UDC 004.89
[1] **Zaid Musbah**
Postgraduate
[2] **Taras Lehinevych**
Master student
[3] **Andrii Glybovets**
Candidate of Physic-Mathematical Sciences, professor, associate professor of Faculty of Informatics
[1] National University of Kyiv-Taras Shevchenko, Kyiv
[2,3] National University "Kyiv-Mohyla Academy", Kyiv

CROSS-LANGUAGE TEXT CLASSIFICATION WITH CONVOLUTIONAL NEURAL NETWORKS

Introduction. Text classification or text categorization problem is currently one of the most observed in the field of information and computer sciences. The task is to assign a text to one or more classes or categories and it becomes more difficult if we have to deal with different languages. This problem is called cross-language text classification problem. In our paper [1] was shown that cross-language multi-label text classification can be handled by a deep learning system without artificially embedding knowledge about words, phrases, sentences or any other syntactic or semantic structures associated with a language.

Model architecture. First of all, the proper dataset was created for cross-language multi-label classification problem based on extracted articles from Wikipedia. Every sample contains article text and categories. In order to feed sample into neural network model the text and categories were converted into the vectors of integers by using mapping index (every work or category corresponds to unique integer number). We proposed the deep neural network model architecture for this problem. The first layer is embedding one. It receives as an input a matrix of integers that corresponds to the text of an article and produce the smaller matrix representation. The following two layers are quite the same. They perform convolutional operations with different window sizes (32 and 16)[2]. The last convolutional layer is the combination of three sub-layers. Every sub-layer gets an input the output from previous layer (so all sub-layers get the same input) and perform convolutional and max-pooling operation with windows sizes (5, 4 and 3). The outputs of three sub-layers are concatenated and feed into fully connected layer, which output is the probability distribution over categories. The whole deep neural network model is trained by using backpropagation algorithm.

The model was trained with the following hyperparameters:
-rectified linear units;
-windows (h) of 32, 16, 5, 4, 3;
-dropout rate (p) of 0.5;
-l2 constraint (s) of 5;
-128 filters per filter size;
-mini-batch size of 35;

-Adam update rule[3];

-one epoch ;

The described model about has multiple benefits. First of all, it does not require parallel data between all languages or bilingual dictionaries, what is often a bottleneck, because in many real world scenarios such parallel data may not be available. Secondly, there is no need to use artificially embedding knowledge about words, phrases, sentences or any other syntactic or semantic structures associated with a language. Finally, representation of text for any language is learned by neural network and could be used in other application.

Results. The accuracy result on the validation set (set of data that was not used during the training or testing of the model) is equal to 84.56 ± 2.6 %. Moreover, we run the model on new language (it was not used for training or testing) in order to validate our hypothesis about transfer learning. The articles collected for German language were used and the model shows accuracy equal to 63.34 += 1.98%.

Conclusion. Despite the fact that the limitation of this approach is requirement that all articles have the same length, the convolutional neural networks show good results for multilingual long-form texts such as multi-language Wikipedia articles. The proposed solution could be used in variety of applications such as search engines, e-libraries, knowledge bases, any information retrieval systems and software that work with multilingual documents.

References

1. Musbah, Z., Lehinevych, T., Glybovets, A., (2017). Cross-language text classification with convolutional neural networks from scratch. EUREKA: Physics and Engineering, (2), 24-33.

2. Zhang, X., LeCun, Y., (2015). Text understanding from scratch. arXiv preprint arXiv:1502.01710.

3. Ko, Y., Seo, J. (2000). Automatic text categorization by unsupervised learning. In Proceedings of the 18th conference on Computational linguistics -Volume 1 (pp. 453-459). Association for Computational Linguistics.