Unsupervised learning methods based on huge data for the needs of supervised learning methods using small training sets

Karol Kaczmarek

Adam Mickiewicz University
Poznań

Applica.ai
Warsaw

2018

Outline

1 Difficulties

2 Neural language model

3 Data sets

4 References

Difficulties

- supervised learning methods using small training sets
  - classification
  - extraction
- increasing the amount of small training sets
- noisy data (types, ORC error, rare words)
- unbalanced training sets
- time and costs

Outline

1 Difficulties

2 Neural language model

3 Data sets

4 References

Neural language model

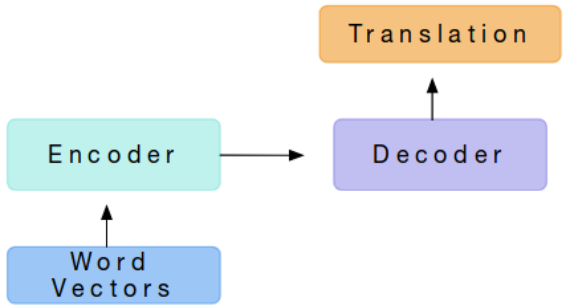
- use different architectures:
  - word level
  - char level
  - sentence level
- evaluation models:
  - perplexity
  - guest gap
- OCR types tolerant

Neural language model

- use embedding of models:
  - CoVe [1]
  - ELMo [2]
- compare with:
  - word2vec [3]
  - fastText [4]

CoVe

- inspired by the pretrained ImageNet
- train an encoder and transfer the encoder to other tasks
- training an attentional sequence-to-sequence model for English-to-German translation



- $CoVe(w) = MT\text{-}LSTM(GloVe(w))$, w - sequence of words

CoVe

	Model	Test		Model	Test
SST-2	P-LSTM [Wieting et al., 2016]	89.2	TREC-6	SVM [da Silva et al., 2011]	95.0
	CT-LSTM [Looks et al., 2017]	89.4		SVM [Van-Tu and Anh-Cuong, 2016]	95.2
	TE-LSTM [Huang et al., 2017]	89.6		DSCNN-P [Zhang et al., 2016]	95.6
	NSE [Munkhdalai and Yu, 2016a]	89.7		*BCN+Char+CoVe (Ours)*	*95.8*
	BCN+Char+CoVe (Ours)	*90.3*		TBCNN [Mou et al., 2015]	96.0
	bmLSTM [Radford et al., 2017]	**91.8**		**LSTM-CNN [Zhou et al., 2016]**	**96.1**
SST-5	MVN [Guo et al., 2017]	51.5	TREC-50	SVM [Loni et al., 2011]	89.0
	DMN [Kumar et al., 2016]	52.1		SNoW [Li and Roth, 2006]	89.3
	LSTM-CNN [Zhou et al., 2016]	52.4		*BCN+Char+CoVe (Ours)*	*90.2*
	TE-LSTM [Huang et al., 2017]	52.6		RulesUHC [da Silva et al., 2011]	90.8
	NTI [Munkhdalai and Yu, 2016b]	53.1		SVM [Van-Tu and Anh-Cuong, 2016]	91.6
	BCN+Char+CoVe (Ours)	***53.7***		**Rules [Madabushi and Lee, 2016]**	**97.2**
IMDb	*BCN+Char+CoVe (Ours)*	*91.8*	SNLI	DecAtt+Intra [Parikh et al., 2016]	86.8
	SA-LSTM [Dai and Le, 2015]	92.8		NTI [Munkhdalai and Yu, 2016b]	87.3
	bmLSTM [Radford et al., 2017]	92.9		re-read LSTM [Sha et al., 2016]	87.5
	TRNN [Dieng et al., 2016]	93.8		btree-LSTM [Paria et al., 2016]	87.6
	oh-LSTM [Johnson and Zhang, 2016]	94.1		600D ESIM [Chen et al., 2016]	88.0
	Virtual [Miyato et al., 2017]	**94.1**		***BCN+Char+CoVe (Ours)***	***88.1***

ELMo (Embeddings from Language Models)

- function of all of the internal layers of the biLM
- lower-level LSTM states model aspects of syntax
- higher-level LSTM states capture context-dependent aspects of word meaning
- ELMo outperforms CoVe

ELMo (Embeddings from Language Models)

Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

Outline

1 Difficulties

2 Neural language model

3 Data sets

4 References

Data sets

- Get more data
  - Common Crawl
  - Wikipedia
- Extract in-domain data sets
- Data augmentation

Outline

1 Difficulties

2 Neural language model

3 Data sets

4 References

References I

[1] B. McCann, J. Bradbury, C. Xiong, and R. Socher, “Learned in translation: Contextualized word vectors,” 2017.

[2] V. Joshi, M. Peters, and M. Hopkins, “Extending a parser to distant domains using a few dozen partially annotated examples,” 2018.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013.

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” 2017.