LSTMs for Semi-Supervised Text Classification

Abhay

Abstract

The following document describes in brief the model by Sachan, Zaheer, and Salakhutdinov 2019. It suggests very simple and robust approach for the task of text classification with both supervised and semi-supervised approaches. On contrary to the complex schemes and models suggested by various researches, this model is based on a simple BiLSTM model, trained with cross-entropy loss. Also, this incorporates the use of various losses via mixed objective function and is able to produce state of the art results for text classification tasks.

1 Introduction

Text classification also known as text tagging or text categorization is the process of categorizing text into organized groups. Text Classification has always been a defining problem in the field of natural language processing. It covers a very broad area of NLP in terms of applications like sentiment analysis, email filtering, language detection and many more. Earliest approaches were based on the extraction of bag of words features followed by either Naive Byes(McCallum, Nigam, et al. 1998) or linear SVM (Joachims 1998). Recently, RNN and CNN models were introduced to utilize the word order and grammatical structure as shown by Kim 2014. The previous state of the art techniques either use pretrained LSTMs or complex computationally expensive model. This model proposes a mixed objective function for semi supervised learning without the requirement of any pretraining step.

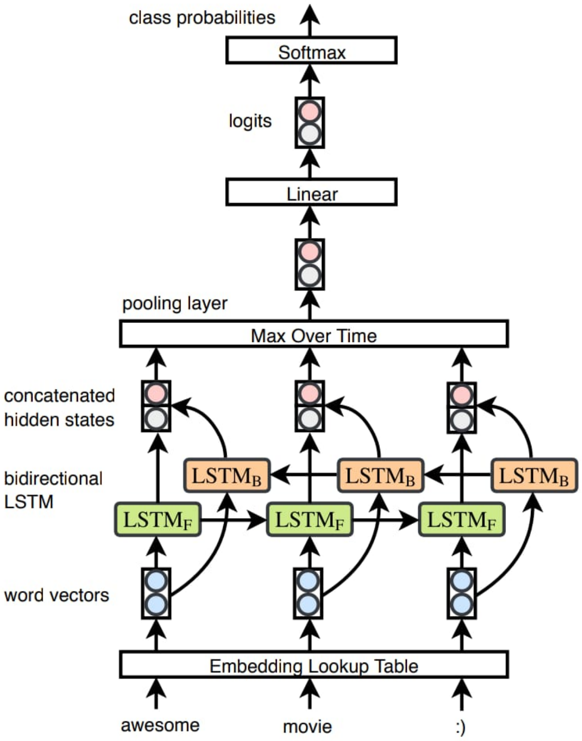

Figure 1: Model

2 Details

2.1 Model Architecture

The Figure 1 represent the classification model

2.2 Methods

2.2.1 Supervised training

Let there be m labeled examples in the training set that are denoted as $(x^{(1)}, y^{(1)}), ..., (x^{(m)}, y^{(m)})$, where $x^{(i)}$ represents a document's word sequence, $y^{(i)}$ represents class label such that $y^{(i)} \in (1, 2, ..., K)$. For supervised training of the classification model, we make use of two methodologies: maximum likelihood estimation and adversarial training:-

$$L_{ML}(\theta) = -1/m_l * \sum_{i=1}^{m_l} \sum_{j=1}^{K} z(y^{(i)} = k) logp(y^i = k|x^i; \hat{\theta}) \tag{1}$$

$$L_{AT}(\theta) = -1/m_l * \sum_{i=1}^{m_l} \sum_{k=1}^{K} z(y^{(i)} = k) logp(y^i = k|v^* i; \hat{\theta}) \tag{2}$$

where v^* corresponds to adversarial embedding corresponding to v

2.2.2 Unsupervised Training

The model also uses experiments two unsupervised methods like Virtual Adversarial training and Entropy Minimization.

$$L_{EM}(\theta) = -1/m * \sum_{i=1}^{m} \sum_{k=1}^{K} p(y^{(i)} = k|x^{(i)}) logp(y^i = k|x^(i)) \tag{3}$$

where $m = m_l + m_u$ and dependence on θ is suppressed

$$L_{VAT}(\theta) = 1/m * \sum_{i=1}^{m} D_{KL}(p(.|v^{(i)}; \theta)||p(.|v^{(*i)}; \theta) \tag{4}$$

where $m = m_l + m_u$

2.2.3 Mixed Objective Function

Thus the proposed mixed objective function using λ_{ML}, λ_{AT}, λ_{EM} and λ_{VAT} as parameters become

$$L_{mixed} = \lambda_{ML} L_{ML} + \lambda_{VAT} L_{AT} + \lambda_{EM} L_{EM} + \lambda_{VAT} L_{VAT} \tag{5}$$

2.3 Results

The model was experimented with the datasets shown in Table 1 The error rates (%) when model is trained using mixed objective functon are shown in Table 2

Dataset	Train	Test	K	L
ACL IMDB	25000	25000	2	268
Elec	25000	25000	2	125
AG-News	120,000	7,600	4	46

Table 1: Summary statistics for text classification datasets; K = number of classes; L = average length of a document.

Model	ACl IMDB	Elec	AG-News
LSTM	5.91	5.40	6.78
oh-LSTM	5.94	5.55	6.57
ULMFit	4.60	-	5.01
L_{Mixed}	4.32	5.24	4.9

Table 2: Error rates (%) of L_{Mixed} compared with previous best methods

3 Conclusion

Sachan, Zaheer, and Salakhutdinov 2019 has shown that a simple BiLSTM model using a combination of various loss functions reported state of the art performance on several text classification datasets and the mixed objective function can be generalise to other tasks such as relation extraction.

References

Joachims, Thorsten (1998). "Text categorization with support vector machines: Learning with many relevant features". In: *European conference on machine learning*. Springer, pp. 137–142.

Kim, Yoon (2014). "Convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1408.5882*.

McCallum, Andrew, Kamal Nigam, et al. (1998). "A comparison of event models for naive bayes text classification". In: *AAAI-98 workshop on learning for text categorization*. Vol. 752. 1. Citeseer, pp. 41–48.

Sachan, Devendra Singh, Manzil Zaheer, and Ruslan Salakhutdinov (2019). "Revisiting lstm networks for semi-supervised text classification via mixed objective function". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 6940–6948.