Long Short-Term Memory with Dynamic Skip Connections

Tao Gui, Qi Zhang, Lujun Zhao, Yaosong Lin, Minlong Peng, Jingjing Gong, Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
Shanghai Insitute of Intelligent Electroics & Systems
{tgui16, qz, ljzhao16, mlpeng16, yslin18, jjgong, xjhuang}@fudan.edu.cn

Introduction

LSTM experiences difficulty in capturing long-term dependencies. In this work, we tried to alleviate this problem by introducing a dynamic skip connection, which can learn to directly connect two dependent words. Since there is no dependency information in the training data, we propose a novel reinforcement learning-based method to model the dependency relationship and connect dependent words. The proposed model computes the recurrent transition functions based on the skip connections, which provides a dynamic skipping advantage over RNNs that always tackle entire sentences sequentially.

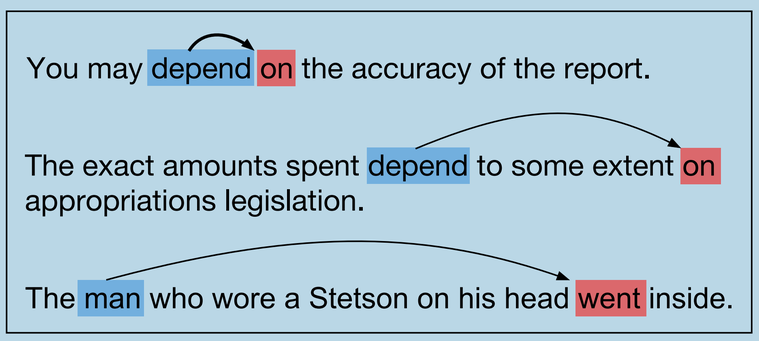


Figure 1: Examples of dependencies with variable length in the language. The same phrase "*depend on*" in different sentences would have dependencies with different lengths. The use of clauses also makes the dependency length uncertain. Therefore, the models using a plain LSTM or an LSTM with fixed skip connections would be difficult to capture such information.

Named Entity Recognition

Model	F1
Huang, Xu, and Yu (2015)	90.10
Chiu and Nichols (2015)	90.91±0.20
Lample et al. (2016)	90.94
Ma and Hovy (2016)	91.21
Strubell et al. (2017)†	90.54 ± 0.18
Strubell et al. (2017)	90.85 ± 0.29
LSTM, fixed skip = 3	91.14
LSTM, fixed skip = 5	91.16
LSTM with attention	91.23
LSTM with dynamic skip	**91.56**

Figure 2: F1-measure of different methods applied to the CoNLL 2003 dataset. The model that does not use character embeddings is marked with †. "LSTM with attention" refers to the LSTM model using attention mechanism to connect two words.

⇑ Vanishing Gradients ⇑

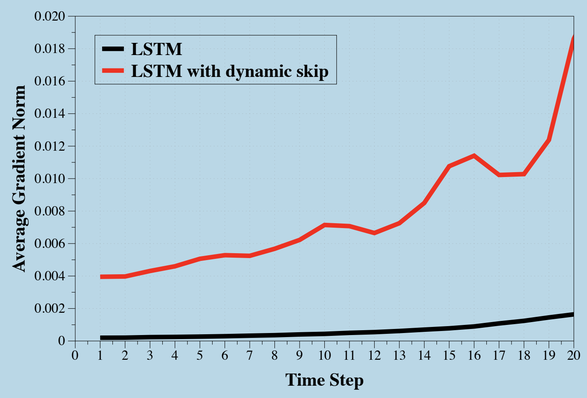


Figure 3: Normalized long-term gradient values $\left\| \frac{\partial L_T}{\partial h_t} \right\|$ tested on CoNLL 2003 dataset. At the initial time steps, the proposed model still preserves effective gradients, which is hundreds of times larger than those in the standard LSTM, indicating that the proposed model have stronger ability to capture long-term dependency.

Dynamic Skip with REINFORCE

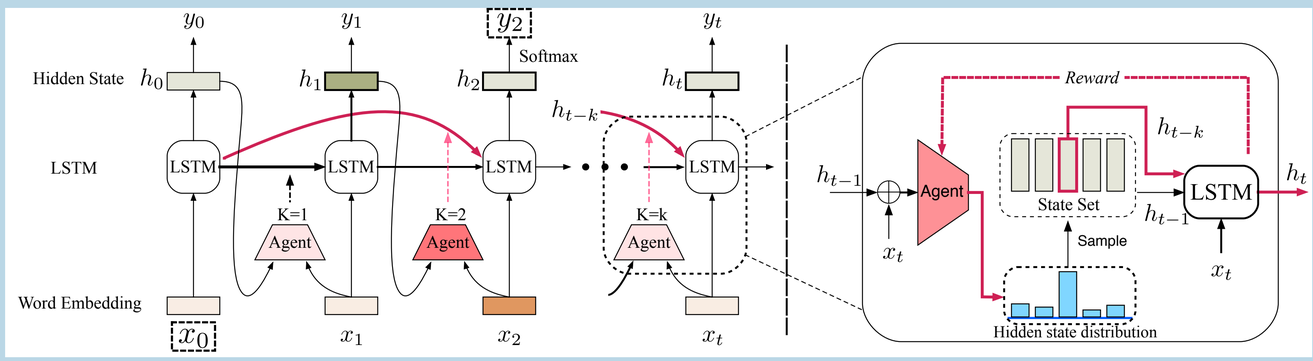


Figure 4: Architecture of the proposed model. At time step t, the agent selects one of the past few states based on the current input x_t and the previous hidden state h_{t-1}. The agent's selections will influence the log-likelihood of the ground truth, which will be a reward or penalty to optimize the agent. Take the phrase "*depend to some extent on*" as an example, the agent should learn to select the hidden state from "*depend*" not "*extend*" to predict "*on*," because selecting "*depend*" receives a larger reward.

Our goal for training is optimizing the parameters of the policy gradient agent θ_a, together with the parameters of standard LSTM and possibly other parameters denoted as θ_l.

$$J_1(\theta_l) = -[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] \quad J_2(\theta_a) = \mathbb{E}_{\pi(a_{1:T})}[R] + H(\pi(a_{1:T}))$$

$$J(\theta_a, \theta_l) = \frac{1}{M}[\sum_{m=1}^{M}(J_1(\theta_l) - J_2(\theta_a))].$$

$$\nabla_{\theta_a} J_2(\theta_a) = \mathbb{E}_{\pi(a_{1:T})}[\sum_{t=1}^{T} \nabla_{\theta_a} \log Pr(a_t|s_t;\theta_a) * (R - \sum_{t=1}^{T} \log Pr(a_t|s_t;\theta_a) - 1)]$$

Language Modeling

Model	Dev.(*PPL*)	Test(*PPL*)	Size
RNN (Mikolov and Zweig 2012)	-	124.7	6 m
RNN-LDA (Mikolov and Zweig 2012)	-	113.7	7 m
Deep RNN (Pascanu et al. 2013)	-	107.5	6 m
Zoneout + Variational LSTM (medium) (Merity et al. 2016)†	84.4	80.6	20 m
Variational LSTM (medium) (Gal and Ghahramani 2016)†	81.9	79.7	20 m
Variational LSTM (medium, MC) (Gal and Ghahramani 2016)†	-	78.6	20 m
Regularized LSTM (Zaremba, Sutskever, and Vinyals 2014)†‡	86.2	82.7	20 m
Regularized LSTM, fixed skip = 3 (Zhang et al. 2016)†	85.3	81.5	20 m
Regularized LSTM, fixed skip = 5 (Zhang et al. 2016)†	86.2	82.0	20 m
Regularized LSTM with attention†	85.1	81.4	20 m
Regularized LSTM with dynamic skip, λ=1, K=5†	**82.5**	**78.5**	20 m
CharLM (Kim et al. 2016)†‡	82.0	78.9	19 m
CharLM, fixed skip = 3 (Zhang et al. 2016)†	83.6	80.2	19 m
CharLM, fixed skip = 5 (Zhang et al. 2016)†	84.9	80.9	19 m
CharLM with attention†	82.2	79.0	19 m
CharLM with dynamic skip, λ=1, K=5†	**79.9**	**76.5**	19 m

⇑Perplexities for different K⇑

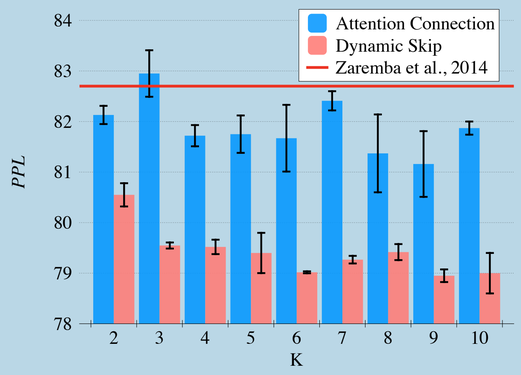
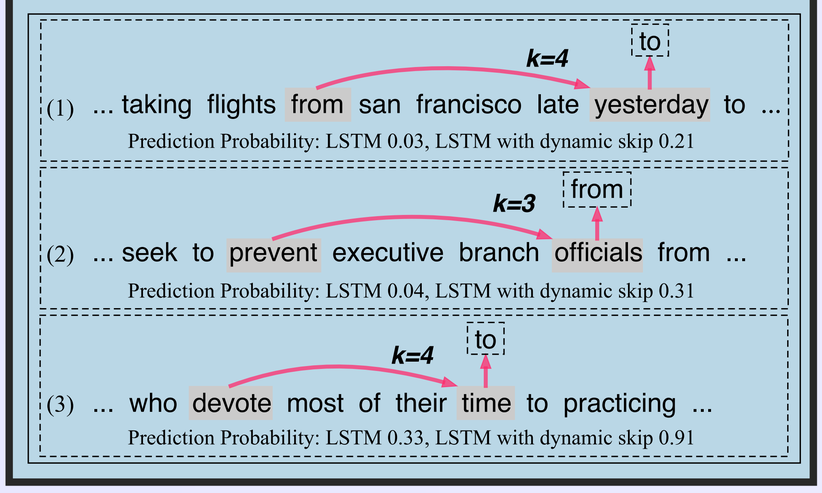


Figure 5: Similar results are shown on λ.

⇑Visualization⇑



Sentiment Analysis

Model	Acc.
LSTM	89.1
LSTM + LA (Chen et al. 2016)	89.3
LSTM + CBA^G (Long et al. 2017)	89.4
LSTM + CBA + LA$_s^G$(*Longetal.*2017)	89.8
LSTM + CBA + LA$_p^G$ (Long et al. 2017)	**90.1**
Skip LSTM (Campos et al. 2017)	86.6
Jump LSTM (Yu, Lee, and Le 2017)	89.4
LSTM, fixed skip = 3 (Zhang et al. 2016)	89.6
LSTM, fixed skip = 5 (Zhang et al. 2016)	89.3
LSTM with attention	89.4
LSTM with dynamic skip, λ=0.5, K=3	**90.1**

Figure 6: Accuracy on the IMDB test set.

Number Prediction

sequence length 11		
Model	**Dev.**	**Test**
LSTM	69.6	70.4
LSTM with attention	71.3	72.5
LSTM with dynamic skip, λ=1, K=10	79.6	80.5
LSTM with dynamic skip, λ=0.5, K=10	**90.4**	**90.5**
sequence length 21		
Model	**Dev.**	**Test**
LSTM	26.2	26.4
LSTM with attention	26.7	26.9
LSTM with dynamic skip, λ=1, K=10	77.6	77.7
LSTM with dynamic skip, λ=0.5, K=10	**87.7**	**88.5**

Figure 7: Accuracies of different methods on number prediction dataset.

Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19) - January 27 – February 1, 2019 - Honolulu, Hawaii, USA