Joint Embedding of Words and Labels for Text Classification

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Heano, Lawrence Carin

Code: bit.ly/LEAM-Duke

Motivation

- Text classification as a label-word joint embedding problem
- Use label information to construct text-sequence representations

Contribution

- **Label Embedding Attentive Model (LEAM)**
- **High accuracy** in standard benchmarks and clinical dataset
- Only basic algebraic operation involved, hence retains interpretability
- **Fewer parameters and less computation**

Traditional Models

- Given training set $\mathcal{S} = \{(\mathbf{X}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$ of pair-wise data, where
  - $\mathbf{X} \in \mathcal{X}$ is the text sequence,
  - $\boldsymbol{y} \in \mathcal{Y}$ is its corresponding label
- Goal: learn a function $f : \mathcal{X} \mapsto \mathcal{Y}$ by

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^{N} \delta(\boldsymbol{y}_n, f(\mathbf{X}_n))$$

- A typical text classification can be presented as a function decomposition

$$f = f_0 \circ f_1 \circ f_2$$

$f_0 : \mathbf{X} \mapsto \mathbf{V}$ represents text sequence as its word-embedding form

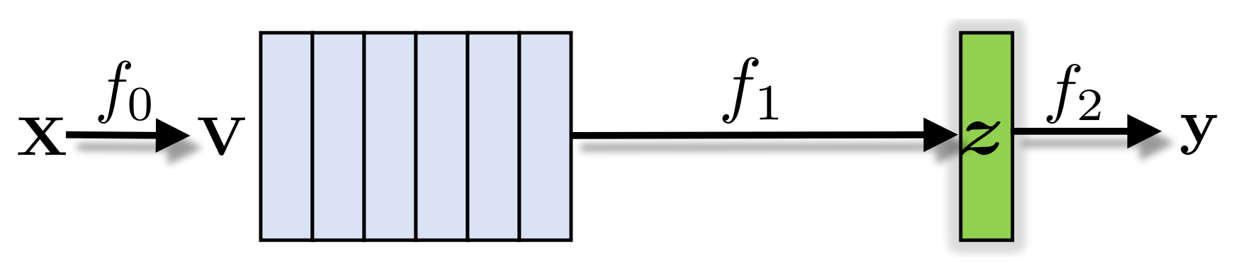
$f_1 : \mathbf{V} \mapsto \boldsymbol{z}$, aggregates word embeddings into a vector representation

$f_2 : \boldsymbol{z} \mapsto \boldsymbol{y}$, annotates the text representation with a label

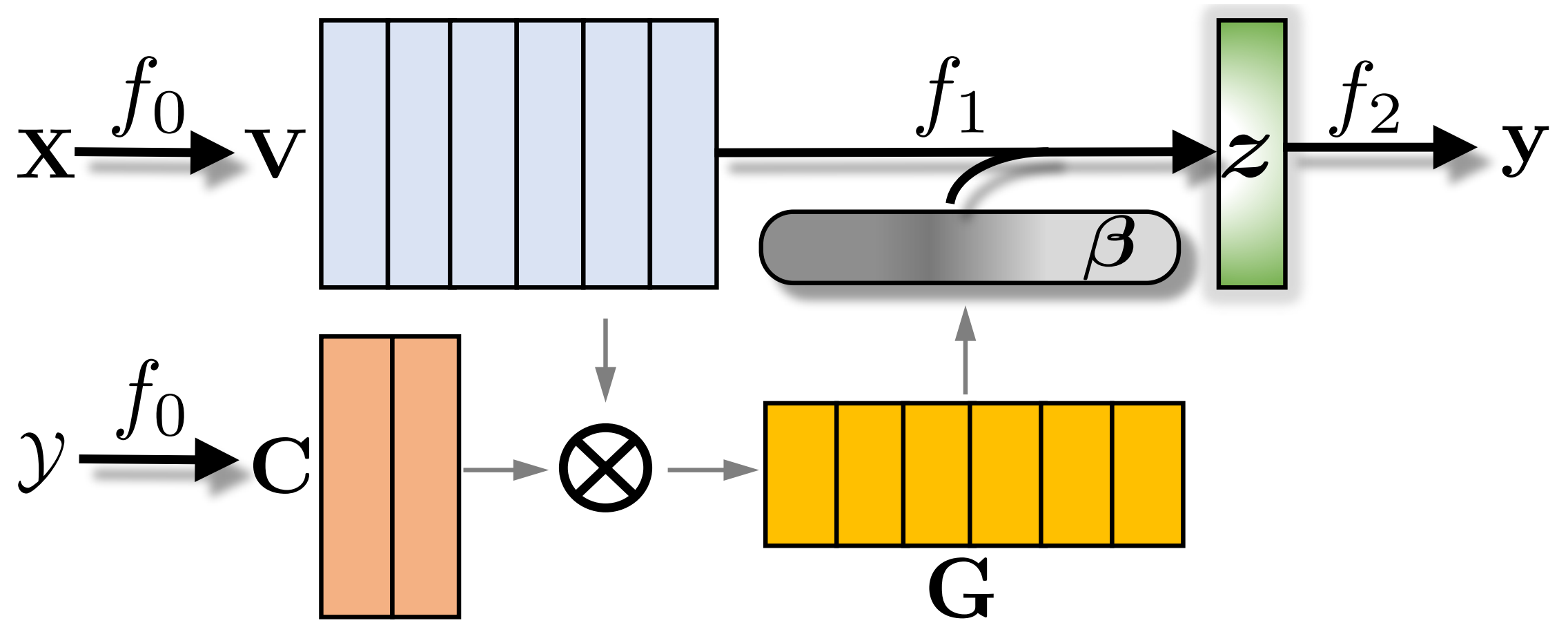
Complexity (Fewer parameters & Less computation)

Model	Parameters	Complexity	Seq.
CNN	$m \cdot h \cdot P$	$O(m \cdot h \cdot L \cdot P)$	$O(1)$
LSTM	$4 \cdot h \cdot (h + P)$	$O(L \cdot h^2 + h \cdot L \cdot P)$	$O(L)$
SWEM	0	$O(L \cdot P)$	$O(1)$
Bi-BloSAN	$7 \cdot P^2 + 5 \cdot P$	$O(P^2 \cdot L^2/R + P^2 \cdot L + P^2 \cdot R^2)$	$O(1)$
LEAM	$K \cdot P$	$O(K \cdot L \cdot P)$	$O(1)$

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



Traditional Models



$\mathbf{X} \xrightarrow{f_0} \mathbf{V}$
f_1
β
z
f_2
$\mathbf{y}$
$\mathcal{Y} \xrightarrow{f_0} \mathbf{C}$
$\otimes$
$\mathbf{G}$

LEAM

Proposed Model : LEAM

- Word embeddings $\mathbf{V}$ and the label embeddings $\mathbf{C}$ in a joint space

1. Measure the *compatibility* of label-word pairs via cosine similarity

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

2. *Local spatial information* among consecutive words (phrases)

$$\mathbf{u}_l = \mathrm{ReLU}(\mathbf{G}_{l-r:l+r}\mathbf{W} + \mathbf{b})$$

3. The *compatibility/attention score* $\quad \beta = \mathrm{Softmax}(\mathbf{m})$

   where the largest compatibility value is pooled: $m_l = \text{max-pooling}(\mathbf{u}_l)$

4. The *text sequence representation is weighted average* of word embeddings weighted by label-based attention score

$$\mathbf{z} = \sum_l \beta_l \mathbf{v}_l$$

- Single-label problem: training object is $\displaystyle \min_{f \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^{N} \mathrm{CE}(\boldsymbol{y}_n, f_2(\boldsymbol{z}_n))$
- Multi-label problem: training object is $\displaystyle \min_{f \in \mathcal{F}} \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} \mathrm{CE}(\boldsymbol{y}_{nk}, f_2(\boldsymbol{z}_{nk}))$
- **Regularization**: To force the label embedding as the anchor points for each classes, we regularize the learned label embeddings to be on its corresponding manifold

$$\min_{f \in \mathcal{F}} \frac{1}{K} \sum_{n=1}^{K} \mathrm{CE}(\boldsymbol{y}_k, f_2(\boldsymbol{c}_k)).$$

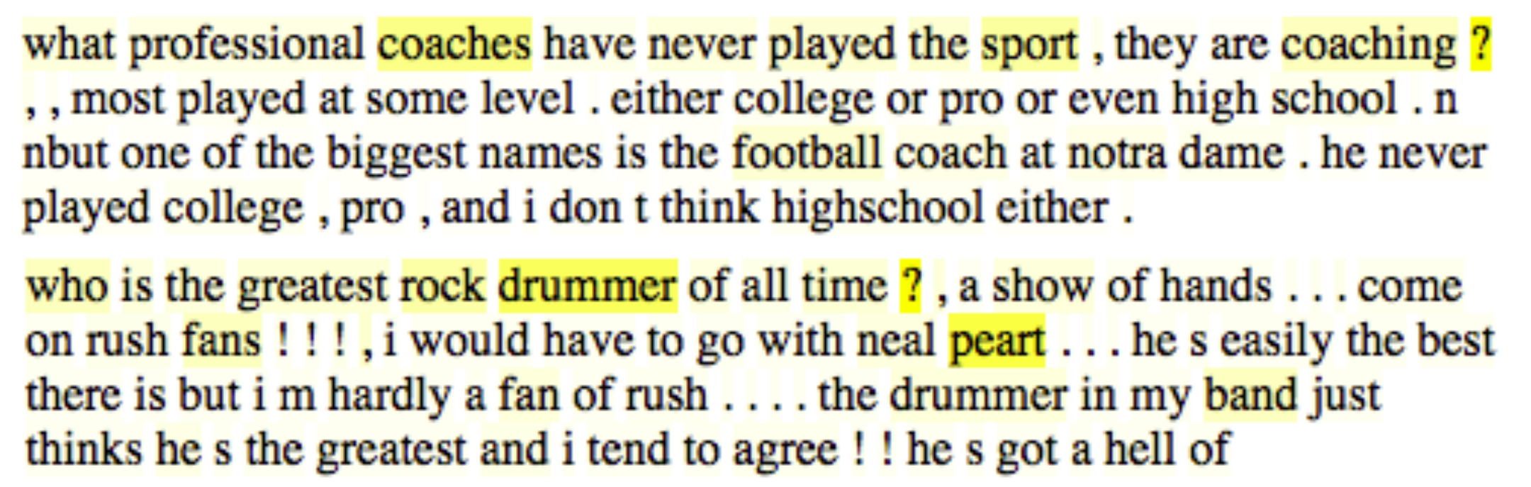
Experimental Results

Benchmark Classification Accuracy

Model	Yahoo	DBPedia	AGNews	Yelp P.	Yelp F.
Bag-of-	68.9	96.6	88.8	92.2	58
CNN	70.94	98.28	91.45	95.11	59.48
LSTM	70.84	98.55	86.06	94.74	58.17
Deep CNN	73.43	98.71	91.27	**95.72**	**64.26**
SWEM	73.53	98.42	92.24	93.76	61.11
fastText	72.3	98.6	92.5	95.7	63.9
HAN	75.8				
Bi-BloSAN	76.28	98.77	**93.32**	94.56	62.13
LEAM	**77.42**	**99.02**	92.45	95.31	64.09

Test Accuracy on document classification tasks, in percentage

Interpretability

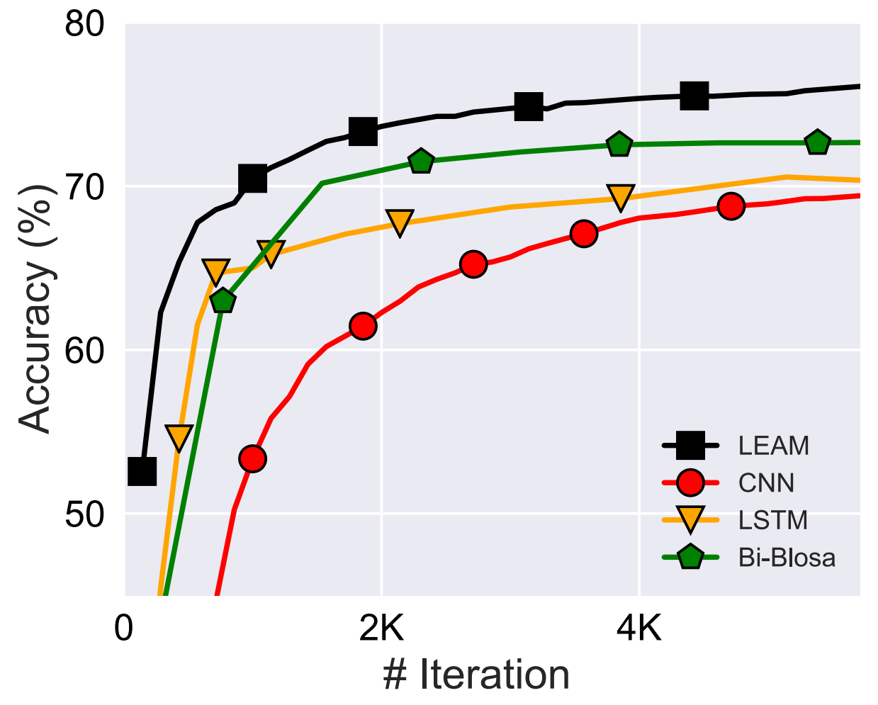


what professional coaches have never played the sport , they are coaching ? , , most played at some level . either college or pro or even high school . n nbut one of the biggest names is the football coach at notra dame . he never played college , pro , and i don t think highschool either .

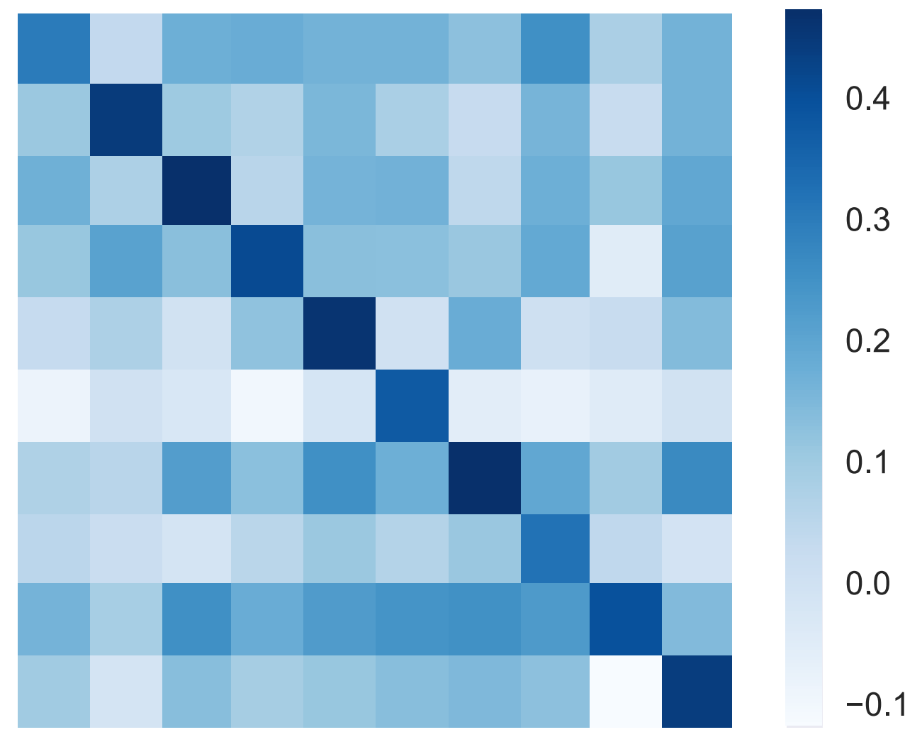
who is the greatest rock drummer of all time ? , a show of hands . . . come on rush fans ! ! ! , i would have to go with neal peart . . . he s easily the best there is but i m hardly a fan of rush . . . . the drummer in my band just thinks he s the greatest and i tend to agree ! ! he s got a hell of

Attention score example

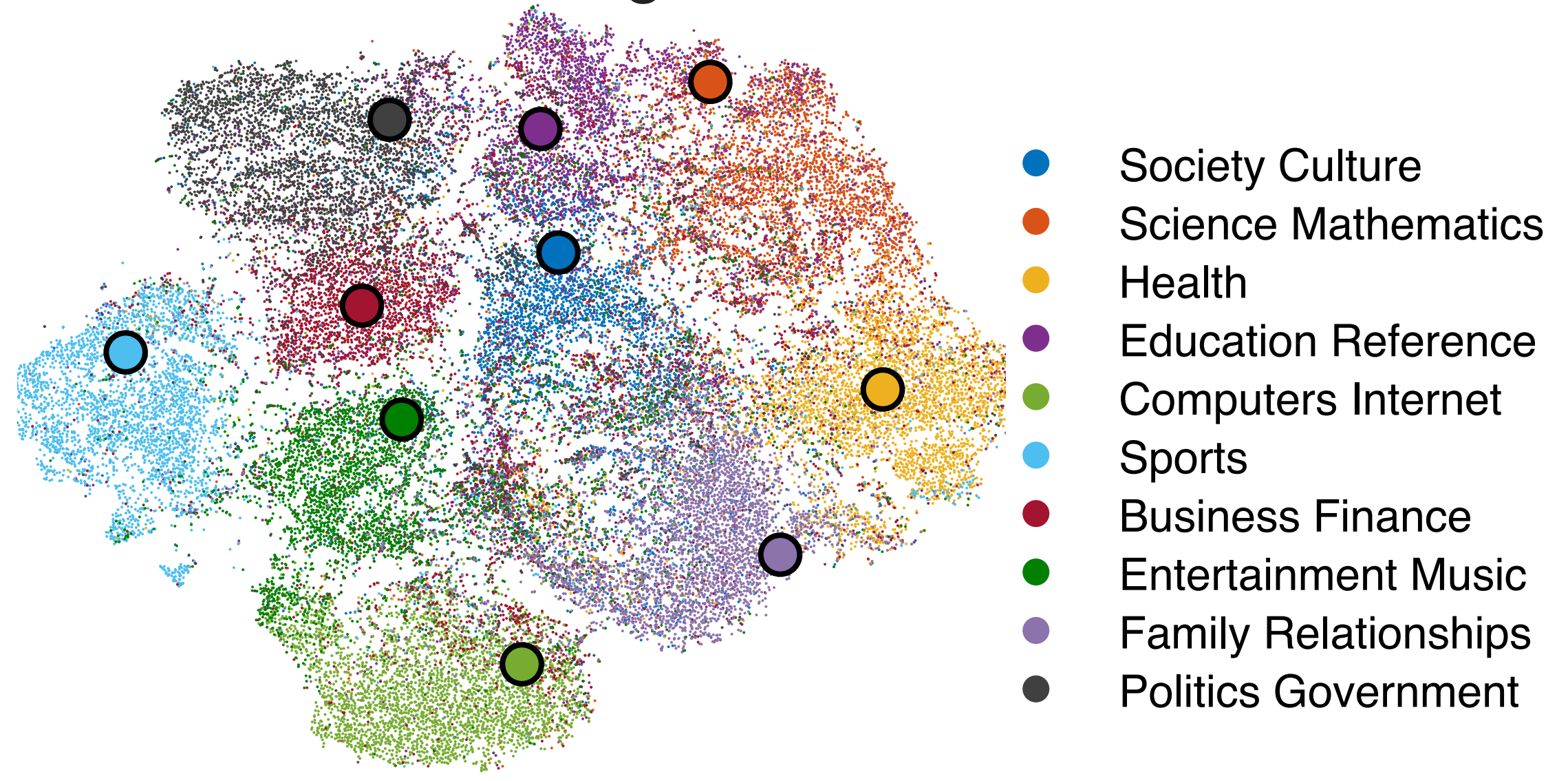
Convergence Speed



Cos similarity matrix: text and label embedding



t-SNE plot of joint embedding, point clouds for texts and large dots for labels


Society Culture
Science Mathematics
Health
Education Reference
Computers Internet
Sports
Business Finance
Entertainment Music
Family Relationships
Politics Government