2):313330.