Two-level Neural Network for Multi-label Document Classification

Ladislav Lenc[1,2] and Pavel Král[1,2]

[1] Dept. of Computer Science & Engineering
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
[2] NTIS - New Technologies for the Information Society
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
{pkral,llenc}@kiv.zcu.cz

Abstract. This paper deals with multi-label document classification using neural networks. We propose a novel neural network which is composed of two sub-nets: the first one estimates the scores for all classes, while the second one determines the number of classes assigned to the document. The proposed approach is evaluated on Czech and English standard corpora. The experimental results show that the proposed method is competitive with state of the art on both languages.

Keywords: Convolutional Neural Networks, Czech, Deep Neural Networks, Document Classification, Multi-label

1 Introduction

This paper is focused on multi-label document classification using neural networks. This task can be seen as the problem to find a model M which assigns a document $d \in D$ a set of appropriate classes $c \in C$ as follows $M : d \rightarrow c$ where D is the set of all documents and C is the set of all possible document classes (labels).

In our previous work [7], we have used standard feed-forward networks and popular convolutional networks (CNNs) with thresholding to obtain the final classification result. We have shown the superior accuracy of these networks without any manually defined features against the state-of-the-art methods.

In this paper, we propose an alternative multi-label document classification approach which uses another neural classifier to identify the number of labels assigned to the document. An original neural network architecture which is composed of two sub-nets is thus proposed: the first one estimates the scores for all classes, while the second one is dedicated to determine the number of classes. To the best of our knowledge, this approach has never been used for multi-label document classification before.

The proposed approach is evaluated on Czech and English standard corpora. The Czech language has been chosen as a representative of highly inflectional Slavic language with a free word order. These properties decrease the performance of usual methods and therefore, sophisticated methods are beneficial. English is used to compare the results of our method with state of the art.

The following section contains a short review of the usage of neural networks for document classification with a particular focus on multi-label classification approaches. Section 3 describes the proposed model. Section 4 deals with experiments realized on the ČTK and Reuters corpora and then analyzes and discusses the obtained results. In the last section, we conclude the experimental results and propose some future research directions.

2 Related Work

Nowadays, "deep" neural nets outperform majority of the state-of-the art natural language processing (NLP) methods on many tasks with only very simple features. These include for example POS tagging, chunking, named entity recognition and semantic role labelling.

Recurrent convolutional neural nets are used for text classification in [5]. The authors demonstrated that their approach outperforms the standard convolutional networks on four corpora in single-label document classification task.

On the other hand, traditional feed-forward neural net architectures are not used for multi-label document classification very often. These models were popular previously as shown for instance in [8]. They build a simple multi-layer perceptron with three layers (20 inputs, 6 neurons in hidden layer and 10 neurons in the output layer, i.e. number of classes) which gives F-measure about 78% on the standard Reuters dataset.

The feed-forward neural networks were used for multi-label document classification in [15]. The authors have modified standard backpropagation algorithm for multi-label learning which employs a novel error function. This approach is evaluated on functional genomics and text categorization.

Le and Mikolov propose in [6] so called *Paragraph Vector*, an unsupervised algorithm that addresses the issue of necessity of a fixed-legth document representation. This algorithm represents each document using a dense vector. This vector is trained to predict words in the document. The results show that this approach for creating text representations outperforms many other methods including bag-of-words models. The authors obtain new state-of-the-art results on several text classification and sentiment analysis tasks.

A recent study on multi-label text classification was presented by Nam et al. in [10]. The authors use cross-entropy algorithm instead of ranking loss for training and they also further employ recent advances in deep learning field, e.g. rectified linear units activation, AdaGrad learning with dropout [9, 13]. The TF-IDF representation of documents is used as network input. The multi-label classification is done by thresholding of the output layer. The approach is eval-

uated on several multi-label datasets and reaches results comparable or better than the state-of-the-art.

Another method [4] based on neural networks leverages the co-occurrence of labels in the multi-label classification. Some neurons in the output layer capture the patterns of label co-occurrences, which improves the classification accuracy. The architecture is basically a convolutional network and utilizes word embeddings as inputs. The method is evaluated on the natural language query classification in a document retrieval system.

An alternative multi-label classification approach is proposed by Yang and Gopal in [14]. The conventional representations of texts and categories are transformed into meta-level features. These features are then utilized in a learning-to-rank algorithm. Experiments on six benchmark datasets show the abilities of this approach in comparison with other methods.

For additional information about multi-label document classification, please refer the survey [12].

3 Network Architecture

We use two types of neural networks that were proposed in [7] as the first sub-net. The first one is a convolutional neural network (CNN) while the second one is a standard feed-forward neural network (FNN). Therefore, using the feature vector F, both networks learn a function $S = f_1(F)$ which assigns a score S to each of possible labels. The values of the output layer were usually thresholded [10] using a fixed threshold. The labels with values higher than this threshold are then assigned to a document.

In this paper, we replace the thresholding method by another neural classifier and then we merge both nets together. Therefore, the output of the first network is used as an input of the second-level feed-forward network which is used to estimate the number of relevant labels l. Finally, the l labels with the highest scores are assigned to the classified document.

1a) Convolutional Neural Network

The input (vector F) of the CNN is a sequence of word indexes from a dictionary. The network requires fixed-length inputs and the documents thus must be shortened or padded to a specified length N. The following layer is an embedding layer which maps the words to real-valued vectors of the size K. In the convolutional layer we employ N_C kernels of the size $k \times 1$. Rectified linear unit (ReLU) activation is used. The next layer performs the max-over-time pooling. The dropout [13] is then applied due to regularization. The output of this layer is fed to a fully-connected layer with ReLU activation function. The output layer of the size C is another fully connected layer which gives the scores for each possible label. We use either *sigmoid* or *softmax* activation function in this layer.

1b) Feed-forward Neural Network

This network is an alternative to the CNN described previously. The input (vector F) is a bag-of-words (BoW) representation of the documents. It is followed by two fully connected layers. Each of them has a ReLU activation with a subsequent dropout regularization. We use the *softmax/sigmoid* activation in the output layer of the size $|C|$.

2) 2nd-level Feed-forward Neural Network

This network is a multi-layer perceptron with one hidden layer. It takes the output from the underlying network (CNN or FNN) S and learns a function $l = f_2(S)$ that maps the vector S to the number of relevant labels l. The output layer has the *softmax* activation.

Figure 1 shows the architecture of the whole network where the CNN and 2nd-level FNN are merged. Due to the space limits, the architecture of the second network which merges together the two FNNs is not depicted.

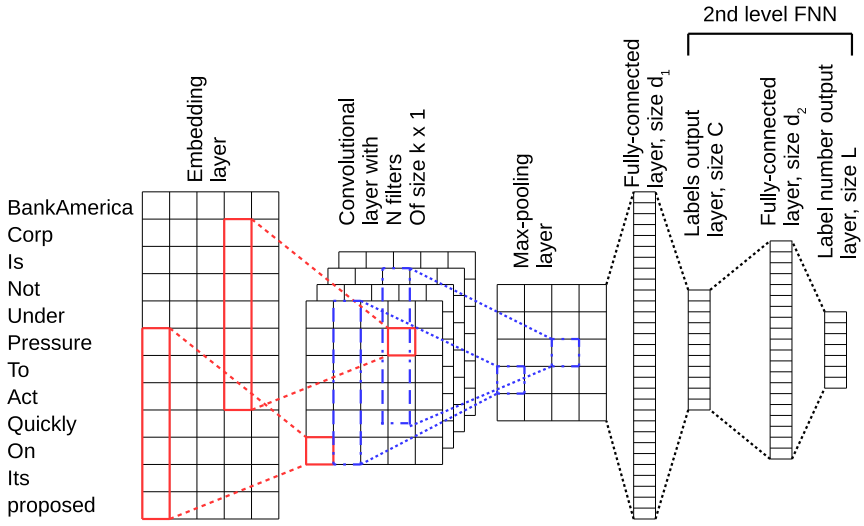

Fig. 1. The architecture of the proposed network - CNN + FNN

The whole network learns the complex function $l, S = f(F) = f_2(f_1(F))$ When we trained the whole network at once, unfortunately, the convergence was not very good. Therefore, we decided to train both sub-nets independently. First, we train the CNN (or FNN) which gives the score S for all labels, then the connected 2nd-level FNN is trained using these scores S. Both sub-nets are learned using adaptive moment estimation (Adam [3]) optimization algorithm.

4 Experiments

4.1 Tools and Corpora

For implementation of all neural nets we used Keras tool-kit [2] which is based on the Theano deep learning library [1]. It has been chosen mainly because of good performance and our previous experience with this tool. For evaluation of the multi-label document classification results, we use the standard recall, precision and F-measure (*F1*) metrics [11]. The values are micro-averaged. To measure the performance of the second sub-net we utilize label accuracy (L-ACC) and mean absolute error (MAE).

Czech Text Document Corpus v 1.0 This corpus is composed of 11,955 documents provided by ČTK and contains 2,974,040 words. The documents are annotated from a set of 60 categories as for instance agriculture, weather, politics or sport out of which we used 37 most frequent ones. The category reduction was done to allow comparison with previously reported results on this corpus where the same set of 37 categories was used. Average number of categories per document is 2.55. 500 randomly chosen documents are reserved for development set while the remaining part is used for training and testing. Left part of Figure 2 illustrates the distribution of the documents depending on the number of labels, while the right part shows the distribution of the document lengths (in word tokens). This corpus is freely available for research purposes at `http://home.zcu.cz/~pkral/sw/`. We use the five-folds cross validation procedure for all experiments on this corpus.

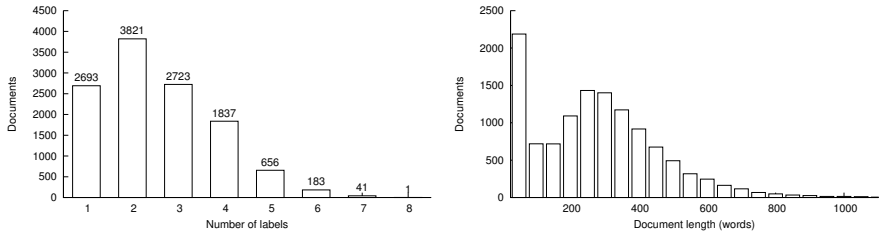

Fig. 2. Distribution of documents depending on the number of labels (left) and distribution of the document lengths (right)

Reuters-21578 English Corpus The Reuters-21578[3] corpus is a collection of 21,578 documents. As suggested by many authors, the training part is composed of 7769 documents, while 3019 documents are reserved for testing. The number of possible categories is 90 and average label/document number is 1.23. This

[3] http://www.daviddlewis.com/resources/testcollections/reuters21578/

dataset is used in order to compare the performance of our method with state-of-the-art approaches.

4.2 System Configuration

In this section we summarize the important parameters that we used in our system configuration. The preprocessing was the same for both Czech and English corpora. We convert all texts to lowercase and replace all numbers by one common token.

The dictionary size is set to 20,000 for both networks. The document length is unified to 400 tokens and the embedding size is 300 for the CNN. The convolutional layer utilizes 40 kernels of the size 16×1. The fully connected layer in CNN has 256 neurons. The two hidden layers of FNN have 1024 and 512 neurons respectively. All dropout rates are set to 20%. In the case of the 2nd-level FNN we use hidden layer with 100 neurons. All the networks are trained for 20 epochs and with the mini-batch size 32.

4.3 Results on the Czech Corpus

The first experiment (see Table 1) shows the performance of the individual networks with the thresholding method. It is realized in order to compare the results of the proposed neural net with state of the art[4]. The threshold values are set on the development data. This table shows that CNN with sigmoid activation function gives the best classification results.

Table 1. Results on Czech corpus with thresholding method, thresholds set on the development corpus

Method	**Prec.**	**Recall**	**F1 [%]**	**Threshold**
CNN sigmoid	87.68	79.09	**83.17**	0.19
CNN softmax	80.84	80.54	80.69	0.06
MLP sigmoid	80.03	83.35	81.66	0.15
MLP softmax	67.78	90.99	77.69	0.04

The second experiment (see Table 2) presents the results obtained with the proposed neural network method. This table shows that this approach performs better when the sigmoid activation function is used. This behavior is not surprising because sigmoid function usually suits better for the multi-label classification problems. This table further shows that this approach outperforms the reference thresholdind method (see Table 1). This experiment also shows that both network topologies (CNN + FNN or FNN + FNN) are comparable. Note that L-ACC is the label accuracy of the second level FNN and MAE is its mean absolute error. It is obvious that there is still room for improvement in the 2nd-level FNN performance.

[4] This approach has been proposed in [7].

Table 2. Results on the Czech corpus using the proposed neural network approach

Method	Prec.	Recall	F1 [%]	L-ACC	MAE
CNN sigmoid	87.20	81.13	**84.06**	63.54	0.46
CNN softmax	84.13	80.20	82.12	60.96	0.53
MLP sigmoid	85.61	82.82	**84.19**	64.47	0.48
MLP softmax	77.28	85.30	81.09	57.11	0.62

4.4 Results on Reuters-21578

The third experiment (see Table 3) shows the performance of the proposed approach on standard English Reuters dataset. This experiment was realized in order to show its robustness across languages and to compare our method with state of the art (SoTa). The results show that especially CNN with the sigmoid activation has very good performance and is comparable with the best performing approach of Nam et al. [10] (SoTa). Note that the authors use TF-IDF representation of documents which is slightly more sophisticated than ours.

Table 3. Results on English Reuters corpus using the proposed neural network approach

Method	Prec.	Recall	F1 [%]	L-ACC	MAE
CNN sigmoid	89.79	84.99	**87.32**	88.17	0.17
CNN softmax	87.52	83.96	85.70	85.80	0.19
MLP sigmoid	85.16	83.22	84.18	86.27	0.19
MLP softmax	81.52	83.24	82.37	81.29	0.23
BR_R [10] (SoTa)	89.82	86.03	**87.89**	-	-

5 Conclusions and Perspectives

In this paper, we have proposed a novel neural network for multi-label document classification. This network is composed of two sub-nets where the first one estimates the scores for all classes, while the second one is used to determine the number of classes. We have evaluated the proposed approach on Czech and English standard corpora. We have experimentally shown that the proposed method is competitive with state-of-the-art methods on both languages

The experiments have shown that the 2nd-level FNN performance could be further improved. This is thus the first perspective. Another possibility for improvement is using manually pre-trained embeddings. However, in this paper, we did not concentrate on this issue and it will thus be solved in our future work. We also would like to experiment with different network types, as for instance LSTM or recurrent CNNs.

Acknowledgements

This work has been supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

References

1. Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: a cpu and gpu math expression compiler. In: Proceedings of the Python for scientific computing conference (SciPy). vol. 4, p. 3. Austin, TX (2010)
2. Chollet, F.: keras. `https://github.com/fchollet/keras` (2015)
3. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
4. Kurata, G., Xiang, B., Zhou, B.: Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In: Proceedings of NAACL-HLT. pp. 521–526 (2016)
5. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification (2015)
6. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: ICML. vol. 14, pp. 1188–1196 (2014)
7. Lenc, L., Král, P.: Deep neural networks for Czech multi-label document classification. In: 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016). Springer, Konya, Turkey (3-9 April 2016)
8. Manevitz, L., Yousef, M.: One-class document classification via neural networks. Neurocomputing 70(7-9), 1466–1481 (2007), `http://www.scopus.com/inward/record.url?eid=2-s2.0-33847410597&partnerID=40&md5=3d75682f283e19695f2857dea9d9f03f`
9. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 807–814 (2010)
10. Nam, J., Kim, J., Mencía, E.L., Gurevych, I., Fürnkranz, J.: Large-scale multi-label text classification - revisiting neural networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 437–452. Springer (2014)
11. Powers, D.: Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies 2(1), 37–63 (2011)
12. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1), 1–47 (2002)
13. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1), 1929–1958 (2014)
14. Yang, Y., Gopal, S.: Multilabel classification with meta-level features in a learning-to-rank framework. Machine Learning 88(1-2), 47–68 (2012)
15. Zhang, M.L., Zhou, Z.H.: Multilabel neural networks with applications to functional genomics and text categorization. Knowledge and Data Engineering, IEEE Transactions on 18(10), 1338–1351 (2006)