Learning FOFE based FNN-LMs with noise contrastive estimation and part-of-speech features

Junfeng Hou, Shiliang Zhang, Lirong Dai

National Engineering Laboratory for Speech and Language Information Processing

University of Science and Technology of China, Hefei, China

Abstract

- Extend FOFE based FNN-LMs:
  - ✧ Add *transitions* of part-of-speech (POS) tags as additional features
  - ✧ Train with noise contrastive estimation (NCE)
- Better performance on PTB & LTCB:
  - ✧ *Transitions* of POS is more meaningful than POS
  - ✧ Dramatically speedup the training speed

Background

- FOFE based FNN-LMs
  - ✧ Encodes each partial sequence (history) based on a simple recursive formula (with $z_0 = 0$) as:

$$\mathbf{z}_t = \alpha \cdot \mathbf{z}_{t-1} + \mathbf{e}_t \quad (1 \le t \le T)$$

  - ✧ A simple example:

$$A=[1,\ 0,\ 0],\ B=[0,\ 1,\ 0],\ C=[0,\ 0,\ 1]$$

$$\{ABC\}=\{\alpha^2,\ \alpha,\ 1\},\ \{ABCBC\}=\{\alpha^4,\ \alpha^3+\alpha,\ \alpha^2+1\}$$

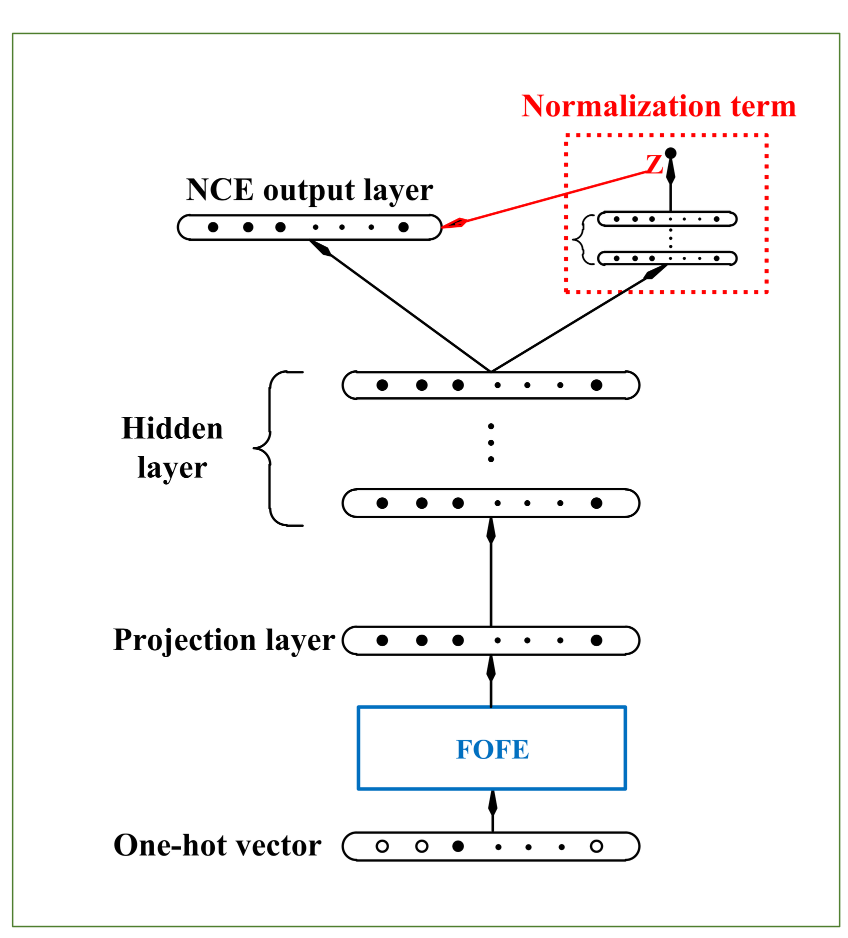
- NCE
  - ✧ NNLM can be trained by the unnormalized probabilities without computing the normalization term of softmax layer
  - ✧ The normalization term is fixed for simplicity

$$p(w\,|\,h,\boldsymbol{\theta})=\frac{1}{\boldsymbol{Z}_{\theta}(h)}\exp(s_{\theta}(w,h))\approx p_{\theta_0}^{h}(w)/\boldsymbol{Z}^{h}$$

$$\boldsymbol{Z}_{\theta}(h)=\sum_{w'}\exp(s_{\theta}(w',h))$$

Main work

- NCE
  - ✧ Context dependent normalization term is used to replace the constant one
  - ✧ Easily scale to huge numbers of observed contexts encountered by the models with large context sizes



- Transitions of POS feature:
  - ✧ Model more *variations* of syntactic information
  - ✧ FOFE code:

$$\begin{bmatrix}\mathbf{z}_{w_t}\\ \mathbf{z}_{pos_t}\end{bmatrix}=\begin{bmatrix}\alpha_{w_t}\cdot\mathbf{z}_{w_t}\\ \alpha_{pos_t}\cdot\mathbf{z}_{pos_t}\end{bmatrix}+\begin{bmatrix}\mathbf{e}_{w_t}\\ \mathbf{e}_{pos_{t-1,t}}\end{bmatrix}\quad(1\le t\le T)$$

  - ✧ A simple example:

Sentence : Only a few books fell in the reading room

POS (totally 43): RB DT JJ NNS VBD IN DT NN NN

TiePOS (totally 1455): <s>_RB RB_DT DT_JJ JJ_NNS NNS_VBD VBD_IN IN_DT DT_NN NN_NN

Experiment

- Two benchmark tasks:

i) Penn Treebank (PTB)

ii) Large Text Compression Benchmark (LTCB)

Corpus	Train	Valid	Test	vocabulary
PTB	930k	74k	82k	10k
LTCB	153M	8.9M	8.9M	80k

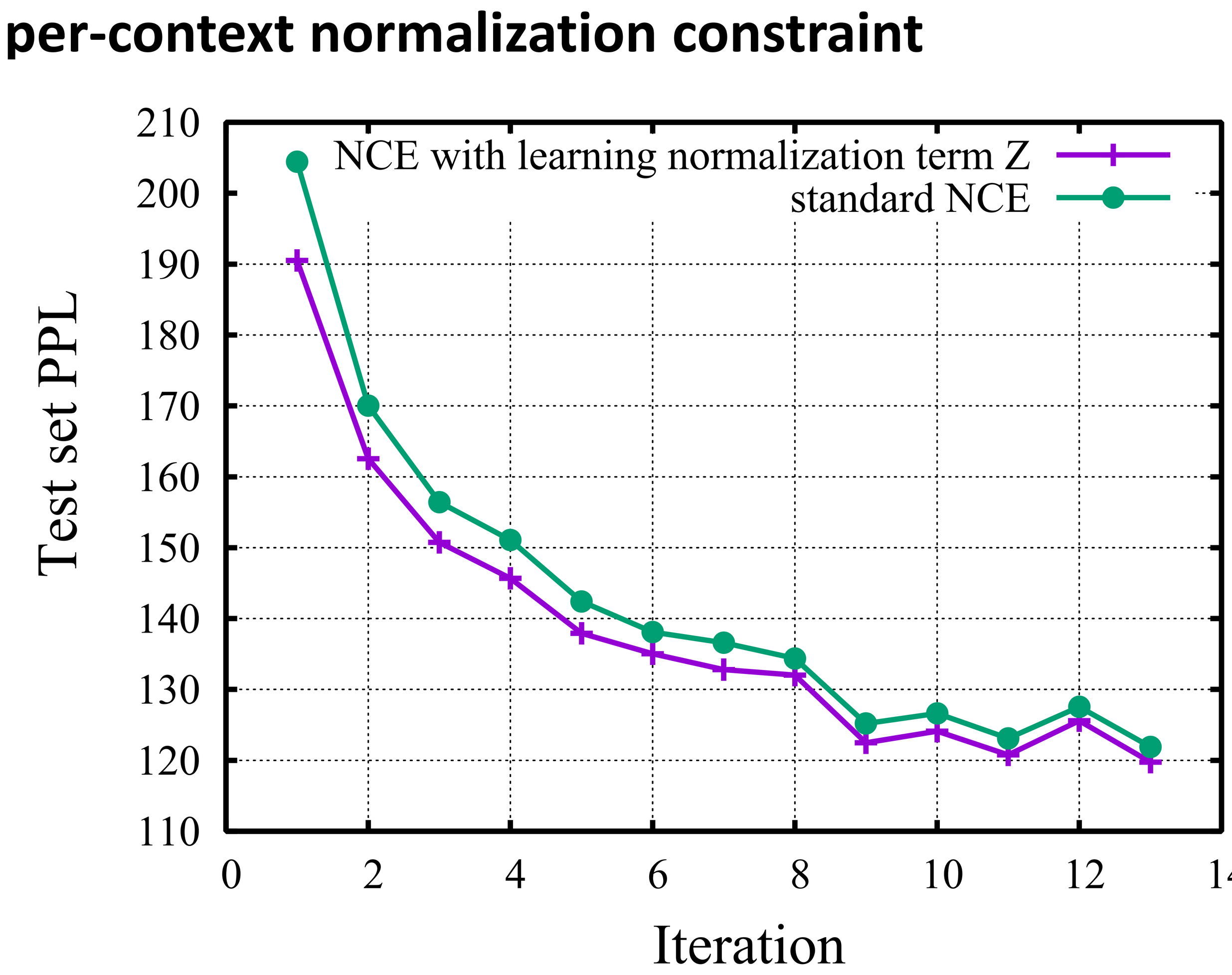
- PTB experiments

Model	Test PPL
trigram FNNLM (Zhang, 2015)	131
RNNLM (Mikolov, 2011)	123
2nd-order FOFE-FNNLM (Zhang, 2015)	108
+MonoPOS	105
+FOFE-MonoPOS	102
+FOFE-tiePOS	*100*

- LTCB experiments

Model	architecture	Test PPL
FNNLM (Zhang, 2015)	[2*200]-400x2-80k	155
RNNLM (Zhang, 2015)	[1*600]-80K	112
FOFE FNNLM (Zhang, 2015)	[2*200]-400x2-80K	112
+FOFE-monoPOS	[2*250]-400x2-80K	109
+FOFE-tiePOS	[2*300]-400x2-80K	*103*
+NCE	[2*200]-400x2-80K	122
+FOFE-monPOS+NCE	[2*250]-400x2-80K	118
+FOFE-tiePOS+NCE	[2*300]-400x2-80K	114
+CD-norm NCE	[2*200]-400x2-80K/400x2-1	120

- ✧ 20x times faster training speed with NCE
- ✧ Model has so many free parameters to meet the approximate per-context normalization constraint



Conclusion

1. Transitions of POS feature can further improve the performance of the FOFE based FNN-LMs

2. Constant normalization term is enough for NCE training and NCE can train the model much faster

Reference

- FOFE : S. Zhang, “The fixed-size ordinally-forgetting encoding method for neural network language”, ACL 2015
- NCE : A. Mnih, “A fast and simple algorithm for training neural probabilistic language models”, arXiv 2012