Deep Learning for Sentiment Analysis
Term Project Report

Chunxu Tang
ctang02@syr.edu

Zhi Xing
zxing01@syr.edu

1 Introduction

Due to the "world of mouth" phenomenon, mining the social media has become one of the most important tasks in Data Mining. Particularly, Sentiment Analysis on social media is useful for various practical purposes such as brand monitoring, stock prediction, etc. Sentiment Analysis is inherently difficult because of things like negation, sarcasm, etc. in texts, but Machine Learning techniques are able to produce accuracy above 90% for multi-class classification in regular texts such as movie reviews, arguably better than human. Unfortunately, the irregularities of social-media texts, such as misspelling, informal acronyms, emoticons, etc., make social-media-oriented Sentiment Analysis, or Text Mining in general, extremely difficult.

The buzzing Deep Learning is dominating pattern recognition in computer vision and voice recognition. As it turned out, it may be good at text classification as well. Various deep neural nets achieve state-of-art sentiment-polarity classification on Twitter data (about 87%) [16, 17, 24]. One of the advantages of Deep Learning is its ability to automatically learn features from data, and this ability leads to lots of interesting designs [9, 16, 17, 23, 24]. In this term project, we studied Recurrent Neural Network and Convolutional Neural Network and their related techniques, implemented and experimented on the networks for Twitter sentiment analysis.

2 Background

2.1 Word Embedding

Image and audio data are rich and high-dimensional, which gives their learning systems lots of things to work with. However, one of the difficulties in Natural Language Processing (NLP) is due to the sparsity of data, because normal word encodings are arbitrary to learning systems and no useful information exists between encodings of any two words. For example, in the certain encoding scheme, the word "good" may be encoded as `Id123` while the word "fantastic" is encoded as `Id456`. These two IDs don't mean anything special to any system even though the two words share similar meanings in lots of contexts. Word embeddings can resolve this issue to certain degree.

A word embedding is a parameterized function mapping words in certain language to high-dimensional vectors. It is grounded on the assumption that words that appear in the same contexts share semantic meanings. The function, which is essentially a lookup table of words to embeddings, is obtained by training a neural network with word-context data. In the high dimensional embedding space, semantically similar words are mapped to points that are nearby each other. For example, the embeddings for "dog" and "cat" are close. In fact, words for animals are in general close to each other. In addition, words' semantic differences are captured by their distances in the embedding space. For instance, the distance vector of "Beijing" to "China" is similar to that of

"Paris" to "France", and the distance vector of "woman" to "man" is similar to that of "queen" to "king". Because word embeddings are able to capture semantic relationships between words, it can provide NLP systems richer data.

2.2 Convolutional Neural Network (CNN)

Deep Learning is pushing the cutting-edge of computer vision, and one of the essential reasons is Convolutional Neural Network (CNN). The key characteristic of CNN that makes it so successful is its ability to automatically select features from inputs. The convolutional layer of a CNN acts like a sliding window over an input matrix. At each step of the sliding, normally referred to as a *stride*, the convolutional layer reduces the submatrix within the window to a single output value. This transformation is done at every stride, and the output values' relative positions are kept. Therefore, at the end, the input matrix is converted to a smaller output matrix. Since the same convolutional layer is applied at every stride, the number of parameters to learn is relatively small, which is why the network can be "deep". As an example, consider a training set of 100×100 pictures, a convolutional layer with window size 10×10 slides over each picture from left to right, top to bottom, and converts every 10×10 submatrix of pixels to a single number. After this layer, the 100×100 picture essentially becomes a smaller one. The actual resulting size depends on the *stride size*, which is the number of pixels the window slides for the next stride. In the example, if the stride size is 1, the output matrix is 91×91.

The set of parameters, once learnt, make the convolutional layer specialize at a certain aspect of the input. In a typical CNN, there can be multiple parallel *filters* in a convolutional layer, each of which specializes at a different aspect [18]. These aspects are the "features" learnt by the CNN. In the field of computer vision, one filter may specialize in detecting horizontal edges, while another may specialize in detecting vertical edges; one filter may specialize in colors, while another may specialize in contrasts. In the context of text mining, the 2-D picture becomes 2-D representation of sentence, which is normally obtained by converting a sentence to a sequence of word embeddings. Since the values in the vector can be combined to obtain different aspects of the word's meanings, the hope is that the convolutional layers can specialize so that CNN is able to automatically detect useful features from the word embeddings.

For word embeddings, we use the `word2vec` model in [21] in this term project. It comes in two flavors, the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model. These two models are algorithmically similar. The difference is that, CBOW model is trained to predict a target word from its context while Skip-Gram model is trained to predict context words from target word. As an example, consider the phrase "the cat sits on the mat". We parse the data one word at a time, so each word gets to be the target word once. If "mat" is the target word, CBOW predicts "mat" from "the cat sits on the", while Skip-Gram predicts "the", "cat", "sits", "on" or "the" from "mat". CBOW model is better for smaller datasets while Skip-Gram model is better for larger ones.

In a typical CNN, a convolutional layer is normally followed by a pooling layer, e.g. max pooling, which selects the most important the features. There can be multiple repetitions of convolution-pooling pairs and the final pooling layer is normally connected to a fully connected layer, which generates outputs for classification.

2.3 Recurrent Neural Network (RNN)

In a traditional neural network, it is assumed that every input is independent of each other. While, in reality, sometimes, the processed information is in sequence. For example, in a sentence, the

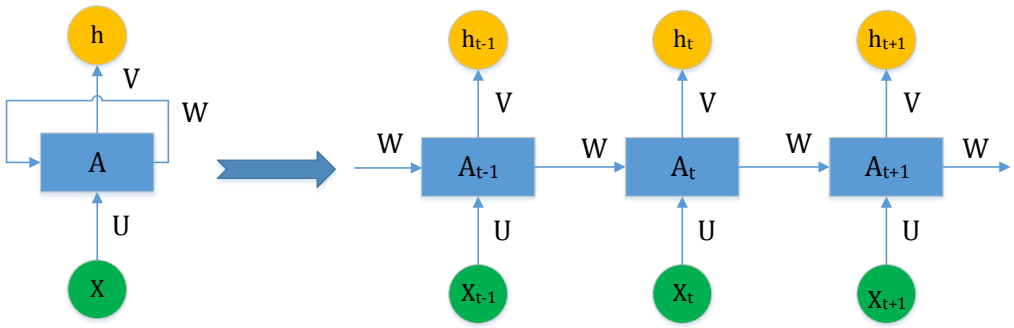

Figure 1: Architecture of Recurrent Neural Network.

words may have relationship with other words. Much as Convolutional Neural Network is especially for data with grid of values, the Recurrent Neural Network is designed to make use of sequential information. The "recurrent" refers to performing the same task for every element of a sequence, and the input of every node depends on previous computation.

The model of a classical RNN is shown in Figure 1. In the left of the graph, there is a computation node **A**, with an input **X**, an output for current node **h**. There is a weight **W** which is utilized to compute next state. In this case, the computation of current state depends on previous computation results. The right side of the figure is the unfolding in time of the computation. From the figure, we can observe that

$$\mathbf{A}_t = f(U\mathbf{X}_t + W\mathbf{A}_{t-1})$$

where $\mathbf{A}_t$ is the hidden state of time **t**, $\mathbf{X}_t$ is the input of current state, and function **f** is the transition function for every time step.

3 Classification with Paragraph Vector for Sentiment Analysis

3.1 Paragraph Vector (doc2vec)

Distributed representation of words was first raised in [**?**], and recently, Quoc Le et al. [19] proposed *Paragraph Vector*, an unsupervised algorithm that learns fixed length representation from variable-length pieces of texts, such as sentences and documents. Currently, this state-of-art technique provides promising results on text classification and sentiment analysis tasks.

Paragraph Vector consists of two representation models: distributed memory (DM) model and distributed bag of words. In the distributed memory model, every paragraph is mapped to an unique vector, and every work is also mapped to an unique vector. The paragraph vector and word vectors are averaged, summed, or concatenated together to predict next word in the context. And for the distributed bag of words (DBOW) model, it is to sample random words from the paragraph for the classification task.

3.2 Experiment

The dataset used in the experiments is from the Stanford Twitter Sentiment corpus [11], which consists of 1.6 million two-class machine-labeled tweets for training, and 498 three-class hand-labeled tweets for test. We removed all of the hashtag labels and urls, used `TweetTokenizer` to tokenize the tweets, and finally removed all of the punctuations.

Parameter	Description	Value
min_count	ignore all words with total frequency lower than this	1
window	maximum distance between the predicted word and context words used for prediction	10
size	dimensionality of the feature vectors	100
sample	threshold for configuring which higher-frequency words are randomly downsampled	1e-5
worker	number of worker threads	12

Table 1: Parameters of `doc2vec`.

Model		Accuracy
DM	dm_mean	73.0
	dm_concat	62.1
	dm_sum	73.5
DBOW		68.5

Table 2: Accuracies(%) of different models with logistic regression.

For the Paragraph Vector, we use implementation of models.doc2vec method in gensim library for the sentiment analysis task. The function supports both of distributed memory and distributed bag of words models. The parameters we select are shown in Table 1. Because of the long time of training the representation model, we randomly extract 10,000 positive tweets and 10,000 negative tweets from training dataset.

In the experiment, we compare the performance of distributed model whose word vectors are summed, averaged and concatenated, and distributed bag of words model. The evaluation result is shown in Table 2. From the table, we observe that dm_sum achieves the highest accuracy of the four approaches. So in the next experiment, which is about comparison of classification algorithms, we use dm_sum model for the training.

For the classification, we evaluate and compare logistic regression, k-nearest neighbors, and random forest. And the performance of these classifiers is demonstrated in Table 3. From the table, we summarize that logistic regression obtains the best classification performance of the classifiers, and the accuracy is 72.7%.

4 Convolutional Neural Network for Twitter sentiment

4.1 Network structure

The structure of the CNN used in our term project is shown in Figure 2, where the parameters are chosen soloy for visualization. The structure is a simplification of the model used in [17]. Although this model has a minimalist design, it has most of the typical layers of a text-mining CNN: word embedding layer, convolutional layer, max-pooling layer, and fully-connected layer.

	LR	SVM	RF	KNN	NB
Accuracy	**73.5**	72.9	58.2	59.6	66.6

Table 3: Accuracies(%) of different classifiers. LR: Logistic Regression. SVM: Support Vector Machine. RF: Random Forest. KNN: K-Nearest Neighbors. NB: Naive Bayes.

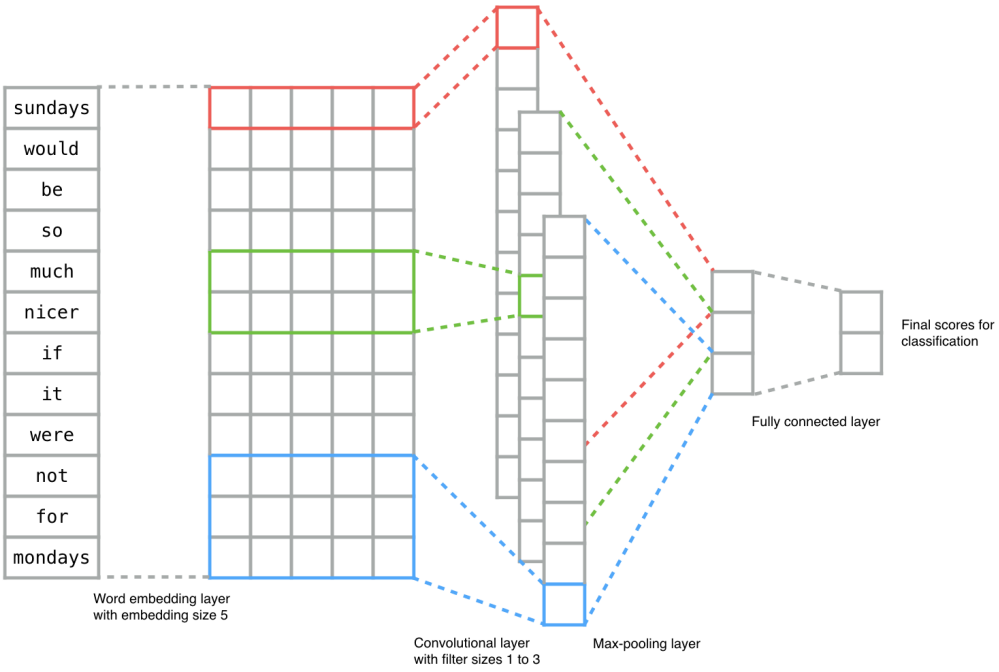

Figure 2: Structure of CNN

The first layer is the word embeddings, which is encoded as a weight matrix that is used as a lookup table. Each row of the weight matrix is a vector of size k represents a unique word in the vocabulary. Given a sentence of length n, padded if necessary, each word is replaced by its corresponding vector and the sentence is converted to an output matrix that is a concatenation of all the embeddings. This output matrix is represented as

$$\mathbf{x}_{1:n} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \ldots \oplus \mathbf{x}_n$$

where $\mathbf{x}_i \in \mathbb{R}^k$ is the embedding of the i-th word in the sentence, and $\mathbf{x}_{i:ij} \in \mathbb{R}^{kj}$ is used to denote the sub-matrix from the i-th word to the j-th word.

The second layer is the convolutional layer. Given a window size h, this layer is encoded by a *filter* $\mathbf{w} \in \mathbb{R}^{hk}$. When this filter is applied to the h-gram $\mathbf{x}_{i:i+h-1}$ of the input sentence, a feature c_i is generated by

$$c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b)$$

where $b \in \mathbb{R}$ is bias and f is the activation function such as rectifier. The filter slides over all the possible h-grams of the input sentence to produce a feature map

$$\mathbf{c} = (c_1, c_2, \ldots, c_{n-h+1})$$

for $\mathbf{c} \in \mathbb{R}^{n-h+1}$.

The next layer is the max-pooling layer. It is applied to the feature map produced by the convolutional layer to produce a single feature $\hat{c} = \max(\mathbf{c})$. Taking the maximum essentially picks the most important feature in the feature map, which is an effective way to deal with variable sentence length, because the special word for padding has 0s in all the dimensions in its word embedding.

The last layer is a fully-connected layer. The $\hat{c}$ is the selected feature by a *single* filter, there're a number of filters with different window sizes. All the selected features from these filters are the

input for this fully-connected layer, which uses softmax function to calculate the score for each class.

4.2 Regularization

The purpose of regularization is to prevent overfitting or co-adaptation. Unlike [17], only dropout is used to prevent co-adaptation, because according to [25] the L2-norm constraint used in [17] generally has little effect on the end result, and we want to keep the network as simple as possible. The dropout is applied to the fully-connected layer of the CNN. It works by randomly setting a proportion p of hidden units to 0 during learning. During testing, the dropout is disabled and the learnt weights are scaled down by p.

4.3 Experiment

Hyperparameter	**Value**
sentence length	200
word embedding size	200
filter window size	1, 2
number of filters per window size	128
dropout probability	0.5

Table 4: Hyperparameters of CNN model

The dataset used in the experiments is the Stanford Twitter Sentiment corpus [11], which consists of 1.6 million two-class machine-labeled tweets for training, and 498 three-class hand-labeled tweets for test. We compose a smaller training set of 25,000 positive and 25,000 negative tweets randomly sampled from the original, a validation set of 2,500 positive and 2,500 negative tweets randomly sampled from the *rest* of the original, and a smaller test set consists of only the 359 positive or negative tweets, excluding the neutral ones, from the original. The validation set is used for setting some of the training parameters such as the number of training epochs.

The data is preprocessed to remove punctuations and other symbols like "#" and "@", and to separate word contractions, e.g. "don't" to "do not". Each sentence is then tokenized by the `TweetTokenizer` provided in Python NLTK [6] library and padded to 200 tokens as necessary.

The CNN is implemented in TensorFlow, Google's deep learning library [1]. The network structure is defined in Python, but the backend is implemented in C++, so the training and testing procedures run as C++ programs. The hyperparameters of the model are listed in Table 4.

The skip-gram `word2vec` model proposed in [21] is used for the word embedding layer. The size of the embeddings is 200. The embedding model is pretrained using a subset of the Google News data used in [21] that consists of 17 million words, with a vocabulary of 71,291 words. After plugged into the CNN, the parameters of the `word2vec` model are set untrainable so it becomes a static lookup table. There're two reasons for fixing the parameters:

1. To reduce the number of parameters need to be learnt.

2. Tweets contain lots of noise, making the layer trainable exposes it to the noises, which may be counterproductive. [1]

[1] We did test the model with the embedding layer set trainable. The performance is slightly worse.

Because of this layer, the vocabulary of the CNN is determined by the `word2vec` model, saved as a word-to-index dictionary. During data preprocessing, each word is converted to an index according to the dictionary.

For the convolutional layer, there can be a number of a filters with different window sizes. In order to keep our model simple and small, only two windows sizes, 1 and 2, are used, and each window size has 128 filters.

The model is trained with data batches of 128 tweets for 10 epochs. On a computer with 2.8GHz quad-core CPU and 16GB of memory, given pretrained `word2vec` model, the CNN model can be trained and tested well under half an hour. We run the train-test iteration 20 times, and obtained an average accuracy of 79.60%. Although this is not a very impressive performance, since our model has a very simple design and it is not optimized at all, it shows great potential of CNN for text mining in social media.

We also tried tweaking the hyperparameters in different ways, but didn't get significant improvement. For example, with "filter window size" set to "1, 2, 3", and the number of training epochs changed to 30, the average accuracy of 20 train-test iterations becomes 78.65%. The number of epochs is increased because the convergence time is longer due to more parameters to learn. The training time in this case is over 4 hours.

5 Recurrent Neural Network for Sentiment Analysis

5.1 Network Architecture

Though RNN is adapted to make use of sequential data, it is usually difficult to train the model. RNN maintains a vector of activations for each time step, which makes a RNN extremely deep [15]. This also leads to the problem that it is difficult to learn long term dependencies with traditional RNN [4]. A recent summary by Pascanu et al. [22] concluded the issues as the vanishing and the exploding gradient problems. There has been some work trying to solve the difficulty of training a RNN model. Hochreiter et al. [14] raised the Long Short-Term Memory (LSTM) architecture, which addresses the problem by re-parametrizing the RNN. Another model, Gated Recurrent Unit (GRU) was first presented by Cho et al. [7], is to make each unit to capture dependencies of different time scales. Chung et al. [8] later evaluated LSTM and GRU on sequence modeling and found that GRU is comparable to LSTM. In our work, regarding RNN, we concentrate on LSTM.

The architecture of LSTM is shown in Figure 3. The first layer for LSTM is the "forget gate layer". In this layer, the model checks h_{t-1} and x_t, and outputs a value to represent how much information it needs to keep from previous states.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

The second step consists the "input gate layer" and a tanh layer. The "input gate layer" is a sigmoid layer, which is used to decide what values the model needs to update. And the tanh layer creates a new vector, $\tilde{C}_t$, which could be added to current unit.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_C)$$

And next, the model will obtain C_t from multiplying old state by f_i, multiplying $\tilde{C}_t$ by i_t and summing them up.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

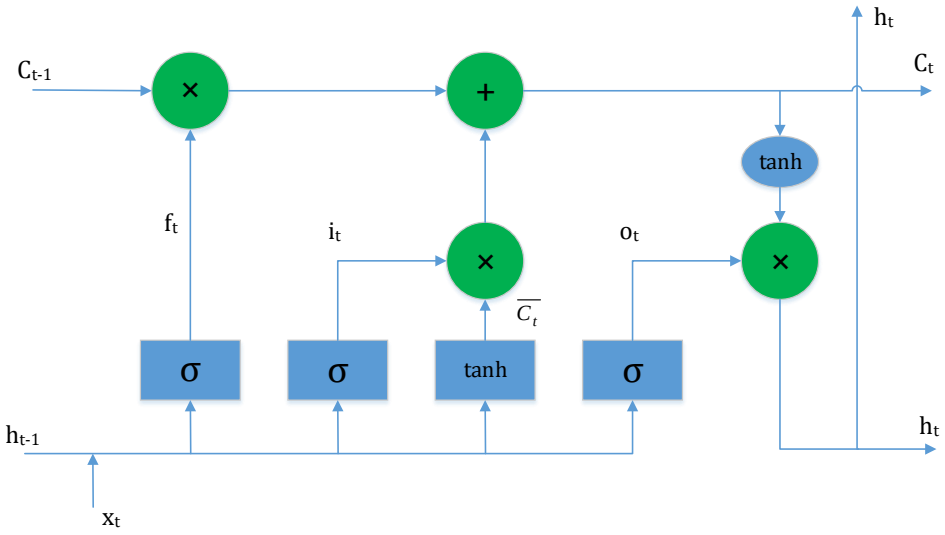

Figure 3: Architecture of Long Short-Term Memory.

Finally, the model needs to decide the output. The output gate layer implements operations on C_t and o_t, and output h_t for the next state.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$
$$h_t = o_t * tanh(C_t)$$

The deep learning model we use in the report is a variance of classical LSTM. And the structure of the LSTM variance is shown in Figure 4. In our model, the output of current state does not depend on current memory cell state C_t, but just on input x_t and h_{t-1}. So, the equation of o_t is updated to

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

Additionally, the model also consists of a mean pooling layer and a logistic regression layer. For every LSTM cell, it outputs information h_i. And by averaging all of the output sequence, mean pooling outputs h and it is fed to the logistic regression layer. Then the logistic regression layer will train the model according to class labels and associated sequences.

5.2 Experiment

For the LSTM model for sentiment analysis, we design and evaluate the model on two different kinds of sentiment data sets. The first one is the Stanford Twitter corpus, which has been described in previous section. And the second data set is the Large Movie Review Dataset [20], which consists of 25,000 highly polar movie reviews for training, and 25,000 reviews for testing. In the 25,000 training reviews, there are 12,500 positive reviews and 12,500 negative reviews.

During the pre-processing step, we utilize the count-based method to represent the words in the data sets. A count-based approach takes advantage of the assumption that words which have similar counts in a text context, share similar semantic meaning. This approach is opposite to context-predicting semantic vectors such as `word2vec` which is used in our experiment of CNN. The difference between the two word representation models were discussed in citebaroni2014.

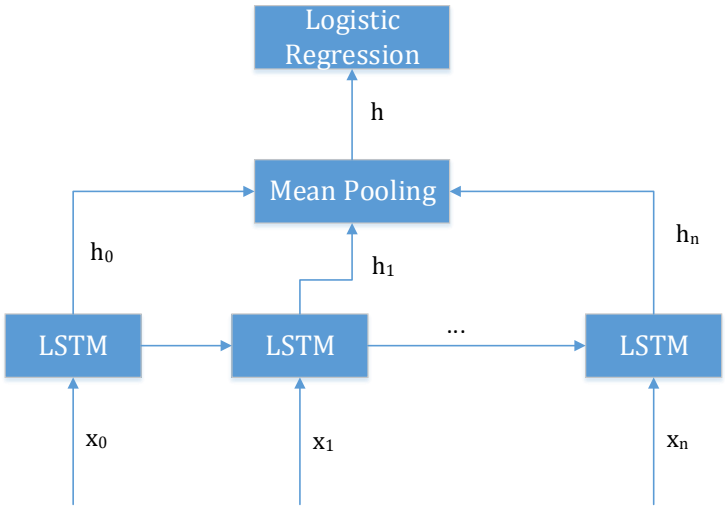

Figure 4: LSTM model used in our report.

The LSTM model is implemented in Theano [3, 5], which is a Python library for deep learning. It allows users to define, optimize and evaluate mathematical expression efficiently and conveniently.

We run the LSTM model on each of the two data sets three times, and calculate the mean value of the accuracies. While, we observe a dramatic difference between the two data sets. The accuracy of Stanford Twitter corpus is as low as 54.6%, but the accuracy of IMDB dataset is as high as 81.7%. To explain the large gap between the results, we think there may be following reasons:

1. The paramters of the LSTM model on Stanford Twitter corpus needs to tuned. Even though we have adjusted the parameters for every specific dataset, we did not find the best settings for Twitter sentiment analysis.

2. LSTM is good at capturing long time dependency information in a corpus. While, for tweets, actually, they are usually very short, and there is little dependency information existed. Our LSTM model may not fit well in this case.

6 Conclusion

References

[1] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247, 2014.

[3] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*, 2012.

[4] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on*, 5(2):157–166, 1994.

[5] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a cpu and gpu math expression compiler. In *Proceedings of the Python for scientific computing conference (SciPy)*, volume 4, page 3. Austin, TX, 2010.

[6] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[7] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[9] Cícero Nogueira dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*, pages 69–78, 2014.

[10] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[11] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12, 2009.

[12] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*, volume 385 of *Studies in Computational Intelligence*. Springer, 2012.

[13] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2342–2350, 2015.

[16] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[17] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[20] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[22] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *arXiv preprint arXiv:1211.5063*, 2012.

[23] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[24] Xin Wang, Yuanchao Liu, Chengjie Sun, Baoxun Wang, and Xiaolong Wang. Predicting polarities of tweets by composing word embeddings with long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1343–1353, 2015.

[25] Ye Zhang and Byron Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.