A Linguistically Informed Convolutional Neural Network

Sebastian Ebert, Ngoc Thang Vu, Hinrich Schütze

Center for Information and Language Processing, University of Munich, Germany

ebert@cis.lmu.de

1. Introduction

- large amount of labeled training data necessary to train Convolutional Neural Network; linguistic knowledge can help to compensate for it
- linguistic knowledge is crucial for polarity classification
- linguistic resources already available, e.g., sentiment lexicons
- question: how to incorporate such knowledge into CNN

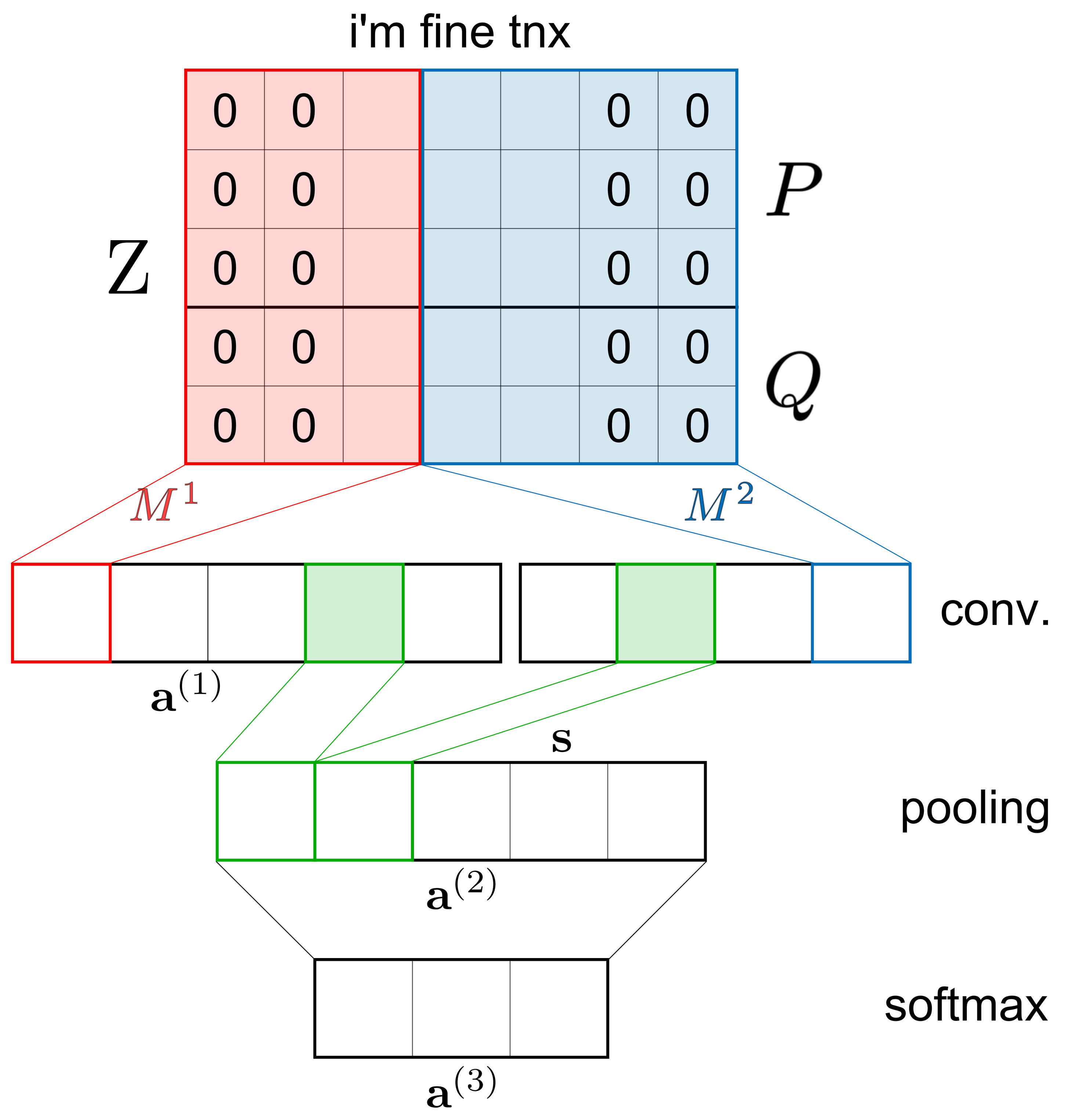
1.1 Contributions

1. incorporation of linguistic features into Convolutional Neural Network (CNN)
   - word-level features: learn interactions between words
   - sentence-level features: learn overall features
2. performance comparable to state-of-the-art on SemEval Twitter data

2. Convolutional Neural Network

2.1 Why CNN?

- work with arbitrary input length
- capture sequential phenomena, i.e., keep word order
- consider words in their contexts
- capture long-distance effects
- goal of CNN: conflate the input sequence into a meaningful representation by finding salient features that indicate polarity



2.2 Input

$$Z = \begin{bmatrix} | & & | \\ LT_{\cdot,t_1} & \cdots & LT_{\cdot,t_n} \\ | & & | \end{bmatrix}$$

- $LT \in \mathbb{R}^{d \times |V|}$: lookup table
- d: length of representation
- V: vocabulary

2.3 2D Convolution

$$\mathbf{a}_o^{(1)} = \sum_{i=1}^{d} \sum_{j=1}^{m} M_{i,j} Z_{i,o+j}$$

- $M \in \mathbb{R}^{d \times m}$: filter matrix
- m: filter size
- $\mathbf{a}_o^{(1)}$: layer's activation at current position $o \in [0, n-m]$ of convolution

2.4 Max Pooling and Non-linearity

- ReLU non-linearity: $\mathbf{a}^{(2)} = \max(0, \mathbf{a}^{(1)} + b^{(2)})$
- $\mathbf{a}^{(1)}$: maximum value of $\mathbf{a}_o^{(1)}$
- $b^{(2)}$: bias

2.5 Softmax

- concatenate sentence features: $\mathbf{a}^{(2)\prime} = [\mathbf{a}^{(2)}\ \mathbf{s}]$
- input to fully connected layer: $\mathbf{z} = W\mathbf{a}^{(2)\prime} + b^{(3)}$
- softmax: $\mathbf{a}_i^{(3)} = \frac{\exp(\mathbf{z}_i)}{\sum_j \exp(\mathbf{z}_j)}$

3. Linguistic Features

3.1 Word-level Features

$$LT = \begin{bmatrix} P \\ Q \end{bmatrix}$$

- $P \in \mathbb{R}^{d_P \times |V|}$: word embeddings; randomly initialized or pre-trained with word2vec on unlabeled Twitter data
- $Q \in \mathbb{R}^{d_Q \times |V|}$: linguistic features

binary sentiment indicators binary polarity label per token; lexicons: MPQA [Wilson et al., 2005], Opinion lexicon [Hu and Liu, 2004], NRCC Emotion lexicon [Mohammad and Turney, 2013]

sentiment scores sentiment score per token (or bigram); lexicons: sentiment 140 lexicon, hashtag lexicon [Mohammad et al., 2013]

binary negation indicator if token is between a negation word and the next punctuation

3.2 Sentence-level Features

counts number of terms that are all upper case; number of elongated words such as 'coooool'; number of emoticons; number of contiguous sequences of punctuation; number of negated words

sentiment scores number of sentiment words in a sentence; the sum of sentiment scores of these words as provided by the lexicons; the maximum sentiment score; the sentiment score of the last word

4. Experiments

4.1 Training Parameters

- trainable parameters: $\theta = \{P, M^*, W, b^{(*)}\}$
- training hyper-parameters: mini-batch stochastic gradient descent with 100 batch size, AdaGrad with initial $lr = 0.01$, ℓ_2 with $\lambda = 5e^{-5}$
- CNN hyper-parameters: $d_P = 60$, 100 filters for each $m \in \{2, 3, 4, 5\}$

4.2 Data

- SemEval 2015 data set [Rosenthal et al., 2015] and test set of Sentiment140 corpus (Sent140) [Go et al., 2009]
- results reported: $F_{1,macro} = \frac{1}{2}(F_{1,positive} + F_{1,negative})$
- preprocessing: tokenization, normalization of user mentions, urls, punctuation

4.3 Baselines

- SVM with bag-of-words and linguistic features [Mohammad et al., 2013]
- Webis [Hagen et al., 2015] (ensemble) and UNITN [Severyn and Moschitti, 2015] (CNN)

4.4 Results

model	features			SemEval	Sent140
SVM	bow			50.51	67.34
	ling.			57.28	66.90
	bow + ling.			59.28	70.21
Webis				64.84	-
UNITN				64.59	-
	emb.	word	sent.		
lingCNN		+		57.83	72.58
		+	+	59.24	74.36
	+			62.72	77.59
	+		+	62.61	79.14
	+	+		63.43	80.21
	+	+	+	**64.46**	**80.75**

5. Error Analysis

5.1 Examples

- "saturday night in with toast , hot choc & <user> on e news #happydays"
- only '#happydays' has sentiment; no embedding because unknown word; but in lexicon
- before misclassified as neutral, now classified as positive

- "shiiiiit my sats is on saturday . i'm going to fail"
- 'fail' is strongly negative, but occurs only 10 times in the training set, i.e., likely not enough to learn a good sentiment-bearing embedding
- before misclassified as neutral, now classified as negative

5.2 Corpus Size

	1000	3000	all
emb.	49.89	58.10	62.72
emb. + word + sent.	60.89	62.51	64.46

Acknowledgments This work was supported by DFG (grant SCHU 2246/10).