Joint Embedding of Words and Labels for Text Classification
ACL2018

修晔良

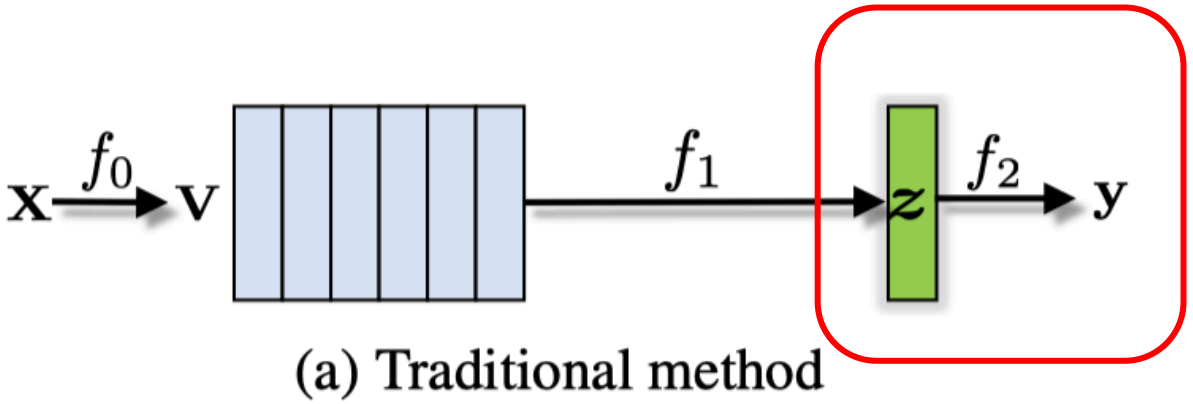
Motivation:

已有的词嵌入方法都是仅基于词特征来构建样本表示，而忽略了标签对构建词嵌入的贡献。

Proposed Method:

- The proposed LEAM is implemented by jointly embedding the word and label in the same latent space, and the text representations are constructed directly using the text-label compatibility

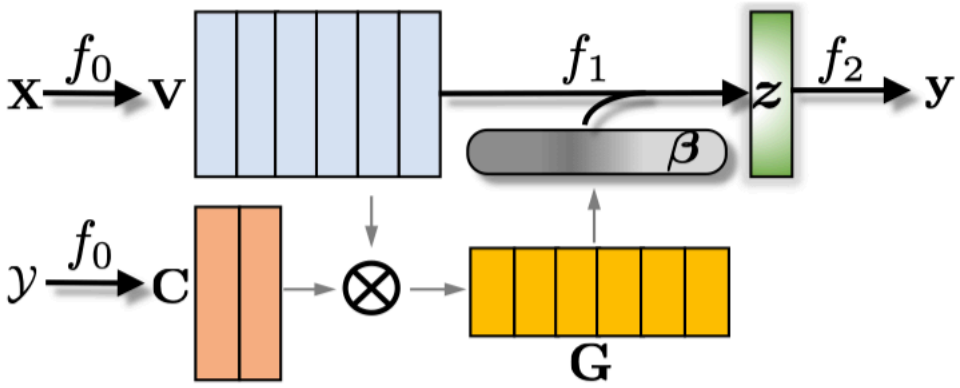
传统文本分类的Pipeline



(a) Traditional method

- $f_0 : \mathbf{X} \mapsto \mathbf{V}$, the text sequence is represented as its word-embedding form $\mathbf{V}$, which is a matrix of $P \times L$.
- $f_1 : \mathbf{V} \mapsto \boldsymbol{z}$, a compositional function f_1 aggregates word embeddings into a fixed-length vector representation $\boldsymbol{z}$.
- $f_2 : \boldsymbol{z} \mapsto \boldsymbol{y}$, a classifier f_2 annotates the text representation $\boldsymbol{z}$ with a label.

已提出算法框架图



(b) Proposed joint embedding method

利用cosine 相似度计算每个label-word之间的相似度

$$\mathbf{G} = (\mathbf{C}^\top \mathbf{V}) \oslash \hat{\mathbf{G}}$$

其中$\hat{\mathrm{G}} \in R^{K*L}$ $\hat{G}$中的每一个元素为 $\hat{\mathrm{g}}_{kl} = ||c_k||||v_l||$

为了更好的获取连续word之间的空间信息，作者在相似度中引入非线性

$\mathrm{C} = [c_1, c_2, \ldots, c_k] \in R^{K*L}$

K表示标签个数，L表示词向量维度

$$u_l = RELU(G_{l-r:l+r} W_1 + b_1)$$ 其中 $\mathrm{u}_l \in R^K$

$$m_l = max\_pooling(u_l)$$

$$z = \sum_l \beta_l v_l$$

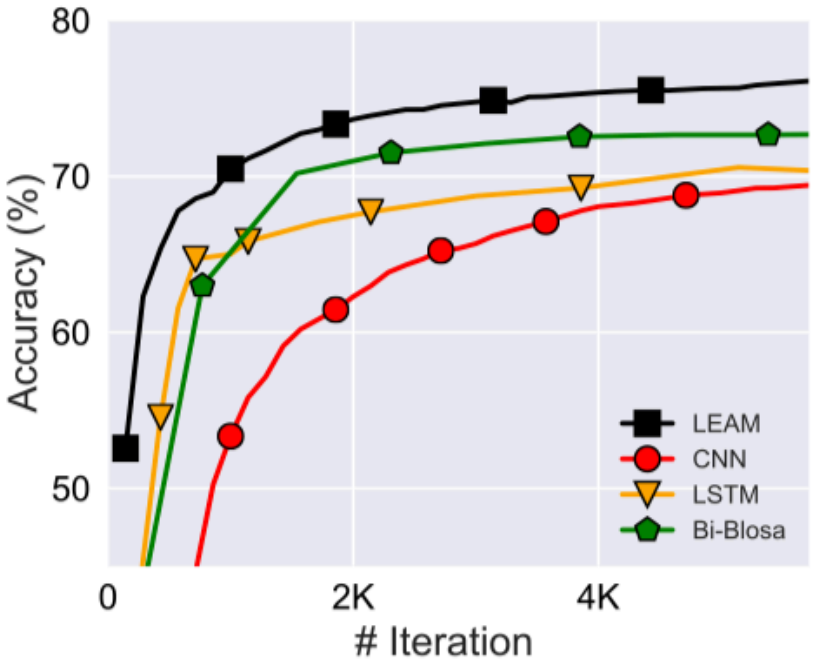
实验结果与分析

Model	# Parameters	Time cost (s)
CNN	541k	171
LSTM	1.8M	598
SWEM	61K	63
Bi-BloSAN	3.6M	292
LEAM	65K	65

Table 4: Comparison of model size and speed.

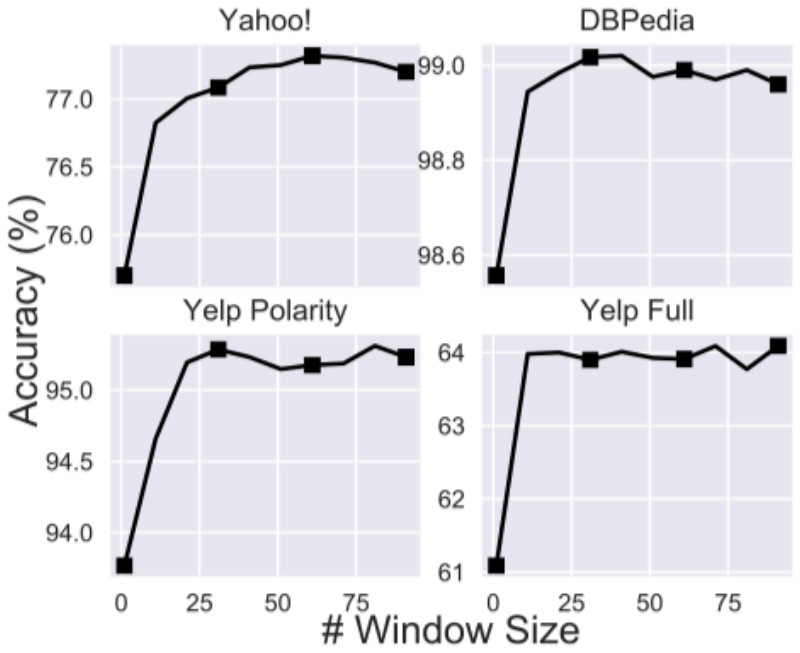
Model	Yahoo	DBPedia	AGNews	Yelp P.	Yelp F.
Bag-of-words (Zhang et al., 2015)	68.90	96.60	88.80	92.20	58.00
Small word CNN (Zhang et al., 2015)	69.98	98.15	89.13	94.46	58.59
Large word CNN (Zhang et al., 2015)	70.94	98.28	91.45	95.11	59.48
LSTM (Zhang et al., 2015)	70.84	98.55	86.06	94.74	58.17
SA-LSTM (word-level) (Dai and Le, 2015)	-	98.60	-	-	-
Deep CNN (29 layer) (Conneau et al., 2017)	73.43	98.71	91.27	**95.72**	**64.26**
SWEM (Shen et al., 2018a)	73.53	98.42	92.24	93.76	61.11
fastText (Joulin et al., 2016)	72.30	98.60	92.50	95.70	63.90
HAN (Yang et al., 2016)	75.80	-	-	-	-
Bi-BloSAN$^\diamond$ (Shen et al., 2018c)	76.28	98.77	**93.32**	94.56	62.13
LEAM	**77.42**	**99.02**	92.45	95.31	64.09
LEAM (linear)	75.22	98.32	91.75	93.43	61.03

Table 3: Test Accuracy on document classification tasks, in percentage. $^\diamond$We ran Bi-BloSAN using the authors' implementation; all other results are directly cited from the respective papers.



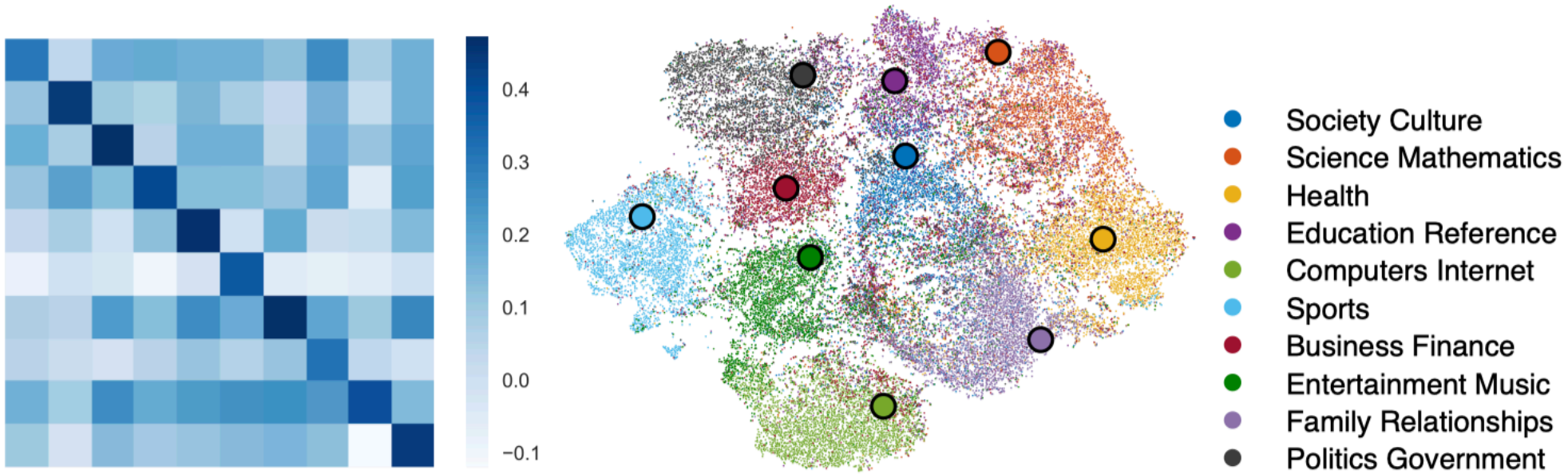
(a) Convergence speed

模型收敛速度分析



(c) Effects of window size

超参数敏感性分析



(a) Cosine similarity matrix

(b) t-SNE plot of joint embeddings

Figure 3: Correlation between the learned text sequence representation z and label embedding $\mathbf{V}$. (a) Cosine similarity matrix between averaged $\bar{z}$ per class and label embedding $\mathbf{V}$, and (b) t-SNE plot of joint embedding of text z and labels $\mathbf{V}$.

Interpretability of attention

what professional coaches have never played the sport , they are coaching ?
, , most played at some level . either college or pro or even high school . n
nbut one of the biggest names is the football coach at notra dame . he never
played college , pro , and i don t think highschool either .

who is the greatest rock drummer of all time ? , a show of hands . . . come
on rush fans ! ! ! , i would have to go with neal peart . . . he s easily the best
there is but i m hardly a fan of rush . . . . the drummer in my band just
thinks he s the greatest and i tend to agree ! ! he s got a hell of

(a) Yahoo dataset

The darker yellow means more important words. The 1st text sequence is on the topic of "Sports", and the 2nd text sequence is "Entertainment".

Contribution

1. 提出了一种利用标签信息来增强样本表示的端到端文本分类算法

2. 提出的算法参数少，具有一定的可解释性。

3. 实验显示已提出的算法具有很好的分类效果。

Thank U