A Generalized Language Model in Tensor Space

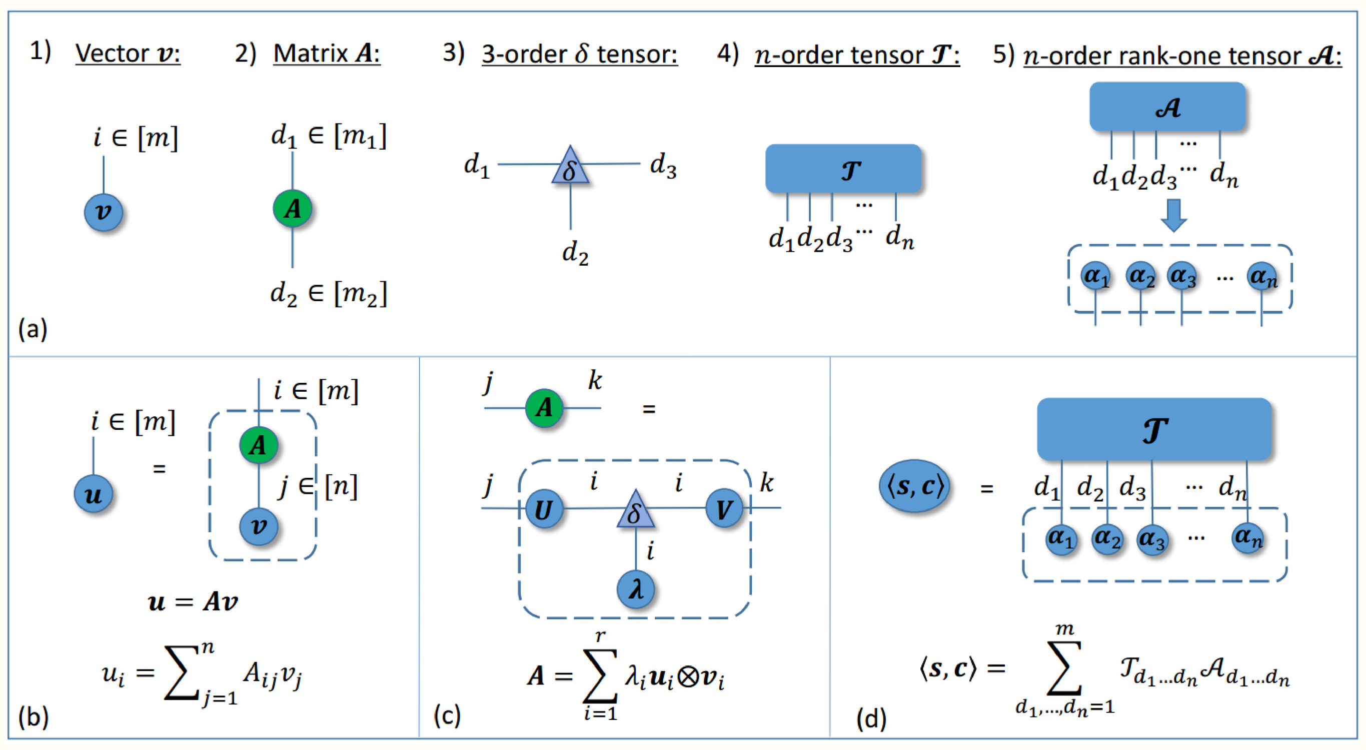
Lipeng Zhang [1], Peng Zhang [1,*], Xindian Ma[1] , Shuqin Gu [1], Zhan Su [1], Dawei Song [2]
1 College of Intelligence and Computing, Tianjin University, China
2 School of Computer Science and Technology, Beijing Institute of Technology, China

1. Introduction

Motivation

- The existing methods usually adopt relatively low-order tensors, which have limited expressive power in modeling language.
- We propose a language model based on relatively high-order tensor representation——Tensor Space Language Model (TSLM).
- Challenges
  - To derive an effective solution for such high order representation;
  - To demonstrate such a solution is a general approach for language modeling;
  - To solve that such a high-order tensor contains exponential magnitude of parameters.

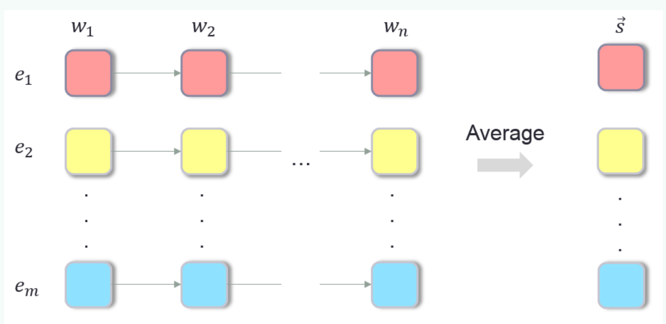
2. Tensor Network



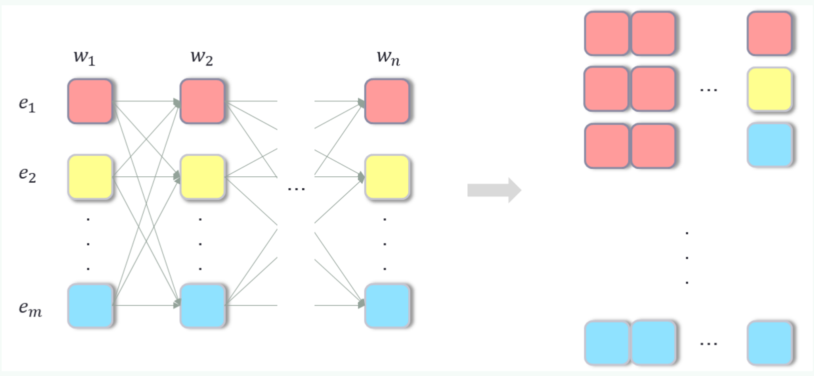
3. Tensor Space Language Model and The Generalization

TSLM Basic Representation

Hypotheses: A sentence has n words. Each word has m semantic meanings.



The sentence still has m meanings.



we will get m^n semantic combinations.

- How to represent a single word

$$\boldsymbol{w}_i = \sum_{d_i=1}^{m} \alpha_{i,d_i} \boldsymbol{e}_{d_i}$$

- How to represent a sentence

$$\boldsymbol{s} = \boldsymbol{w}_1 \otimes \cdots \otimes \boldsymbol{w}_n = \sum_{d_1,\cdots,d_n=1}^{m} \mathcal{A}_{d_1\cdots d_n} \boldsymbol{e}_{d_1} \otimes \cdots \otimes \boldsymbol{e}_{d_n}$$

- Assume that each sentence s_i appears with a probability p_i. We can denote the corpus as:

$$\boldsymbol{c} = \sum_i p_i \boldsymbol{s}_i = \sum_{d_1,\cdots,d_n=1}^{m} \mathcal{T}_{d_1\dots d_n} \boldsymbol{e}_{d_1} \otimes \cdots \otimes \boldsymbol{e}_{d_n}$$

- The sentence probability:

$$p(\boldsymbol{s}) = \langle \boldsymbol{s}, \boldsymbol{c} \rangle = \sum_{d_1,\cdots,d_n=1}^{m} \mathcal{T}_{d_1\dots d_n} \mathcal{A}_{d_1\dots d_n}$$

The Generalization of N-Gram Language

- **Claim 1**:
  In our TSLM, when we set the dimension of vector space $m = |\mathrm{V}|$ and each word w as an one-hot vector, the probability of sentence s consist of words $d_1, \dots, d_n$ in vocabulary is the entry $\mathcal{T}_{d_1\dots d_n}$ of tensor $\mathcal{T}$.
- **Claim 2**:
  In our TSLM, we define the word sequence $w_1^i = (w_1, w_2, \dots, w_i)$ with length i as:

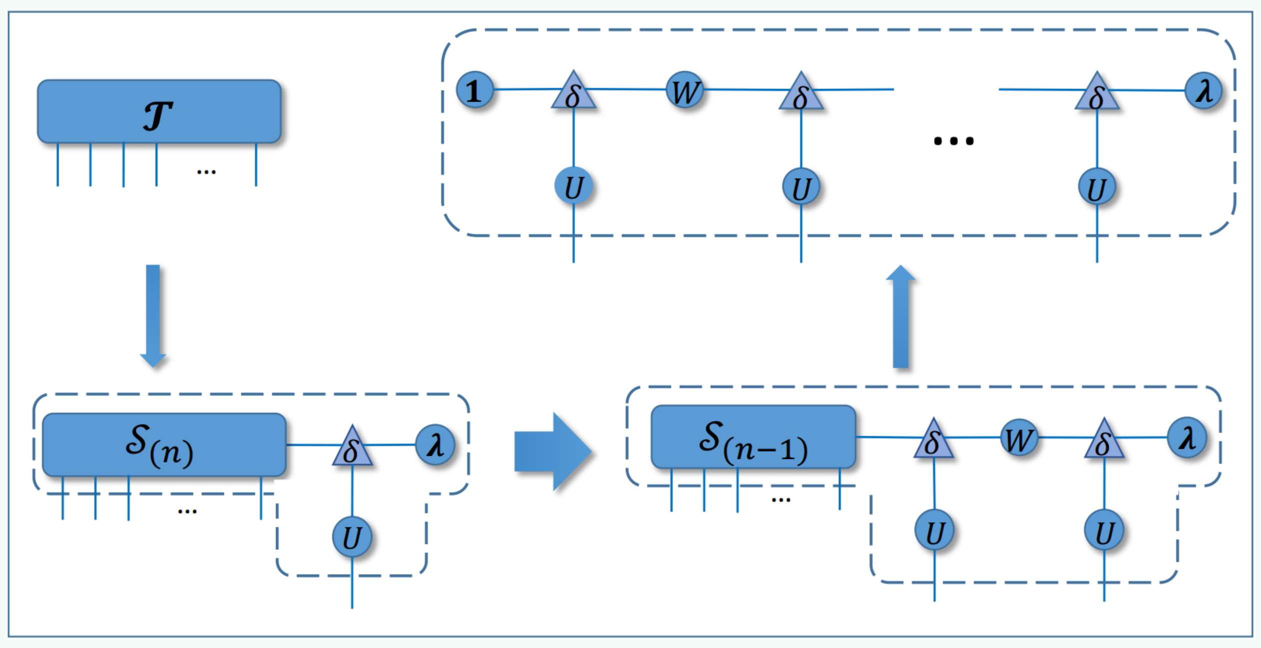
$$\boldsymbol{w}_1^i = \boldsymbol{w}_1 \otimes \cdots \otimes \boldsymbol{w}_i \otimes \mathbf{1}_{i+1} \otimes \cdots \otimes \mathbf{1}_n$$

  Which means that the sequence w_1^i is padded via using full one vector $\mathbf{1}$. Then, the probability $p(w_1^i) = \langle \boldsymbol{w}_1^i, \boldsymbol{c} \rangle$.
- In TSLM, the conditional probability $p(w_i|w_1^{i-1})$ can be computed as:

$$p(w_i|w_1^{i-1}) = \frac{p(w_1^i)}{p(w_1^{i-1})} = \frac{\langle \boldsymbol{w}_1^i, \boldsymbol{c}\rangle}{\langle \boldsymbol{w}_1^{i-1}, \boldsymbol{c}\rangle}$$

4. Deriving Recursive Language Modeling Process from TSLM

Recursive Tensor Decomposition



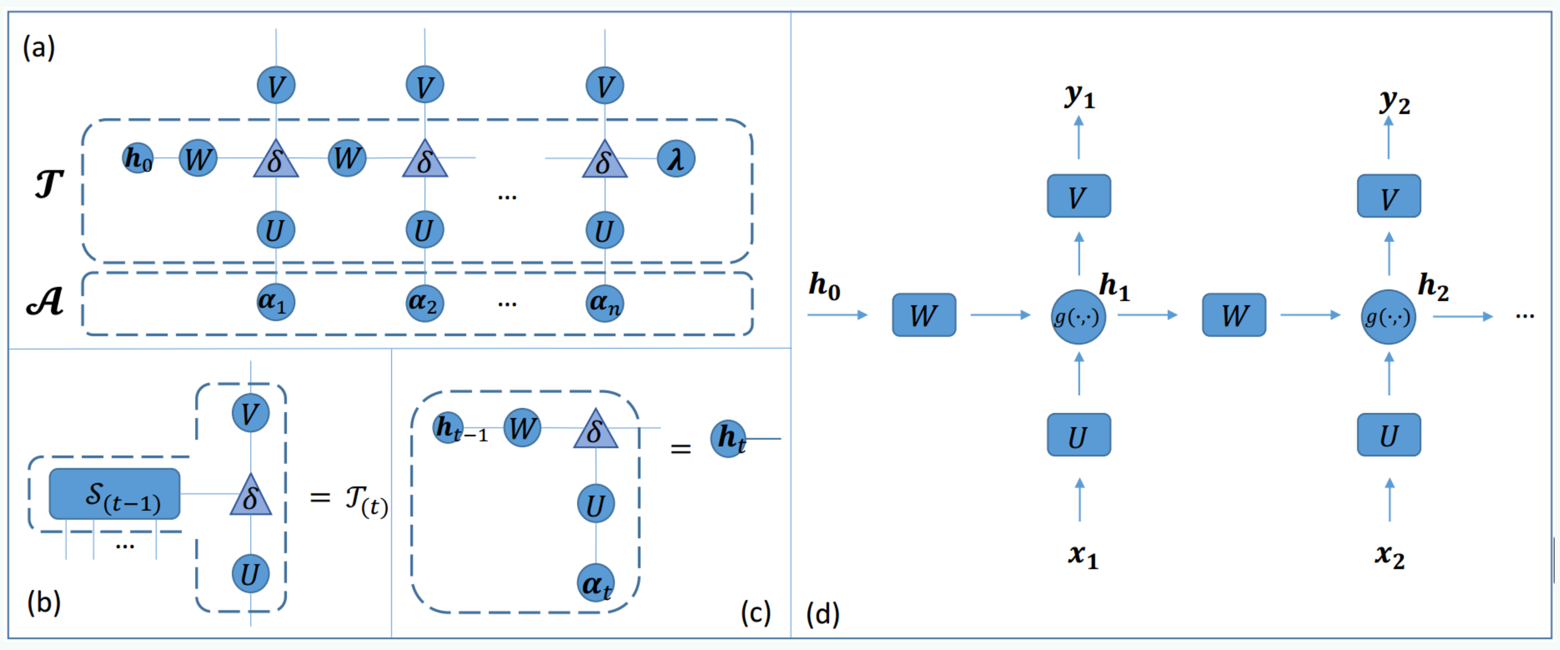
- Decomposition :

$$\mathcal{T} = \sum_{i=1}^{r} \lambda_i \mathcal{S}_{(n),i} \otimes \boldsymbol{u}_i$$

$$\mathcal{S}_{(n),k} = \sum_{i=1}^{r} W_{k,i} \mathcal{S}_{(n-1),i} \otimes \boldsymbol{u}_i$$

$$\cdots$$

$$\mathcal{S}_{(1)} = \mathbf{1}$$



- Denote : $\boldsymbol{h}_0 = W^{-1}\mathbf{1}$
- Computing $\boldsymbol{h}_t$ recursively :

$$\boldsymbol{h}_t = W \boldsymbol{h}_{t-1} \odot U \boldsymbol{\alpha}_t$$

- Constructing a tensor mapping to vocabulary by a matrix $V \in \mathbb{R}^{r \times |\mathrm{V}|}$:

$$\mathcal{T}_{(t),k} = \sum_{i=1}^{r} V_{k,i} \mathcal{S}_{(t-1),i} \otimes \boldsymbol{u}_i$$

- Therefore, computing the conditional probability recursively:

$$p(w_t|w_1^{t-1}) = softmax(\boldsymbol{y}_t)$$

$$\boldsymbol{y}_t = V \boldsymbol{h}_t$$

$$\boldsymbol{h}_t = g(W \boldsymbol{h}_{t-1}, U \boldsymbol{\alpha}_t)$$

$$g(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a} \odot \boldsymbol{b}$$

5. Empirical Evaluation and Conclusion

Experiment Results

	PTB				WikiText-2			
Model	**Hidden size**	**Layers**	**Valid**	**Test**	**Hidden size**	**Layers**	**Valid**	**Test**
KN-5(Mikolov and Zweig 2012)	-	-	-	141.2	-	-	-	-
RNN(Mikolov and Zweig 2012)	300	1	-	124.7	-	-	-	-
LSTM(Zaremba, Sutskever, and Vinyals 2014)	200	2	120.7	114.5	-	-	-	-
LSTM(Grave, Joulin, and Usunier 2016)	1024	1	-	82.3	1024	1	-	99.3
LSTM(Merity et al. 2017)	650	2	84.4	80.6	650	2	108.7	100.9
RNN†	256	1	130.3	124.1	512	1	126.0	120.4
LSTM†	256	1	118.6	110.3	512	1	105.6	101.4
TSLM	256	1	**117.2**	**108.1**	512	1	**104.9**	**100.4**
RNN+MoS†(Yang et al. 2018)	256	1	88.7	84.3	512	1	85.6	81.8
TSLM+MoS	256	1	**86.4**	**83.6**	512	1	**83.9**	**81.0**

Conclusion

- We propose a novel language model, named Tensor Space Language Model, aiming to consider high-order dependencies of words via tensors and tensor networks.
- We prove that TSLM is a generalization of the n-gram language model.
- We can derive a recursive calculation of conditional probability for language modeling via tensor decomposition in TSLM.