Journal of Computers Vol. 29 No. 5, 2018, pp. 250-259
doi:10.3966/199115992018102905020

Sentence Classification Using Novel NIN

Yan-Ping Fu[1*], Yun Liu[1*], Zhen-Jiang Zhang[2]

[1] Department of Electronic and Information Engineering, Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing, China
{17111012, liuyun}@bjtu.edu.cn

[2] Department of Software Engineering, Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing, China
zhenjiangzhang@bjtu.edu.cn

Received 7 January 2018; Revised 7 July 2018; Accepted 7 August 2018

Abstract. Sentence classification is basic problem of natural language processing, thus the ability of accurately modelling sentences is pivotal to classification performance. We describe a novel NIN (Network In Network) to train on top of word embedding for sentence-level classification tasks. The noval NIN consists a conventional convolutional layer, a micro neural networks layer named perceptron layer, global average pooling and a softmax classification layer. We modify conventional NIN with decreasing the layer of perceptron and applying an effective activation function to adopt for the modelling of sentences classification. We demonstrated the excellent classification performances with novel NIN on two datasets: small scale binary sentiment prediction and THUC news.

Keywords: data mining, deep learning, NIN, sentence classification, text classification

1 Introduction

As one of the natural language processing issues, text classification has been widely researched with different methods [1-2]. With the popularity of short text in the social networks, sentence classification have become research hotspots in text categorization. It has been solved by many deep learning models and has achieved remarkable results in recent years [3-5].

Text classification mainly relied on manual annotation features and statistical classification methods at early stages [6-7]. Recently, with the development of technology, deep learning method can obtain text features through self-training without manual annotation and complete classification tasks through supervised learning. To structure natural language word vector representations, deep learning models can train compatible network parameter to structure language models. Then, it applies word representation that have been trained to implement text classification [8-9]. As the composition of sentence feature, word embedding is a vector representation with a 1-V (V is the vocabulary size) lower dimensional space map, which encode contextual semantic information in their dimensions [10-11]. Using deep learning complex network, it is more significant to train large language data set than the simple models.

Deep learning models have their advantages in aspect of natural language processing, which mainly relies on their perfect self-learning ability to construct language model. Some deep learning models commonly are used for text classification such as recurrent neural network (RNN) [12], convolutional neural network (CNN) [13], gated recurrent unit (GRU) [14], Long Short-Term Memory (LSTM) [15] and their variants [16-18]. At first, these models generally obtain word representation of text through adjusting the parameters of word embedding. Then, the classification model is conducted by training the parameters of the network.

CNN is a central class of deep learning models based on convolving filters [19], originally invented for computer vision [20], have been generally applied to natural language processing such as question chunk

* Corresponding Author

parsing [21], sentence modeling [2], and other basic NLP assignments [22]. For sentence classification, the convolution layers apply one-dimensional filters convolving each row of feature vector in the sentence matrix, and the convolution filters commonly have the same dimension with each word embedding vector to maintain the independence of position features in the sentence. A convolutional layer followed by a max-polling layer and a non-linearity function to constitute sentence feature map, which are combined with a full-connected layer for softmax classifier.

The core convolution filter in CNN is a generalized linear model (GLM) for the patch data, and this construction leads the ability of information extraction is low with GLM. Lin [23] proposed a novel deep network structure named network in network (NIN) to enhance the abstraction ability of the CNN model and discriminability for local patched. To instead of simple GLM in CNN, NIN builds networks with more complex structures to abstract input data in the receptive field. Its structure is a convolution filter followed a multilayer perceptron, and for the sake of less overfitting, NIN utilizes global average pooling over feature map to replace max-pooling and fully connected layers in the classification stage of CNN.

In the task of sentence classification, we define the novel NIN model that adopted for the semantic modelling of sentences. Considering of word semantic complexity [24-26], we reconstruct the NIN model including its multilayer perceptron, the distribution of activation function and regularization. In the research of image identification, the features of image are multidimensional and there is not much correlation between adjacent features. Different from image identification, the feature of text is one-dimensional and the semantic relations of adjacent words need to be considered. Thus, in order to perform the sentence classification, we propose novel NIN structure composed of the multichannel one-dimension convolution layer, followed one-layer perceptron, activation function, average pooling layer and softmax classification layer.

The network handles one-dimensional convolutional layers having same dimension with the word embedding of the sentence feature and apply a more efficient activation function with Elu function [27] to scan the input rather than conventional Relu function [28]. In order to reduce the computational complexity without loss of performance. Followed one-layer perceptron network abstract sentence information instead of stacking multilayer perceptron in classical NIN. With reducing the probability of overfitting, we add a dropout layer behind one-layer perceptron. Then via a global average pooling layer, the spatial average of the feature maps are output and fed into the softmax classification layer. Compared with conventional NIN, we only apply one-layer perceptron for reducing network parameters without any performance loss and combine a more effective activation function named Elu with a dropout layer that is deficient in the NIN to achieve excellent performance of sentence classification.

We experiment with the modified NIN in two data set. The experiments involve binary predicting the sentiment of movie reviews with English datasets [29] and THUCNews data with Chinese datasets [30]. In the experiment, we adopt the word-embedding matric as the feature of the sentence to operate classification training with the proposed network and make some comparisons with other models to demonstrate the superiority of the classification performance.

2 Related Works

2.1 CNN Structure

The convolutional neural networks for sentence classification was firstly proposed in 2014 by Kim [31], whose research was based on the contribution of Razavian et al. [32] about CNN of image classification. As the issues of text processing, the semantic relationship of a statement is important. The greatest advantage of the CNN model is that it can capture the information pertinence of local region, which is pivotal for the task of sentence classification.

CNN model for sentence classification mainly include four layers such as input layer, convolution layer, max pooling layer, fully connected and softmax output layer. For the task of a sentence classification, Yoom showed the architecture of CNN model for a sentence as Fig. 1. The input layer is the k-dimensional word vector corresponding to the i-th word in sentence. In the convolution layer, the filters of different channels convolute the input $n \times k$ representation of sentence to obtain the convolution features of different channel. Next, to capture the most important feature for each feature map, the max-pooling layer extract the highest value of feature in each channel to complete feature extraction. Finally, the fully connected and softmax output layer pass the feature to softmax classifier whose output is the

probability distribution over labels. Thus, the whole model is to conduct language model for word feature and to handle classification model for sentence category.

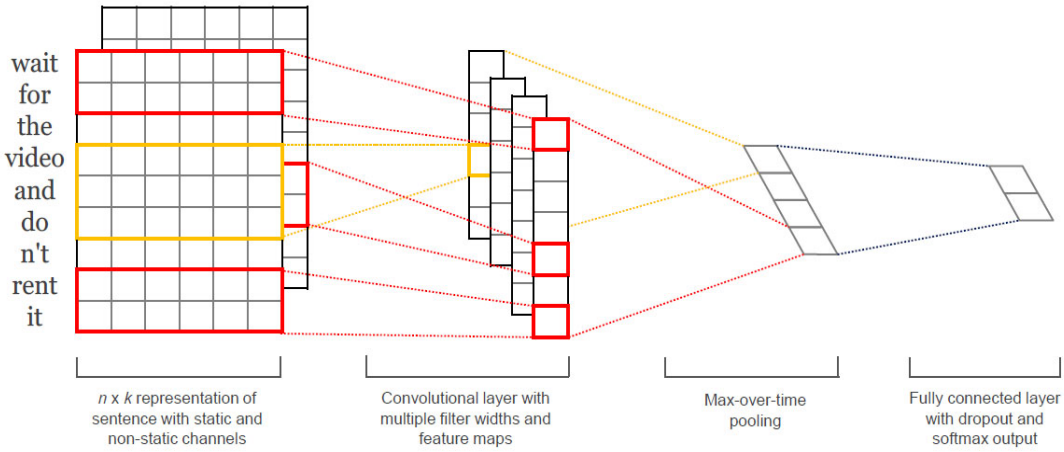

Fig. 1. Model architecture with two channels for the classification of a sentence

This model is a simple CNN structure that uses one layer of convolution to extract features of sentence and complete the classification without deep network. This simple structure can establish the language model perfectly and classify the feature through softmax classifier. The experiments of literature [31] added to the well-established evidence that CNN network is an excellent classification for natural language.

2.2 Conventional NIN Structure

In 2014, Lin proposed a NIN model applied on the image processing, and reconstruct the conventional CNN model and achieve expected performance on image identification. As the convolution filter in CNN is a generalized linear model, its level of abstraction is low. To enhance the abstraction ability of the local model, NIN model was conducted by a nonlinear function approximator.

The structure of NIN is show in Fig. 2. Some stacking of a linear convolutional layer and mlpconv layers map the local receptive field to an output feature vector and a average pooling layer is followed to generate the final classification feature. Compared with CNN, this model adds multilayer perceptron consisting of multiple fully connected layers with nonlinear activation functions. The perceptron is fully connected network that can represent the nonlinear relationship of networks. It converts linear extraction to nonlinear extraction so that the ability of feature extraction is improved. At the same time, instead of max-pooling layer and fully connected layer of classification in CNN, the model of NIN applied a global average pooling layer output the spatial aver-age of the feature maps to classification layer. This layer directly output the spatial average of the feature maps from the last mlpconv layer as the confidence of categories. This structure is a regularizer to prevent overfit-ting of CNN.

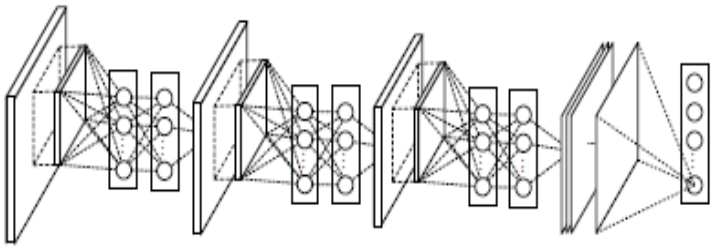

Fig. 2. The overall structure of Network In Network. This figure is proposed by MinLin [23]

In the experiment of image processing, the NIN is composed of the overall deep network with three stacked layers, the mlpconv layers are followed by spatial max pooling layer and dropout is applied on the outputs of all but the last mlpconv layers. For all dataset of image classification in experiments, the performance of NIN model is state-of-the-art and the feature maps of NIN were confidence maps of the

categories. These experiences illustrated the superiority of NIN model.

3 Novel NIN Model

In this paper, we propose the novel NIN model for sentence classification. The initial NIN model was proposed for image identification, however, the feature of word in sentences is totally different from the feature of image. The feature of image is multidimensional in aspect of color and pixels are locally related, but the feature of word in sentences is one-dimensional and the relation between adjacent words is closely related. Thus, for the task of sentence classification, the conventional NIN model is inapplicable and needed to be transformed suitably.

In this section, we firstly introduce the structure of novel NIN model and give an example to illustrate the process in part 3.1. Secondly, we present one-layer perceptron of the novel NIN model in part 3.2. Thirdly, we propose to use a new activation function for whole network in part 3.3. Lastly, the dropout regulation is introduced to entrance the robustness of network.

3.1 Structure

The model architecture applied to sentence classification, show in Fig. 3, is a slight variant of NIN architecture proposed by Lin et al. It is different from scanning the local pixel feature of the image identification that sentence classification is scanning some rows as basic feature unit, since sentence is presented by a matric with a word embedding vector of every row. The feature of sentence is convoluted by the multichannel filters on the convolution layer, and one-layer perceptron make nonlinear mapping to the characteristics of convolution. A dropout layer is followed to reduce the overfitting and a global average pooling layer is applied to obtain the classification feature for the final softmax classify.

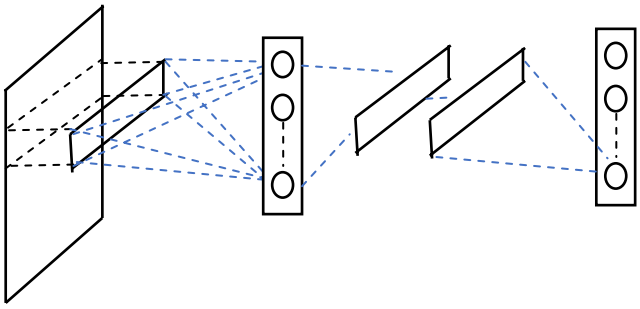

Fig. 3. The overall structure of NIN for sentence classification

In our model, one-layer perceptron is applied for sentence classification because it can fit any function. It maps word features in different channel to the joint characteristics through one-layer perceptron. This structure is simple and effective for feature extraction of sentence without two-layer perceptron of NIN. To reduce the computational complexity, the whole network is designed simply. Therefore, in order to guarantee the robustness of the network, the dropout layer is added to the network.

The specific example of the classification process is as Fig. 4. As in part (a), a sentence can be represented $n \times k$ matric, in which n is the length of a sentence (padded where necessary) and k is dimension of word embedding vector. Part (b) contains 6 convolution filters defined as same dimension with word vector for holding independence of each word. Column vector value in part (c) is results of the convolution operation between (a) and (b), containing same length with the sentence and the number of eigenvalues depends on the total number of convolution filter. A nonlinear activation function named Elu followed convolution layer to abstractly scan the input data.

For the sake of reducing structural complexity, we apply a layer perceptron in (d) to generate feature map (e), which is equivalent to 3 1×1 filters in (d) operate convolution to 6 vector value in (c). The conventional deep NIN used two 1×1 convolution layers as multilayer perceptron while we decrease one layer for a light structure with enough ability of representing information between words simultaneously. By the same token, the number of the feature map in (e) is equal to the number of 1×1 filter, which means (c) layer decides the final classification numbers. And a Relu activation function is applied for the

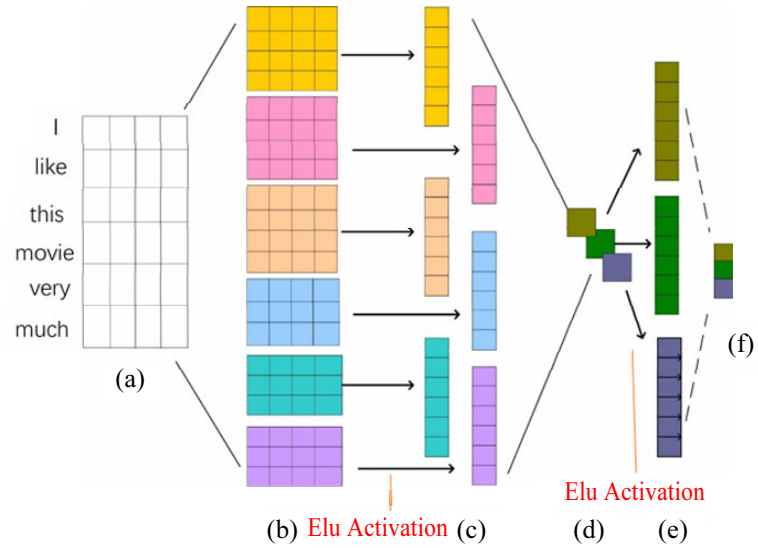

Fig. 4. The structure of the modified NIN for sentence classification

feature map. Then we directly output the spatial average of the feature maps from the one perceptron layer as the confidence of categories via a global average pooling. This layer enforces correspondence between feature maps and categories, and in order to reduce the tendency of overfitting, a dropout regularization is added for overall structure.

3.2 One-layer Perceptron

After the convolution of the features, we build a micro neural network with a layer perceptron, which is a potent function approximator. In conventional NIN, multilayer perceptron is applied for the local patched of image. Instead, we only use one layer because one-layer perceptron can approximate any function with less operational complexity. Compared with image identify, the word embedding feature of sentence classification is more concise, our algorithm can meet the needs of feature extraction.

One-layer perceptron is fully connected network for the features of different channel convolution. Because the convolution layer is a generalized liner model, the level of feature abstraction is low. The perceptron network can improve the ability of abstraction with its property of fitting any function. The structure of one-layer perceptron is showed as Fig. 5. We represent one-dimensional features as neuron. And sentence feature from 6 channels is connected by one-layer perceptron with three channels to obtain final feature.

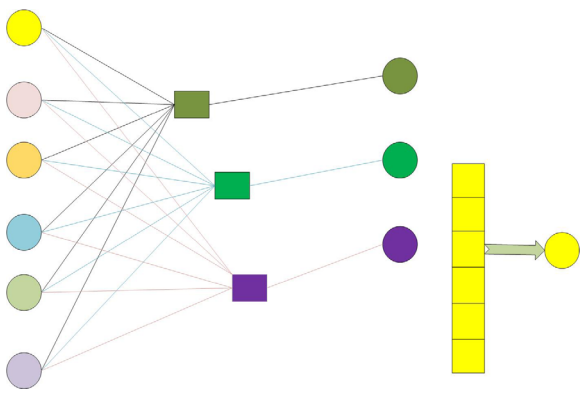

Fig. 5. The structure of one-layer perceptron

3.3 Activation Function

In conventional NIN, activation function is Relu function which is used in the most algorithm with its fast convergence speed. However, Relu function has the disadvantages of dead neuron with the value equaling zero when x is less than 0. Instead, we apply the Elu function in the sentence classification,

since it is more robust for input change or noise in the negative half shaft. Thus the Elu activation function can obtain more stable and excellent performance for the proposed model.

The formula are follows as (1) and (2).

$$f(x)=\begin{cases} x, if\ x \geq 0 \\ 0, if\ x < 0 \end{cases} \tag{1}$$

$$f(x)=\begin{cases} x, if\ x \geq 0 \\ \alpha(e^{x}-1), if\ x < 0 \end{cases} \tag{2}$$

3.4 Dropout Regulation

Dropout is a regularization method of neural network model proposed by Srivastava et al. [33]. Dropout randomly ignores some neurons during training. In other words, their contribution to downstream neurons is temporarily eliminated in the process of the forward propagation, and the neuron does not have any weight updates when it is Back Propagation.

With the continuous studying of the neural network model, the weight of the neuron matches the context of the entire network. The weights of neurons are tuned for certain characteristics, which have some specializations. The surrounding neurons would depend on this specialization, and if they were too specialized, the model would become brittle because of the training overfitting. Applying the dropout method, the network model is less sensitive to the specific weight of neurons. This increases the generalization ability of the model, and it is not easy to overfit the training data.

4 Experiment

4.1 Overview

We test modified NIN model in two different datasets, including binary sentiment prediction in movie reviews in the Stanford Sentiment Treebank with English data set and THUCNews collected by Tsinghua University with Chinese data set. In the experiments, we define each word as 64 dimensions word embedding vector in a sentence and apply the padding to maintain same length of the sentence for one dataset. The top of networks is convolution between feature matric of a sentence and 128 filters to generate feature maps following a Elu activation function, the second layer is one layer perceptron that the number of 1×1 filters keep the same with final categories. After a Elu activation function and a dropout regularization, the feature maps turn into the feature vectors of softmax classifier that predict the probability distribution over classes given the input sentence.

The network is trained through minimizing the cross-entropy of true output and predicted distributions [34]. The adaptive moment estimation (Adam) optimizer [35] is based on gradient back-propagation with mini-batches. In the processing of the training, the dropout parameter is set as 0.5 to obtain less overfitting. To exploit the faster speed of the operations, we experiment the network on a GPU and tensorflow implementation processes.

4.2 Movie Reviews

Movie reviews dataset is sentiment classification involving positive or negative reviews detection with one sentence per review [36]. In this experiment, we use 4031 sentences to train, 1000 sentences to test and 300 sentences to validate for each category. The size of vocabulary in the dataset is 18767.

In Table 1, Yoon Kim's convolutional neural network model initialized randomly (CNN-rand) obtain the accuracy of 76.1%, the result of Socher's recursive Autoencoders with pre-trained word vectors (RAE) is 77.7%, the combinatorial category autoencoders with combinatorial category grammar operator proposed by Hermann and Blunsom (CCAE) get the accuracy of 77.8% and Yang and Cardie's conditional random fields with posterior regularixation (Tree-CRF) achieve the result of 77.3%. From Table 1, we can get the conclusion, compared with some conventional models, our modified NIN model with dropout obtain more excellent result of sentence classification with 78.1% accuracy while without

dropout our model only more efficient than CNN-rand and Tree-CRF model with the result of 77.6%. the experiment results not merely demonstrate the superiority of our modified NIN model but also represent the importance of dropout to classification performance.

Table 1. Accuracy of Movie Reviews classification with other models

Model	Accuracy
CNN-rand [31]	76.1%
RAE (Socher et al., 2011) [37]	77.7%
CCAE (Hermann and Blunsom, 2013) [38]	77.8%
Tree-CRF (Nakagawa et al., 2010) [39]	77.3%
Modified NIN (without dropout)	77.6%
Modified NIN (with dropout)	78.1%

4.3 THUCNews

We choose the Chinese dataset named THUCNew as the basic data which is collected by Tsing Hua university. THUCNews is based on the filtered Sina News historical data including 740,000 news documents and having 14 classes to choose. We choosed 10 classes of THUCNews to operate the experiment including lottery, house property, sports, recreation, home furnishing, education, science and technology, fashion, current politics and game. 5000 news are chosen from each category for training, 1000 news for testing and 500 news for validating.

In this task, we validated the modified NIN model with dropout or without dropout respectively compared with Hang Zhuang's convention network model using pinyin format and strokes format [40]. From Fig. 6, we observe the modified NIN model significantly outperforms the Hang Zhuang's two models and our best result is more 6.5% than the convention network model using strokes format. The modified NIN with dropout preforms better than without dropout due to the regularization of the dropout.

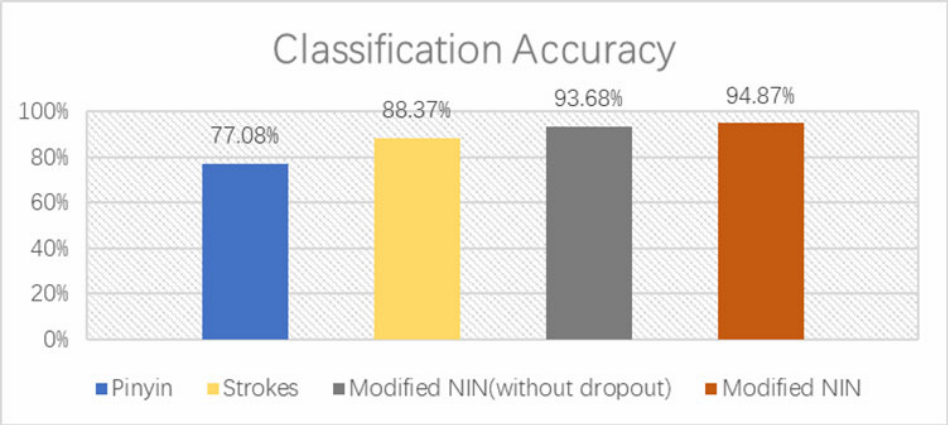

Fig. 6. Sentence classify accuracy on the THUCNews dataset

4.4 Activation Function Experiment

As already mentioned, our modified NIN structure applied the activation function to provide the nonlinear modeling ability. In the experiments, we verify different activation functions including Relu function and Elu function. We present average results from 50 trials in Table 2. For Movie reviews dataset, it shows the accuracy of the model using Elu function is more 1.75% than using Relu function. For THCUNews dataset, the result is more 1.30%. This result proves Elu function has more superior performance than Relu function due to its left soft saturability.

Table 2. Accuracy of Modified NIN model on different activation function

Data set	Relu function	Elu function
Movie Reviews	76.26%	78.01%
THCUNews	93.47%	94.87%

4 Conclusion

We proposed a novel NIN model for sentence classification tasks in this paper. This new structure is a variant of the NIN proposed by Min to process the image classification. Our model consists of a convolution layer, a perceptron layer and a global average pooling layer involving the efficient activation function and dropout regularization. With one-layer perceptron, the network structure is simple and the ability of nonlinear feature extraction is perfect. The Relu activation function further enhance ability of nonlinear. Appling dropout regularization, the simplified structure is more robust for big corpus data.

We demonstrated the state-of-the art performance on movie reviews dataset and THUCNews dataset, at the meantime, we show a small increase of classification accuracy, which prove effectiveness of the Elu activation function and the network stability of dropout regularization. Through constructing the novel NIN model for sentence classification, we conclude the feature maps of sentence were confidence maps of the categories, and this motivates the researches of novel NIN model in other fields.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (No. 2017YFC0840200), the Fundamental Research Funds for the Central Universities (No. 2017JBZ107) and Academic Discipline, Post-Graduate Education Project of the Beijing Municipal Commission of Education.

References

[1] T. Joachims, Transductive inference for text classification using support vector machines, in: Proc. the 16th International Conference on Machine Learning, 1999.

[2] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, R. Socher, Ask me anything: dynamic memory networks for natural language processing, in: Proc. International Conference on Machine Learning, 2016.

[3] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: Proc. the 25th International Conference on Machine Learning, 2008.

[4] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, in: Proc. Advances in Neural Information Processing Systems 28, 2015.

[5] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: Proc. the 52nd Annual Meeting of the Association for Computational Linguistics, 2014.

[6] A. McCallum, K. Nigam, A comparison of event models for naive bayes text classification, in: Proc. AAAI-98 Workshop on Learning for Text Categorization, 1998.

[7] K. Nigam, A.K. McCallum, S. Thrun, T. Mitchell, Text classification from labeled and unlabeled documents using EM, Machine learning 39(2-3)(2000) 103-134.

[8] Y. Bengio, R. Ducharme, P. Vincent, A neural probabilistic language model, Journal of Machine Learning Research 3(2003) 1137-1155.

[9] T. Mikolov, K. Chen, G.S. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proc. International Conference on Learning Representations (ICLR), 2013.

[10] O. Levy, Y. Goldberg, Neural word embeddings as implicit matrix factorization, in: Proc. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, 2014.

[11] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. <https://arxiv.org/

abs/1301.3781>, 2013.

[12] P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning. <https://arxiv.org/abs/1605.05101>, 2016.

[13] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, in: Proc. Advances in neural Information Processing Systems, 2015.

[14] Z. Wu, S. King, Investigating gated recurrent neural networks for speech synthesis. <https://arxiv.org/abs/1601.02539>, 2016.

[15] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, B. Xu, Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. <https://arxiv.org/abs/1611.06639>, 2016.

[16] C. Zhou, C. Sun, Z. Liu, F. Lau, A C-LSTM neural network for text classification. <https://arxiv.org/abs/1511.08630>, 2015.

[17] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, H. Hao, Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification, Neurocomputing 174(2016) 806-814.

[18] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proc. the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[19] S. Lawrence, C. LeCun, C. Nohl, A.D. Back, Face recognition: a convolutional neural network approach, IEEE Transactions on Neural Network 8(1)(1997) 98-113.

[20] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, in: Proc. Advances in Neural Information Processing Systems 25, 2012.

[21] W. Yih, X. He, C. Meek, Semantic parsing for single-relation question answering, in: Proc. ACL 2014, 2014.

[22] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, in: Proc. Advances in Neural Information Processing Systems 27, 2014.

[23] M. Lin, Q. Chen, S. Yan, Network in network, in: Proc. International Conference on Learning Representations, 2014.

[24] S. Kiran, C.K. Thompson, The role of semantic complexity in treatment of naming deficits: training semantic categories in fluent aphasia by controlling exemplar typicality, Journal of Speech, Language, and Hearing Research 46(4)(2003) 773-787.

[25] M. Daneman, R. Case, Syntactic form, semantic complexity, and short-term memory: influences on children's acquisition of new linguistic structures, Developmental Psychology 17(4)(1981) 367.

[26] S. Kiran, Typicality of inanimate category exemplars in aphasia treatment: further evidence for semantic complexity, Journal of Speech, Language, and Hearing Research 51(6)(2008) 1550-1568.

[27] D.A. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (ELUs). <https://arxiv.org/abs/1511.07289>, 2015.

[28] F. Agostinelli, M. Hoffman, P. Sadowski, P. Baldi, Learning activation functions to improve deep neural networks. <https://arxiv.org/abs/1412.6830>, 2014.

[29] B. Pang, L. Lee, Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales, in: Proc. Meeting of the Association for Computational Linguistics (ACL-2005), 2005.

[30] D. Huang, J. Wang, An approach on Chinese microblog entity linking combining baidu encyclopaedia and word2vec, in: Proc. Procedia Computer Science, 2017.

[31] Y. Kim, Convolutional neural networks for sentence classification, in: Proc. the 2014 Conference on Empirical Methods in

Natural Language Processing (EMNLP), 2014.

[32] A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: an astounding baseline. <https://arxiv.org/abs/1403.6382>, 2014.

[33] N. Srivastava, G.E. Hinton, A. Krizhevsky, I Sutskever, R Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15(1)(2014) 1929-1958.

[34] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12(2011) 2493-2537.

[35] D Kingma, J. Ba, Adam: A method for stochastic optimization. <https://arxiv.org/abs/1412.6980>, 2014.

[36] V.K. Singh, R. Piryani, A. Uddin, P. Waila, Sentiment analysis of movie reviews: a new feature-based heuristic for aspect-level sentiment classification, in: Proc. 2013 International Multi-conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013.

[37] R. Socher, J. Pennington, E. Huang, A. Ng, C. Manning, Semi-supervised recursive autoencoders for predicting sentiment distributions, in: Proc. the Conference on Empirical Methods in Natural Language Processing, 2011.

[38] K. Hermann, P. Blunsom, The role of syntax in vector space models of compositional semantics, in: Proc. the 51st Annual Meeting of the Association for Computational Linguistics, 2013.

[39] T. Nakagawa, K. Inui, S. Kurohashi, Dependency tree-based sentiment classification using CRFs with hidden variables, in: Proc. the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010.

[40] H. Zhuang, C. Wang, C. Li, Q. Wang, X. Zhou, Natural language processing service based on stroke-level convolutional networks for Chinese text classification, in: Proc. 2017 IEEE International Conference on Web Services (ICWS), 2017.