# Authorship identification using deep learning
# Deep learning alapú szerzőazonosítás

Füleki Fábián, Jani Balázs Gábor, Torner Márton
*Students of Budapest University of Technology and Economics*
*Project work for Deep Learning course (VITMAV45)*
*Team: LoremIpsum*

***Abstract*— In this paper we are publishing a proof of concept of a Bidirectional Long Short-Term Memory network to identify the author of a given article. Every author has its own writing style that can be found in his/her written articles. The network is aiming to learn the commonly used phrases and the typical structure of sentences for each author. The dataset consists of 50 authors with 50 articles from each of them, which gives us a baseline accuracy of 2%. Despite this low value the best presented network reaches a 94% overall accuracy. We are also demonstrating an own approach based on a quite new research scope, which is called Attention. This mechanism can provide a better solution for the vanishing gradients problem.**

***Absztrakt*— Jelen publikáció keretei között bemutatunk egy kétirányú hosszú rövidtávú memória (BiLSTM) alapú neurális hálót, amely képes meghatározni egy adott cikk íróját. Minden szerző rendelkezik egy saját, majdhogynem egyedi stílussal. Ezeket a jellemző mondatstruktúrákat, gyakran használt szófordulatokat és egyedi kifejezéseket a hálózat könnyen meg tudja tanulni és fel tudja ismerni. Az adathalmazunk 50 szerzőből áll, mindegyikükhöz pedig 50 cikk tartozik. A véletlenszerű tippelés esetén tapasztalható 2%-os pontossághoz képest a legjobb hálózatunk képes 94%-os pontosságot elérni. Bemutatunk továbbá egy saját megközelítést is, mely a rekurrens neurális hálókban alkalmazott, nemrégiben kutatott, úgynevezett "attention" mechanizmuson alapszik. Ez a megközelítés képes az LSTM hálózatoknál előforduló, úgynevezett elenyésző gradiens problémára megoldást nyújtani, sőt, még a "time distributed" rétegeknél is jobban teljesíthet.**

## I. INTRODUCTION

Authorship identification is one of the major problems of written media and forensic science. Identifying someone as the author of an article or even just a few sentences can be harmful and helpful as well. Our goal here is to provide a quite reliable author identifier tool of printed text. Using handwriting would be even more accurate, but we would like to create a general solution for the problem. Modern media is based on online, written articles anyway. Author identification could be easily done manually by using Stylometry techniques, because most of the authors have personal habits, commonly used phrases, preferred locutions. We attempt to create a neural network to automate this identification process. This task corresponds to the well-known classification problem, where all the samples belong to one of the known class, or author in this case. In this paper we do not attempt to add an outlier class (unknown author) to the samples. We are going to use the Reuter_50_50 dataset, which consists of fifty authors and fifty articles for all of the authors. Fifty authors do not appear to be a big number, but it will be adequate for the proof of concept. Fifty authors mean that we are going to have a pretty low baseline of 2%. Our solution is based on a Bidirectional Long Short-Term Memory network. LSTM can really help, because the phrases and locutions usually consist of more than two words, and so the network has to remember for the previous words. Because sentences usually contain anaphoric and cataphoric references, we attempt to use Bidirectional LSTM, which converges faster than LSTM and reaches a better accuracy, although every epoch takes roughly twice as much time. Certain phrases can overlap sentences, so a good solution would be to process more sentences at once, in our research we are using three following sentences as an input. This solution has achieved an accuracy improvement from around 60% to 90% and also proves that not the content but the phrases are used for the identification.

## II. SUBJECT, PREVIOUS SOLUTIONS

Handwriting is a widely used data source for author identification. But we are going to use written text as input, so we have to approach this problem completely different. Our baseline paper will be the "Deep Learning based Authorship Identification" publication from students of the Stanford University, California. In this paper we can find two type of neural networks. The first one is a gated recurrent unit network and the second one is a long short-term memory network. In this paper we are focusing on using a long short-term memory network. In the baseline article the Reuters_50_50 dataset has been used as well, so we can compare the results of the tests.

## III. IMPLEMENTATION

### A. Dataset

#### 1) Source of the dataset

Our primary dataset is the Reuters_50_50 (C50), which is a subset of Reuters Corpus Volume I(RCVI). The RCV1 is archive of categorized newswire stories, made public for research purposes by Reuters, Ltd. The C50 collection consist of 50 texts for each of the 50 top author, for training and separately the same amount for testing purpose (5000 texts in total). This dataset has been previous used by previous studies of authorship recognition, so we can compare our results at the end of the training.

*2) Preprocessing of the articles*

Our goal here is to format the sentences into data points, which can be used as input for the neural network. We are using Spacy for representing each word as a series of numbers. This is called Global Vectors for Word Representation (GloVe), which uses co-occurrence matrices for the word representations. Spacy can determine for each word which part of the sentence it is as well. This information is called part of speech and can be very useful. If a word cannot be processed by Spacy because it isn't in the language pack, we don't include it in the processed sentence. Stop words and punctuations are not really useful in our scenario, so we are removing them as well.

We don't want to lose any information about coherence between sentences, so every input data will consist of three sequential sentences. Therefore, every sentence will be included three times in the dataset, this is some kind of data augmentation as well.

## B. Neural Network

In our project we tried to use 3 different networks to solve the authentication problem. The inputs are sentences (or sentence sequences as they contain more semantic information) so a timestep contains one word represented as a (301 dimension) vector.

Our first try was a simple LSTM model (Long Short-Term Memory) because recent studies have shown the power of these networks at sequence analysis. The first layer is an LSTM layer since we did our embedding as it was described in the previous sections. We don't want to use only the last output as it compresses all the information from 75 timesteps in one vector so we used a time distributed version of LSTM not to lose potentially important data in the long process. The next is an average pooling layer which is followed by a softmax classifier fully-connected layer.

The next network tries to improve performance with a BiLSTM model (Bidirectional Long Short-Term Memory). The conception is that in long articles there are not only cataphoric but anaphoric references and BiLSTM networks different from make their prediction by processing the sentence sequences in both directions (forward and backward). The two output vectors (per timestep since we use time distributed version here also) are summed by element. This part is visualized on Figure 1. The following layers are the same as at the previous network.

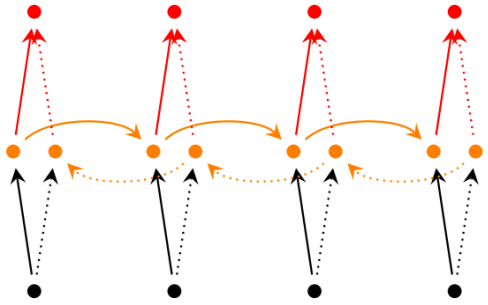

*Figure 1*

The third model is a bit more complex based on our ideas. Despite LSTM networks are more powerful at sequence analysis than previous solutions, some problems have yet to be solved. One of the biggest issues is that they suffer from data loss at analyzing long sequences as they compress the information in one vector/matrix. Time distributed models give a solution to this problem but not a perfect one. Researches have shown the advancement we can make with Attention Mechanisms. The examples are mainly related to CNNs but we tried to implement a simplified model based on the concept. The first part here is a BiLSTM layer, but the final outputs of the two directions are not summarized, they are concatenated along with a third vector containing the time distributed sum of the outputs at each time step. This concept tries to extract more information from the computations. The concatenation does not wrap the knowledge and the time distributed merge creates peeks at points, where the calculation in each (or most of the) timestep(s) predict the same. This part is followed by the softmax classification.

## C. Approaches

In this project we tested our models efficiency in different conditions. The first approach used only one sentence as the input but with this we couldn't reach high accuracy as it only learns the sentence level references and in long articles the style of an author is also strongly marked by the connections across the sentences. Even the simplest model can achieve a huge improvement by creating bundles (more than one sentence) and using them as outputs. This is shown in Figure 2.


*Figure 2*

The training process time can shorten with a Convolutional layer put before LSTM, but it lowers the accuracy of the model. Using more LSTM layers also makes the efficiency worse.

In Figure 3, 4 and 5 we can see how more neurons in the LSTM/BiLSTM layer make each model better.

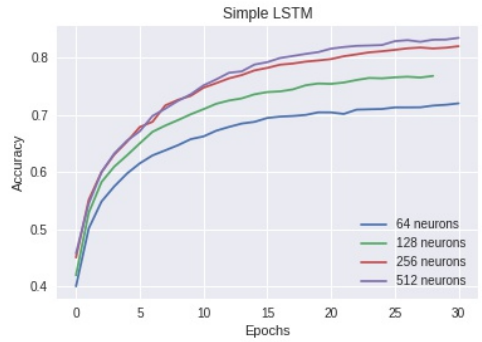


*Figure 3*

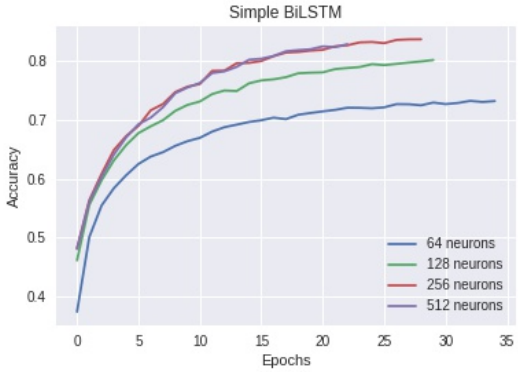


*Figure 4*

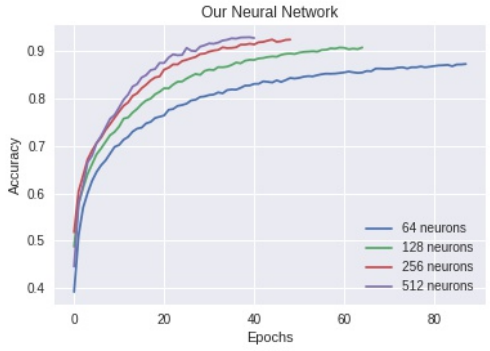


*Figure 5*

The best configuration of each neural network is compared on Figure 6. This shows that the third model can reach the highest accuracy (but this takes the most time to train).

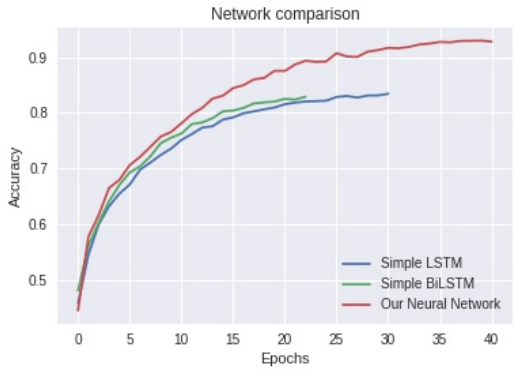


*Figure 6*

The test accuracies of these networks are:

| | Simple LSTM | Simple BiLSTM | Our Neural Network |
|---|---|---|---|
| Test accuracy | 83.70% | 84.00% | 93.56% |

We tested training our network on 10 articles per author, and it has shown similar results.

### *D. Final evaluation, conclusions*

In this project, we studied different deep learning models on authorship identification. We designed and implemented 3 models for authorship identification and shown that simple LSTM and BiLSTM solutions have limitations and can be outperformed by using new techniques. A model with 94% accuracy on identifying the real author of an article from 50 authors could be used to solve real world problems.

### *E. Future plans*

Many future works could be done to improve accuracy such as implementing a more complex Attention mechanism or using named entity relation tags above part-of-speech tagging to express more semantic information from the articles. The models could be extended with an Unknown class for better usage.
Based on our research and models the problem of grouping authors by style or generating texts with the style of an author could be solved.

## REFERENCES

Qian, Chen, Tianchang He, and Rao Zhang. "Deep Learning based Authorship Identification."
Available: http://web.stanford.edu/class/cs224n/reports/2760185.pdf

Bagnall, Douglas. "Author identification using multi-headed recurrent neural networks." *arXiv preprint arXiv:1506.04891* (2015)
Available: https://arxiv.org/abs/1506.04891

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. “Overview of the Author Identification Task at PAN 2015”
Available: http://ceur-ws.org/Vol-1391/inv-pap3-CR.pdf

Jeffrey Pennington, Richard Socher, Christopher D. Manning. "GloVe: Global Vectors for Word Representation", Computer Science Department, Stanford University, Stanford, CA 94305
Available: https://nlp.stanford.edu/pubs/glove.pdf

Douglas Bagnall. "Author identification using multi-headed recurrent neural networks",
Available: https://arxiv.org/pdf/1506.04891.pdf

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, Eduard Hovy. “Hierarchical Attention

Networks for Document Classification", Carnegie Mellon University, Microsoft Research, Redmond Available: http://www.cs.cmu.edu/~./hovy/papers/16HLT-hierarchical-attention-networks.pdf