Comparison of different LSTM models on UCI News aggregator Dataset

AVNI SHARMA
Dept. Electrical and computer engineering
Michigan State University
East Lansing, 48823 USA
sharm104@msu.edu

***Abstract*—In this work two slim LSTM models are compared with different activation functions and learning rates. The data is balanced and 10000 counts of news titles have been taken from each category. The model also uses a layer of bidirectional LSTM. The two LSTM models are evaluated on the UCI News aggregator data set.**

I. Introduction

Long-Short-Term Memory(LSTM) models are a type of Recurrent Neural Networks(RNNs). They have the ability to learn over long sequences through the use of gates. Gates regulate the information flow of the network. RNNs comes up with a lot of set backs like :

- Short term memory where they tend to discard information from earlier time steps as moving on towards later in the network this may result in the loss of important information permanently.
- Gradients are important in order to update weights in a neural network however if the gradient value becomes too small it doesn't contribute towards the learning process at all. RNNs pose a vanishing gradient probelem as they back propagate through time.
- RNNs can also pose an exploding gradient problem where they assign unreasonably high important to the weights.

To solve these problems, researchers proposed many improved networks, such as the ESN (Echo State Network), the Gated recurrent units(GRU), and so on. And one of the most successful applications is LSTM which accumulates long-term relationships between distant nodes by designing weight coefficients as per the connection. An LSTM cell consists of 3 main components input gate, forget gate and output gate.

An LSTM unit utilizes a "memory" cell (denoted by C_t) that decides whether the 'information' is useful or not and a gating mechanism that contains three non-linear gates: (i) an input (denoted by i_t, (ii) an output (denoted by $o_t$and (iii) a forget gate (denoted by f_t The gates regulate the flow of signals into and out of the cell to adjust long-term dependencies effectively and achieve successful RNN training. The standard equations for LSTM memory blocks are given as given by equations(1)-(6).

In order to make the network more efficient the network can be trained in both time directions simultaneously using bidirectional RNNs. Bidirectional RNNs have proved to overcome the shortcomings offered due to conventional RNNs[].

$$i_t = \sigma_{in}(U_i h_{t-1} + W_i x_l + b_i) \tag{1}$$
$$f_t = \sigma_{in}(U_f h_{t-1} + W_f x_t + b_f) \tag{2}$$
$$o_t = \sigma_{in}(U_o h_{t-1} + W_o x_t + b_o) \tag{3}$$
$$\tilde{c}_t = \sigma(U_c h_{t-1} + W_c x_t + b_c) \tag{4}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{5}$$
$$h_t = o_t \odot \sigma(c_t) \tag{6}$$

Fig. 1. Network Equations for Basic LSTM

Basically the Bidirectioal RNN splits the RNN states into two states, forward and backward states, the input output of the states are not connected. Basic structure is depicted in the .

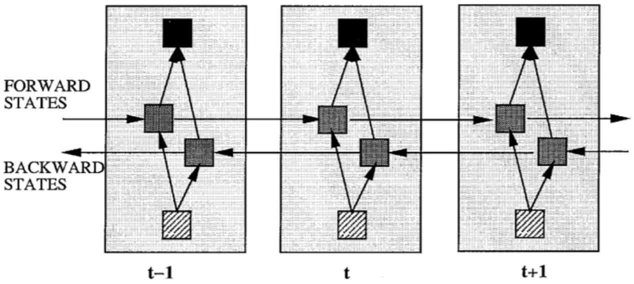

Fig. 2. Network architecture for the bidirectional RNN

The Bidirectional RNN (BRNN) was successfully tested by [1] on the TIMIT dataset wth BRNN structure giving out the best performance. BRNN along with LSTM have peen adopted by various researched around the globe and it have proven to given significant results. Thi-Nga Ho et al. [2] implemented Bidirectional LSTM (Bi-LSTM)network for Sentence Unit Detection in Automatic Speech Recognition transcription.Ali Mert Ertugrul et al. [3] utilized Bi-LSTM to classify Movie genre from plot summaries where Bi-LSTM proved to out perform basic Recurrent Neural Networks (RNNs) and Logistic Regression (LR) as a baseline. M Riza Alifi [4] utilize Bi-Lstm for information extraction of traffic Condition from Social Media.

II. Literature review

Junyoung Chung et al. [5] examines and evaluates three different recurrent neural networks, that are: LSTM-RNN, GRU-RNN and tanhRNN on the task of sequence modeling on two datasets: polyphonic music data and raw speech signal data. The results show that the convergence of gating units (GRU-RNN and LSTM-RNN) are much faster and their final solutions tend to be better compared with the traditional tanh-RNN on both dataset. In order to compare the performance of LSTM and GRUs more concretely, paper [8] examines the performance of LSTM and GRU on other datasets and also find other three best architecture discovered by the search procedure (named MUT1, MUT2, and MUT3). They also evaluated an LSTM without input gates (LSTM-i), an LSTM without output gates (LSTMo), and an LSTM without forget gates (LSTM-f). The mainly results are that: 1) GRU outperformed the LSTM on all tasks with exception of language modeling. 2) the LSTM with large forget bias outperformed both LSTM and the GRU on almost all tasks. 3) that the LSTM-i and LSTM-o achieved the best results on the music dataset when dropout is used. Therefore, the LSTM-RNN and its derived structure proved to be the optimal option to deal with the sequence model.

III. Simplified LSTM model

While the LSTM model has demonstrated impressive performance in applications involving sequence-to-sequence relationships, a criticism of the standard LSTM resides in its relatively complex model structure with 3 gating signals and the number of its relatively large number of parameters [see eqn (5)].In order to reduce the adaptive parameter number in each gate, F. M. Salem [6] proposed simplifications to the standard LSTM result in five LSTM variants by removing some of the parameters in the selected blocks The gates in fact replicate the parameters in the cell. Here, three simplifications to the standard LSTM result in two LSTM variants we refer to them here as simply, LSTM10 and LSTM11. There variants are obtained by removing signals, and associated parameters in the gating eqns (1)-(3). For uniformity and simplicity, we apply the changes to all the gates identically.

A. The LSTM model 10:

Model has No Input Signal and No Bias and can be given by the following equations

$$i_t = \sigma_{in}(u_i \odot h_{t-1}) \tag{8}$$
$$f_t = \sigma_{in}(u_f \odot h_{t-1}) \tag{9}$$
$$o_t = \sigma_{in}(u_o \odot h_{t-1}) \tag{10}$$
$$\tilde{c}_t = \sigma(u_c \odot h_{t-1} + W_c x_t + b_c) \tag{11}$$
$$c_t = f_t \odot c_{t-1} \odot \tilde{c}_t) \tag{12}$$
$$h_t = o_t \odot \sigma(c_t) \tag{13}$$

Fig. 3. Network equations for LSTM model 10

B. The LSTM model 11:

Model has No Input Signal and No Bias and can be given by the following equations

$$i_t = \sigma_{in}(u_i \odot h_{t-1} + b_i) \tag{14}$$
$$f_t = \sigma_{in}(u_f \odot h_{t-1} + b_f) \tag{15}$$
$$o_t = \sigma_{in}(u_o \odot h_{t-1} + b_o) \tag{16}$$
$$\tilde{c}_t = \sigma(u_c \odot h_{t-1} + W_c x_t + b_c) \tag{17}$$
$$c_t = f_t \odot c_{t-1} \odot \tilde{c}_t) \tag{18}$$
$$h_t = o_t \odot \sigma(c_t) \tag{19}$$

Fig. 4. Network equations for LSTM model 10

IV. Dataset

The dataset comes from the UCI machine learning repository provided by the Artificial Intelligence Lab at Faculty of Engineering, Roma Tre University in Italy. The dataset contains headlines, URLs, publisher and categories for 422,937 news stories collected by a news web aggregator between March 10th, 2014 and August 10th, 2014. The columns of the dataset contains the following :
ID: Numeric ID
TITLE: News title
URL: Url
PUBLISHER: Publisher name
CATEGORY: News category (b = business, t = science and technology, e = entertainment, m = health)
STORY: Alphanumeric ID of the cluster that includes news about the same story
HOSTNAME: Url hostname
TIMESTAMP: Approximate time the news was published, as the number of milliseconds since the epoch 00:00:00 GMT, January 1, 1970

A. Data Balancing

The given dataset was highly unbalanced. Training on unbalanced dataset might increase the efficiency for entertainment news and reduce for other according to their counts in the training dataset. Fig5 shows the count of news title under different categories. the dataset is balanced by taking 10000 counts from each category fig10 shows te count of news title under different category after the data is balanced.

B. Text processing

All the words were turned into lower format , and the characters were restricted with only English letters and numbers in order to reduce the amount of generated features.

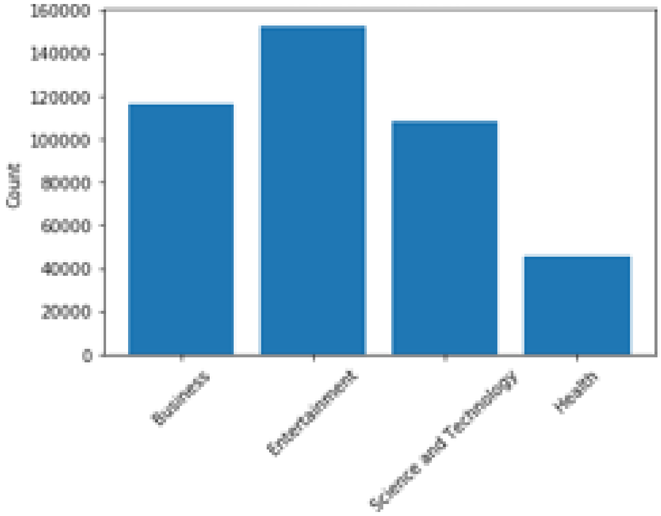

Fig. 5. News title counts for different categories before balancing

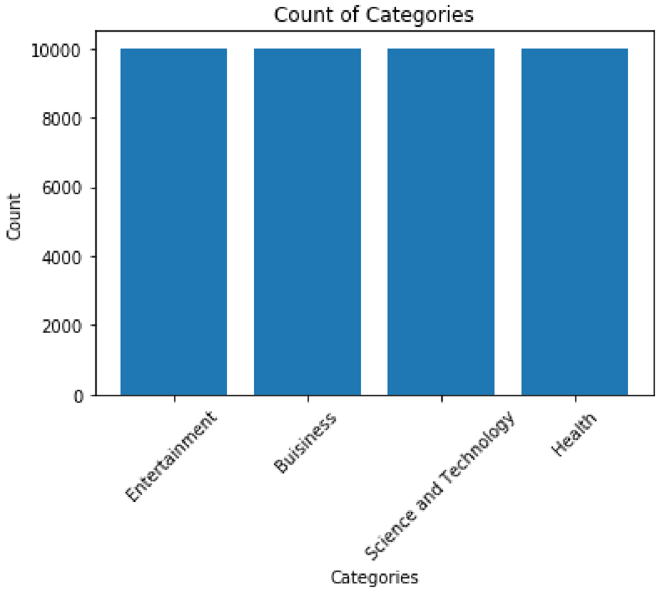

Fig. 6. News title counts for different categories after balancing

C. Tokenizing

In order to train the model better the sentences were turned into numerical sequences. Each sequence was then padded with 0 to unify the dimensions of these sequences.

D. One hot encode

The data is a categorical data with 4 categories namely business(b), entertainment(e), health(m), science and technology(t). For an efficient and better implementation the data was one hot encoded as [1,0,0,0]-b ,[0,1,0,0]-e,[0,0,1,0]-m,[0,0,0,1]-t.

V. Network Architecture

Keras framework was chosen for the implementation and thus Sequential Model, Dropout, LSTM and Dense layers were imported. Adam optimizer was chosen. With categorical crossentropy. The code runs for 2 different learning rates,for two different activation functions and compares the base LSTM with LSTM10 and LSTM11 for training and testing accuracies. The code was run for 15-20 epochs. VM instance at google cloud platform was used for the computations ans training.

VI. Results and conclusion

The results (model accuracy curve) for two different learning rates and activation function comparing base LSTM with LSTM layer 10 and 11 are depicted in Fig7. to Fig10.

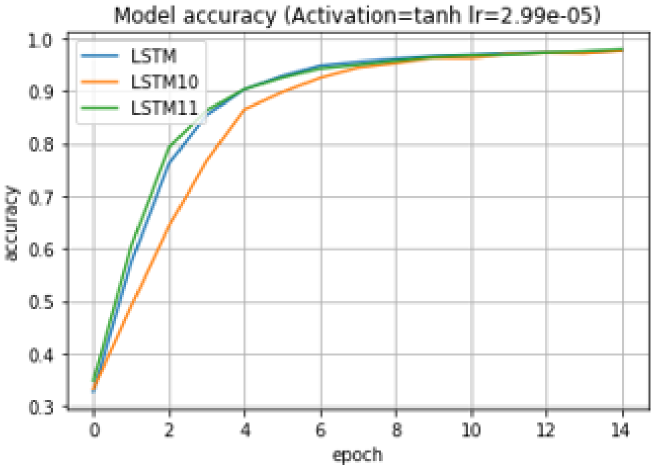

Fig. 7. Model Accuracy with tanh activation and 2.99e-05 learning rate

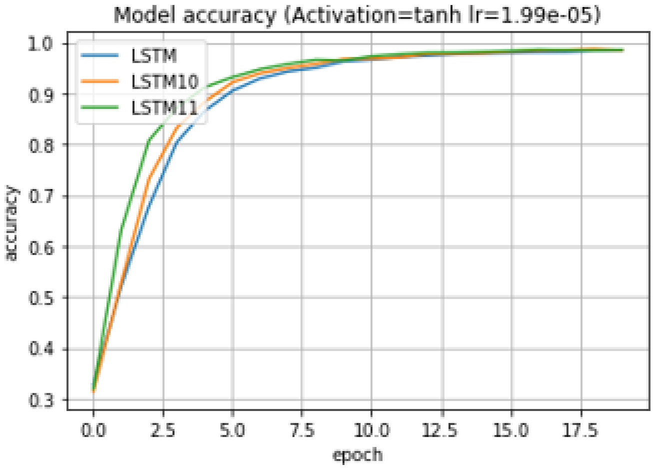

Fig. 8. Model Accuracy with tanh activation and 1.99e-05 learning rate

The results for the two different learning rates and activation functions can be summarized in the four tabels given.

VII. conclusion

The paper discussed two simplified LSTMs models (LSTM10 and LSTM11) that were defined by eliminating

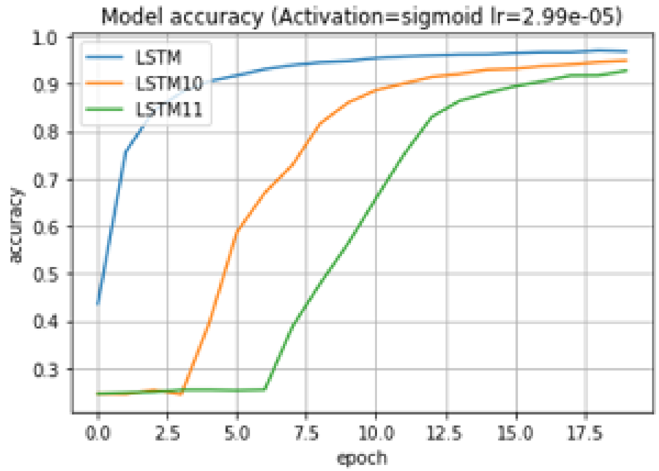

Fig. 9. Model Accuracy with sigmoid activation and 2.99e-05 learning rate

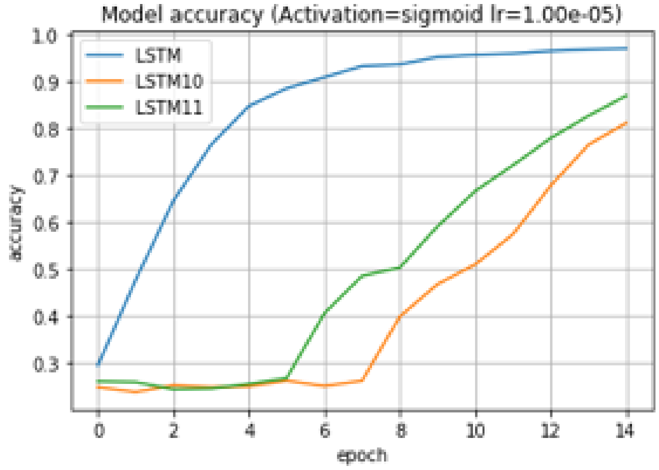

Fig. 10. Model Accuracy with sigmoid activation and 1.00e-05 learning rate

TABLE I

	ACTIVATION = tanh , LEARNING RATE=2.99e-05	
	Training Accuracy	***Validation Accuracy***
LSTM0	0.9744	0.78
LSTM10	0.9763	0.77
LSTM11	0.9793	0.78

TABLE II

	ACTIVATION = tanh , LEARNING RATE=1.00e-05	
	Training Accuracy	***Validation Accuracy***
LSTM0	0.9858	0.80
LSTM10	0.9855	0.79
LSTM11	0.9861	0.79

TABLE III

	ACTIVATION = Sigmoid , LEARNING RATE=2.99e-05	
	Training Accuracy	***Validation Accuracy***
LSTM0	0.9686	0.82
LSTM10	0.9487	0.81
LSTM11	0.9275	0.80

TABLE IV

	ACTIVATION = Sigmoid , LEARNING RATE=1.00e-05	
	Training Accuracy	***Validation Accuracy***
LSTM0	0.9686	0.77
LSTM10	0.8113	0.65
LSTM11	0.8691	0.69

input signal, bias and/or hidden units from their gate signals in the standard LSTM RNN, The results of the evaluations show that LSTM RNNs show that tanh activation has better convergence with learning rate 1.99e-05. In case of Sigmoid function base LSTM performed better and LSTM 10 and LSTM 11 and LSTM 10 outperforms LSTM11. One important observation is that the model is overfitting as the training accuracy is almost higher in all the cases when compared to the validation accuracy (approx. 20% higher).This issue can be resolved by increasing the count taken from each category. Currently only 10000 counts are taken from each category this count can go up to 45000 with data being still balanced. Also varying this data split for training and testing can help reducing the overfitting problem. The current split is 33%.

References

[1] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673–2681, Nov 1997.

[2] T. Ho, D. Can, and E. Chng, “An investigation of word embeddings with deep bidirectional lstm for sentence unit detection in automatic speech transcription,” in *2018 International Conference on Asian Language Processing (IALP)*, pp. 139–142, Nov 2018.

[3] A. M. Ertugrul and P. Karagoz, “Movie genre classification from plot summaries using bidirectional lstm,” in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pp. 248–251, Jan 2018.

[4] M. R. Alifi and S. H. Supangkat, “Information extraction of traffic condition from social media using bidirectional lstm-cnn,” in *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 637–640, Nov 2018.

[5] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *CoRR*, vol. abs/1412.3555, 2014.

[6] Y. Lu and F. M. Salem, “Simplified gating in long short-term memory (lstm) recurrent neural networks,” in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1601–1604, Aug 2017.