Combination of Neural Networks 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
{llenc,pkral}@kiv.zcu.cz

Abstract. This paper deals with multi-label classification of Czech documents using several combinations of neural networks. It is motivated by the assumption that different nets can keep some complementary information and that it should be useful to combine them. The main contribution of this paper consists in a comparison of several combination approaches to improve the results of the individual neural nets. We experimentally show that the results of all the combination approaches outperform the individual nets, however they are comparable. However, the best combination method is the supervised one which uses a feed-forward neural net with sigmoid activation function.

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

1 Introduction

This paper deals with multi-label document classification by neural networks. Formally, this task can be seen as the problem of finding a model M which assigns a document $d \in D$ a set of appropriate labels $l \in L$ as follows $M : d \rightarrow l$ where D is the set of all documents and L is the set of all possible document labels. In our previous work [1], we have compared standard feed-forward networks (i.e. multi-layer perceptron) and popular convolutional networks (CNNs).

The resulting F-measures of these nets were high, however these values are still far from perfect. Therefore, in this paper, we use several approaches to combine individual networks in order to improve the final classification score. The main contribution of this paper thus consists in a comparison of classifier combination methods for multi-label classification which has, to the best of our knowledge, never been done on this task

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

before. The methods are evaluated on the documents in Czech language, being a representative of highly inflectional Slavic language with a free word order. These properties decrease the performance of usual methods and therefore, a more sophisticated parametrization is beneficial. This evaluation is another contribution of this paper.

The rest of the paper is organized as follows. Section 2 describes the combination methods. Section 3 deals with experiments realized on the ČTK corpus and then discusses the obtained results. In the last section, we conclude the experimental results and propose some future research directions.

2 Networks and Combination Approaches

2.1 Individual Nets

We use a feed-forward deep neural network (FDNN) and a convolutional neural net (CNN) with two different activation functions, namely *sigmoid* and *softmax*, in the output layer. Our CNN is motivated by Kim [2], however we used only one-dimensional convolutional kernel. The topologies of our nets are detailed in our previous work [1].

2.2 Combination

We consider that the different nets keep some complementary information which can compensate recognition errors. We also assume that similar network topology with different activation functions can bring some different information and thus that all nets should have its particular impact on the final classification. Therefore, we consider all the nets as the different classifiers which will be further combined.

Two types of combination will be evaluated and compared. The first group does not need any training phase, while the second one learns a classifier.

Unsupervised Combination The first combination method compensates the errors of individual classifiers by computing the average value from the inputs. This value is thresholded subsequently to obtain the final classification result. This method is called hereafter *Averaged thresholding*.

The second combination approach first thresholds the scores of all individual classifiers. Then, the final classification output is given as an agreement of the majority of the classifiers. We call further this method as *Majority voting with thresholding*

Supervised Combination We use another neural network of type multi-layer perceptron to combine the results. This network has three layers: $n\times$ 37 inputs, hidden layer with 512 nodes and the output layer composed of 37 neurons (number of categories to classify). n value is the number of the nets to combine. This configuration was set experimentally on the preliminary results. We also evaluate and compare, as in the case of individual classifiers, two different activation functions: *sigmoid* and *softmax*. These combination approaches are hereafter called *FNN with sigmoid* and *FNN with softmax*.

3 Experiments

3.1 Tools and Corpus

For implementation of all neural-nets we used Keras tool-kit [3] which is based on the Theano deep learning library [4].

For the following experiments we used the Czech text documents provided by the ČTK. This whole corpus contains 2,974,040 words belonging to 11,955 documents. 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. We have further created the development set which is composed of 500 randomly chosen samples removed from the entire corpus. 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 following experiments, where 20% of the corpus is reserved for testing and the remaining part for training of our models. The optimal value of the threshold is determined on the development set. For evaluation of the multi-label document classification results, we use the standard recall, precision and F-measure (*F1*) metrics. The results are micro-averaged.

3.2 Results of the Individual Networks

The first experiment (see Sec. 1 of Table 1) shows the results of the individual neural nets with sigmoid and softmax activation functions. These results demonstrates very good classification performance of all individual networks.

Approach	Prec.	Recall	F1 [%]
1. Invidual networks			
(a) FDNN with softmax	84.4	82.1	83.3
(b) FDNN with sigmoid	83.0	81.2	82.1
(c) CNN with softmax	80.6	80.8	80.7
(d) CNN with sigmoid	86.3	81.9	84.1
2. Unsupervised combination			
network (a) & (c) & (d) combined by *averaged thresholding*	86.7	83.5	85.1
network (a) & (b) & (d) combined by *majority voting with thresholding*	87.5	82.6	85.0
3. Supervised combination			
all networks combined by *FNN with softmax*	85.7	83.6	84.6
all networks combined by *FNN with sigmoid*	88.0	82.8	**85.3**

Table 1. Experimental results

3.3 Results of Unsupervised Combinations

The second experiment shows (see Sec. 2 of Table 1) the results of *Averaged thresholding* and *Majority voting with thresholding* methods. These results confirm our assumption that the different nets keep complementary information and that it is useful

to combine them to improve classification scores of the individuals networks. These results further show that the performance of both methods are comparable.

Note that due to the space limit, only the best performing combination for each method is reported in this table.

3.4 Results of Supervised Combinations

The following experiments show the results of the supervised combination method with an FNN (see Sec. 2.2). We have evaluated and compared the nets with both sigmoid and softmax (see Sec. 3 of Table 1) activation functions.

These results show that these combinations have also positive impact on the classification and that sigmoid activation function brings better results than softmax. Moreover, as supposed, this supervised combination slightly outperforms both previously described unsupervised methods.

4 Conclusions & Future Work

In this paper, we have used several combination methods to improve the results of individual neural nets for multi-label document classification of Czech text documents. We have shown that it is useful to combine the nets to improve the classification score of the individual networks. We have also proved that the thresholding is a good method to assign the document labels of multi-label classification. We have further shown that the results of all the approaches are comparable. However, the best combination method is the supervised one which uses an FNN with sigmoid activation function. The F-measure of this approach is 85.3%.

We further analyzed the final results and discovered that the classification should be still improved if the number of classes is known for every document. Therefore, the first perspective is to build a meta-classifier to provide this information. The consecutive multi-label classification will be using the class dependent thresholds. The next perspective consists in proposing a novel combination method based on deep neural network. The main challenge of this work will be to found an optimal network topology with a reasonable number of parameters to avoid the overfitting. We also would like to experiment with confidence measures to improve the final classification results.

References

1. 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), Konya, Turkey, Springer (2016)
2. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
3. Chollet, F.: keras. https://github.com/fchollet/keras (2015)
4. 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). Volume 4., Austin, TX (2010)