ADL 2020 PROJECT-II REPORT

Mayank Gupta

mayankg@iisc.ac.in

ABSTRACT

A lot of tasks in NLP such as machine translation, sentiment-classification, etc. uses bi-directional LSTM ([1]) based models. In the standard bi-directional LSTM models, both left-to-right and right-to-left models predict hidden features independently, and we concatenate the output from these two networks to get the bi-directional model output. So, the standard left-to-right LSTM model does not take into account future words while taking the current word as input and vice-versa. So in this work, I propose a new bi-directional model such that the left-to-right LSTM model takes into account future words while taking the current word as input. I also pre-train this network for masked language modeling task and next sentence prediction task so that it can be fine-tuned for any down-stream applications.

Index Terms— Bi-directional LSTM, Masked Language Modeling, Pre-training.

1. INTRODUCTION

Language model pre-training has been shown to be effective for improving many NLP tasks ([2],[3]). In [2], authors demonstrated the importance of bidirectional pre-training for language representations and achieved better performance than [3] which uses shallow concatenation of independently trained left-to-right and right-to-left language models and therefore does not utilize left and right context efficiently.

2. TECHNICAL DETAILS

In this work, I implemented a new type of bi-directional LSTM network which better utilizes left and right context while producing an output hidden vector. Main idea is to feed the hidden states of the backward LSTM into the input of the forward LSTM along with current input vector. After implementing this network, I pre-train this network for masked language modeling and next sentence prediction task.

2.1. New LSTM network

This LSTM network takes input sequence $\mathbf{x} = (x_1, x_2, ..., x_T)$ and returns $\mathbf{h} = (h_1, h_2, ..., h_T)$, where $\mathbf{h_t} = (\overrightarrow{h_t}, \overleftarrow{h_t})$ and T represents the total time steps. From the equations below, we can see that output of left-to-right LSTM depends on complete input sequence.

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h}_{t+1}) \tag{1}$$

$$p_t = [x_t, \overleftarrow{h_t}] \tag{2}$$

$$\overrightarrow{h_t} = LSTM(p_t, \overrightarrow{h}_{t-1}) \tag{3}$$

2.2. Model Pre-training

Unsupervised pre-training of the network is done using 2 tasks: Masked Language Modeling and Next Sentence Prediction. These both tasks are same as in [2], but this model uses vocabulary of 10,000 words instead of 30,000 sub-words as used by [2]. In MLM task, we mask 15% of the symbols randomly and at the output we predict the masked words. In NSP, we are given 2 sentences and have to classify if the second sentence is next to first sentence or not.

3. RESULTS

2-layer networks are used for experiments on IMDB and SNLI datasets, 1-layer networks are used for pre-training experiment on WikiText-2. New LSTM uses around 1.2x parameters than normal LSTM network in experiments. Loss of MLM task is reported in the table for WikiText-2 , loss for NSP task was similar for both the models. We can see from the table that performance of both models is similar.

Dataset used	New LSTM		LSTM	
	Loss	Accuracy	Loss	Accuracy
IMDB	0.351	85.71	0.368	84.3
SNLI	0.0023	77.8	0.0023	77.42
WikiText-2	5.803	-	5.837	-

Table 1. Comparison of models.

4. CONTRIBUTIONS

Proposed a new type of LSTM architecture which better utilizes left and right context while predicting an output hidden

Thanks to Google for colab platform.

state. I also implemented pre-training of new LSTM architecture using Masked Language modeling and Next Sentence Prediction so that it can be fine-tuned for downstream task.

5. RESOURCES

1. Codebase: https://github.com/bentrevett/pytorch-sentiment-analysis
2. http://d2l.ai/

6. REFERENCES

[1] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.