Individual document classification; promising results from Convolutional Neural Networks and Graph Embeddings.

Bart Thijs

KU Leuven, ECOOM & Dept MSI, Naamsestraat 61, 3000 Leuven, Belgium
bart.thijs@kuleuven.be

Introduction

Classification schemes used in most of the commercial multidisciplinary bibliographic databases are journal-based systems and lacking a proper document-based classification system. The WoS Core Collection uses their 'Web of Science Categories' which comprises about 250 subject areas and each journal indexed in this database is assigned to one or more subject category. In addition, Clarivate Analytics is applying their system of 22 research areas for the calculation of the Essential Science Indicators. This, too, is a journal-based system with each journal have a single assignment. Analogously, Scopus has a two-level journal classification scheme 'All Science Journal Classification' with 304 categories at the most fine-grained level and 27 top level categories. Several attempts have been conducted in the past to circumvent these shortcomings. Recently, Dimensions was released as the first bibliographic database with a document-based classification scheme but not without questions on is reliability and validity (Bornmann, 2020 or Singh et al., 2020). The topic has also been quite popular in some AI contests or hackathons. In this context, convolutional neural networks are often proposed for this task as CNN have a proven track record for text classification in other applications. Other approaches for the classification of individual scientific papers are based on citation links (Subelj et al, 2016; Waltman & van Eck, 2012 or Glänzel et al., 1999). In fact, this study tries to identify the added value of network data for the classification task as is done with hybrid approaches for document clustering and unsupervised learning (Thijs & Glänzel, 2018). The study does not present a ready-to-use article-based classification but tries to identify possible opportunities offered by the advent of new graph-based techniques and attempts to shed some light on possible shortcomings and hindrances that affects the reliability, validity and applicability. Important to mention here is the absence of a proper ground truth of the document classification.

Data

Two publications sets have been used. First, a set of 40.790 publications indexed in the Web of Science Core Collection between 2007 and 2019 and assigned to a set of ISI subject categories compromising the field 'Non-internal medicine' of the Leuven-Budapest classification scheme and attributed to one of the nine disciplines within this field. The selected papers have between 3 and 7 citation links to other papers in the same field. The direction of the citation is neglected in this study. The construction of the datasets augments the probability that the paper is indeed properly assigned to the provided class. Consequently, the obtained classification models will not be applicable to a broad set of papers without additional training or adaptation. A second data set, mainly used for validation purposes, is a set of publications published in multidisciplinary journals during the same time period and citing or being cited by documents from the first set.

Methods

Three classification models are compared in the study. First, two deep learning models, built around the combination of a convolutional neural network and a pooling layer preceded by a word embedding, are created within the Keras framework. The first one is a simple model with only one combination while the second one is more complex with three concatenated CNN's with each different filter size. The models are trained on a random selection of 70% of the first paper set. Both models produce for each input document an output vector with a predicted score for each of the 9 possible labels.

The third model is based on the StellarGraph implementation of the GraphSage framework (Hamilton, 2017) which creates low level node embeddings using not only the features of the node itself but also the features and labels of the neighborhood by applying a random walk selection procedure. An additional advantage of this framework is that it allows to efficiently generate representations of unseen data. The feature vector for the graph embedding is the result of a document embedding using Doc2Vec from the Gensim library with a dimensionality of 150.

Results

The first model is trained in 5 epochs and reaches an accuracy of 99.95% for the training set being an overfitted result as the validation set has an accuracy of 81.33%. The timing took 2h and 20 minutes. Figure 1. plots the accuracy of the predictions on the training and validation set for the first model.

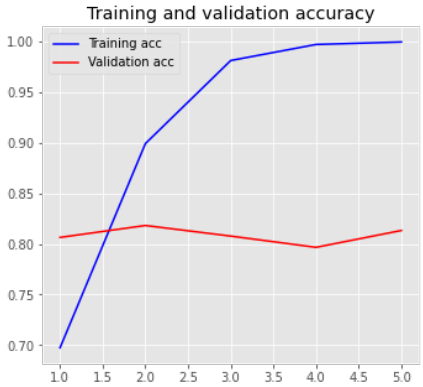

Figure 1: Accuracy of the Model 1 predictions

The second, more complex model took 27 hours and 23 minutes to reach a similar overfitted result for the training set and an accuracy of 82.97% for the training set. As the model is much slower in reaching its best results, the training took 25 epochs.

The third model starts from a Doc2Vec paragraph embedding and is using a random walk neighbor selecting mechanism with up to 10 first order neighbors and up to 5 in the second order. The accuracy of the training set is 80.34% and for the validation set 78.23%. This shows clearly absence of an overfitting problem. The results are

plotted in figure 2. This result was obtained after only 6 minutes with 40 epochs.

In order to have a fair comparison between the time needed for the training, the doc2vec embedding has to be taken into account. But also, being less the 40 minutes the total time from text and network data to prediction is well below one hour.

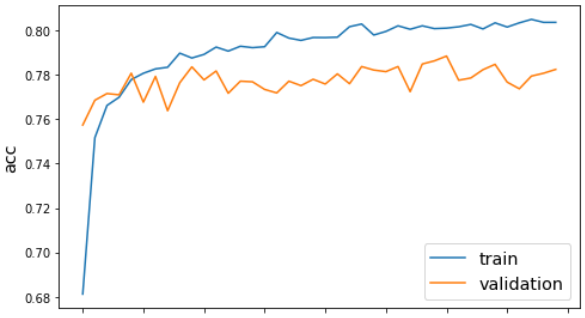

Figure 2. Accuracy of the Model 3 predictions

In fact, a TSNE plot (Figure 3) of both the document embedding using Doc2Vec and the node embedding from GraphSage shows the improvement of the classification adding the network information, the labels of citing and cited documents.

In a last step, the classification models have been applied on a selection of publications from multidisciplinary journals and manually validated. The results are promising as prediction scores can be used and deviations between the three predictions can be used for highlighting deviating cases.

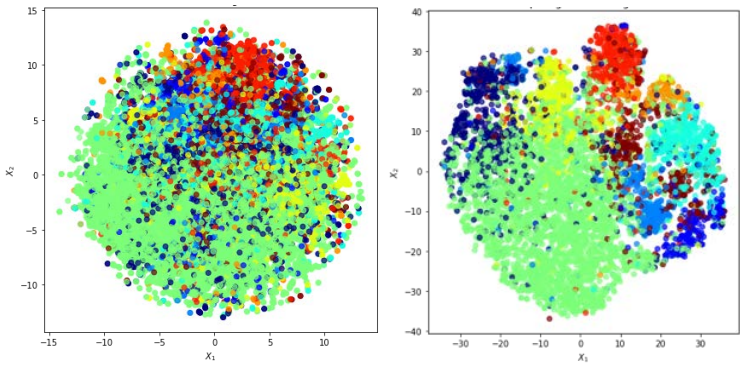

Figure 3. TSNE plot of Doc2Vec (left) and GraphSage (right) embeddings

References

Bornmann, L., (2020). Field classification of publications in Dimensions: a first case study testing its reliability and validity. *Scientometrics*, 117 (1), 637-640.

Glänzel, W., & Schubert, A. & Czerwon, H.J., (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics*, 44(3), 427-439.

Hamilton, W.L., Ying, R., Leskovec, J. (2017) Inductive Representation Learning on Large Graphs. arXiv:1706.02216.

Singh, P., Piryani, R., Singh, V.K. & Pinto, D., (2020). Revisiting subject classification in academic databases: A comparison of the classification accuracy of Web of Science, Scopus & Dimensions. *Journal of Intelligent & Fuzzy Systems*. 39(2), 2471-2476.

Subelj, L, van Eck, N.J., Waltman, L. (2016). Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods, *PLOS ONE*, 11 (4), e0154404.

Thijs, B. & Glänzel, W. (2018). The contribution of the lexical component in hybrid clustering, the case of four decades of "Scientometrics". *Scientometrics*, 115 (1), 21-33.

Waltman, L., & van Eck, N.J., (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378-2392.