Efficient Authorship Assignment

Nishanth TA
Department of Computer Science
Shiv Nadar University
Noida, India
nt608@snu.edu.in
Sachin Sriramagiri
Department of Computer Science
Shiv Nadar University
Noida, India
ss804@snu.edu.in

May 17, 2019

Abstract

Authorship attribution supported by statistical or computational methods has a long history starting from 19th century and marked by the seminal study of Mosteller and Wallace (1964). During the last decade, this scientific field has been developed substantially taking advantage of research advances in areas such as machine learning, information retrieval, and natural language processing. In this paper, we present a method to efficiently ascertain an author using salient differentiating features learnt through a variety of methods and contrast these methods with one another. We also discuss the scope of these methods and list their limitations subsequently.

1 Introduction

Authorship assignment stems from the ability of feature detectors to discern the essential hallmarks of the works of a given author. The majority of employed methods for authorship attribution currently use manual feature extraction methods and algorithms based on lexical, character and syntactic features, with the requirement of deep domain knowledge and vulnerabilities to changes in writing style. The pioneering work on authorship assignment can be attributed to Mendenhall (1887) on the plays of Shakespeare to discern their originality. Statistical modeling was extensively used in the first half of the 20th century, with a multitude of papers involving these statistical studies. Later, the detailed study by Mosteller and Wallace (1964) [1] on the authorship of 'The Federalist Papers' (a series of 146 political essays written by John Jay, Alexander Hamilton, and James Madison, twelve of which claimed by both Hamilton and Madison) was undoubtedly the most influential work in authorship attribution. Their method was based on Bayesian statistical analysis of the frequencies of a small set

of common words (e.g., 'and', 'to', etc.) and produced significant discrimination results between the candidate authors.

2 Related Work

The work of Mosteller and Wallace (1964)[1] led to to non-traditional authorship attribution studies as opposed to traditional human expert-based methods. Since then and until the late 1990s, research in authorship attribution was dominated by attempts to define features for quantifying writing style, a line of research known as 'stylometry' (Holmes, 1994; Holmes, 1998)[2]. Hence, a great variety of measures including sentence length, word length, word frequencies, character frequencies, and vocabulary richness functions had been proposed. Rudman (1998)[3] estimated that nearly 1,000 different measures had been proposed that far. The authorship attribution methodologies proposed during that period were computer-assisted rather than computer-based, meaning that the aim was rarely at developing a fully-automated system. the main problem of that early period was the lack of objective evaluation of the proposed methods. In most of the cases, the testing ground was literary works of unknown or disputed authorship (e.g., the Federalist case), so the estimation of attribution accuracy was not even possible. The main methodological limitations of that period concerning the evaluation procedure were the following:

- The textual data were too long (usually including entire books) and probably not stylistically homogeneous.
- The number of candidate authors was too small (usually 2 or 3).
- The evaluation corpora were not controlled for topic.
- The evaluation of the proposed methods was mainly intuitive (usually based on subjective visual inspection of scatterplots).
- The comparison of different methods was difficult due to lack of suitable benchmark data

Since the late 1990s, things have changed in authorship attribution studies. The vast amount of electronic texts available through Internet media (emails, blogs, online forums, etc) increased the need for handling this information efficiently. This fact had a significant impact in scientific areas such as information retrieval, machine learning, and natural language processing (NLP) . The development of these areas affected authorship attribution as described below.

- Information retrieval research developed efficient techniques for representing and classifying large volumes of text.
- Powerful machine learning algorithms became available to handle multi-dimensional and sparse data allowing more expressive representations. Moreover, standard evaluation methodologies have been established to compare different approaches on the same benchmark data.
- NLP research developed tools able to analyze text efficiently and providing new forms of measures for representing the style (e.g., syntax-based features).

More importantly, the plethora of available electronic texts revealed the potential of authorship analysis in various applications in diverse areas including intelligence (e.g., attribution of messages or proclamations to known terrorists, linking different messages by authorship) (Abbasi Chen, 2005)[4], criminal law (e.g., identifying writers of harassing messages, verifying the authenticity of suicide notes) and civil law (e.g., copyright disputes) (Chaski, 2005; Grant, 2007)[5][6], computer forensics (e.g., identifying the authors of source code of malicious software) (Frantzeskou, Stamatatos, Gritzalis, Katsikas, 2006)[7], in addition to the traditional application to literary research (e.g., attributing anonymous or disputed literary works to known authors) (Burrows, 2002; Hoover, 2004a)[8][9]. Emphasis is now given to factors such as the training text size , the number of candidate authors, and the distribution of training texts over the candidate authors.

In the typical authorship attribution problem, a text of unknown authorship is assigned to one candidate author, given a set of candidate authors for whom text samples of undisputed authorship are available. From a machine learning point-of-view, this can be viewed as a multi-class single-label text categorization task. This task is also called authorship (or author) identification usually by researchers with a background in computer science. Several studies focus exclusively on authorship attribution (Stamatatos, Fakotakis, Kokkinakis, 2001) [10], while others use it as just another testing ground for text categorization methodologies (Zhang Lee 2006) [11]. Beyond this problem, several other authorship analysis tasks can be defined, including the following:

- Author verification (i.e., to decide whether a given text was written by a certain author or not).
- Plagiarism detection (i.e., finding similarities between two texts).
- Author profiling or characterization (i.e., extracting information about the age, education, sex, etc. of the author of a given text).
- Detection of stylistic inconsistencies (as may happen in collaborative writing).

3 Network Architectures used

3.1 Artificial Neural Networks

Artificial neural networks (ANN) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains. The neural network itself is not an algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data inputs. The original goal of the ANN approach was to solve problems in the same way that a human brain would. Artificial neural networks have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games and medical diagnosis.

With the meteoric rise in computing power and available data, the extensive use of neural network with an increase in the number of hidden layers and neurons to learn complex functions resulted in a wave of *deep learning*, which has

resulted in improved State-of-the-Art algorithms across the charts[12], especially with tasks that require extensive training datapoints such as image recogition and speech classification algorithms. In deep learning, each level learns to transform its input data into a slightly more abstract and composite representation, thereby leading to increasingly complicated feature maps which then can be used to accurately determine required outputs, such as presence of a particular feature or comparison and clustering of features in a lower dimensional space.

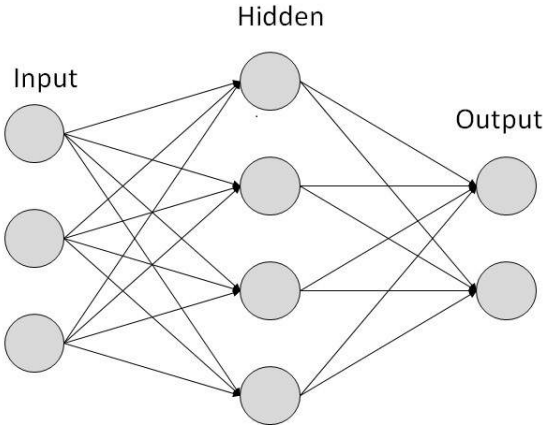

Figure 1: ANN with one hidden layer

3.2 Siamese Networks

Siamese Networks refer to twin neural networks designed to learn relationships between feature encodings in a low dimensional space. In a Siamese Convolutional Network, these encodings are generated by Convolutional Neural Networks (CNNs) and are then compared using a distance metric to check for similarity. Siamese neural networks were used in 1994 by Bromley et al. (1994) [13] to verify written signatures and have hence been extensively used for image verification [14][15][16]. The architecture of a Siamese Network is given below.

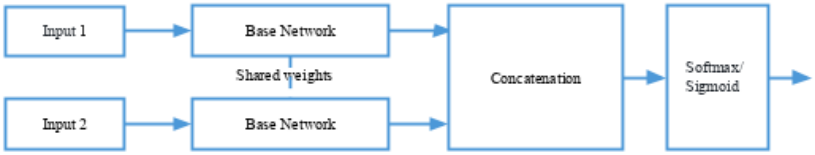

Figure 2: A generalized Siamese Network with weight sharing

3.3 Convolutional Neural Networks

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery. CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. This independence from prior knowledge and human effort in feature design is a major advantage. A convolutional neural network consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutional layers, activation function, pooling layers, fully connected layers and normalization layers.

The primary reason for employment of CNNs in computer vision problems is their ability to share parameters. This parameter sharing stems from the fact that the filters used in CNNs are learned matrices that are convolved with the image by combining local activations of multiple points in the feature maps. This results in a significant reduction in parameters and more importantly gives CNNs the attribute of spatial invariance, which ensures that spatial perturbations in the input do not affect the final output.

Shown below is the architecture of Oxford's VGG 19 network[17] trained on the ImageNet dataset[18], which is a 1000 class image classification problem.

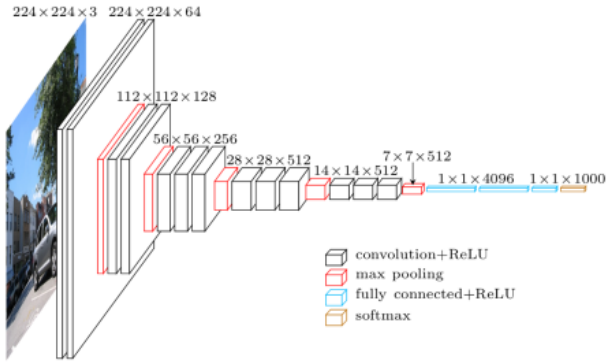

Figure 3: VGG-19 architecture

We aim to use the properties of spatial invariance provided by CNNs to the problem of authorship attribution to develop a robust system designed to capture authorial signatures while also developing resistance to trivial changes in lexical orientation. Previous work in this area has attempted to capture word order patterns[19] and character level similarities[20].

4 Preprocessing employed

Each text was scanned and a vocabulary of all words in the text was generated. All the words in the vocabulary were then replaced by their corresponding word embeddings before feeding into the network. The word embeddings used in this project are GloVe word embeddings by Stanford NLP Group.

5 Experimentation

Model training and evaluation for all three proposed architectures were carried out on a CPU hosted by Google Colab. The dataset used was the Stanford Ad-hoc author attribution conpetition (`http://www.mathcs.duq.edu/~juola/problems/problemD/index.html`), consisting of a corpus with 11 texts from 3 different authors and 4 test examples.

5.1 Artificial Neural Network

The embeddings obtained using the methods discussed in Section 4 were fed into a two hidden layer neural network to perform a 3-class classification to determine the author of the embedded text. The training and validation results hence obtained are displayed below.

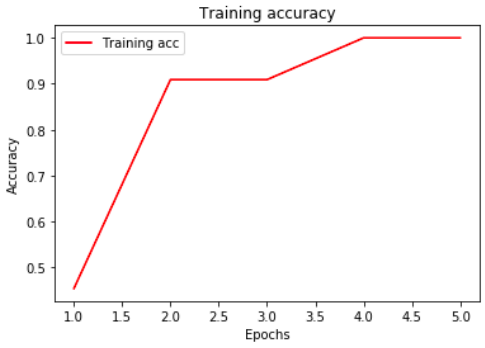

Figure 4: Training accuracy for the Neural Network

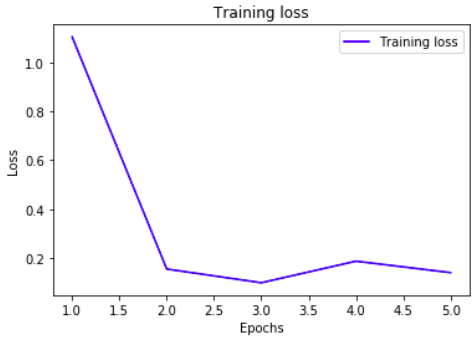

Figure 5: Training loss for the Neural Network

5.2 Convolutional Neural Network

The embeddings obtained in Section 4 are passed into a network comprising of a 1D Convolutional layer with 5 filters, imputed into a Fully Connected layer and passed to the output layer activated with softmax. The training and validation results hence obtained are displayed below.

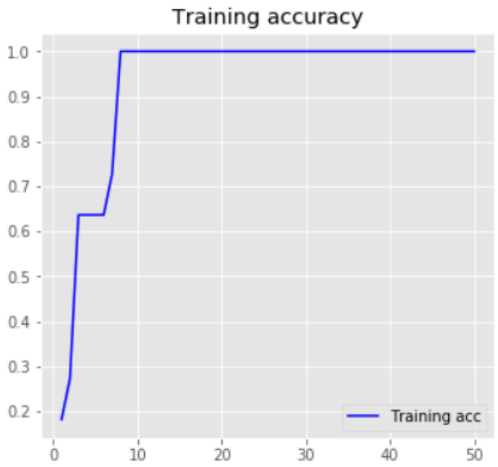

Figure 6: Training accuracy for the Convolutional Neural Network

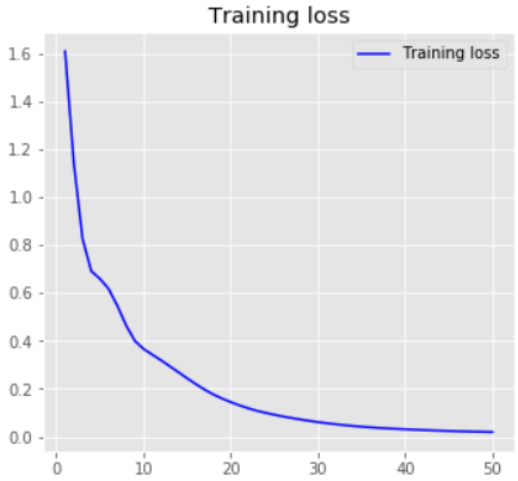

Figure 7: Training loss for the Convolutional Network

5.3 Siamese Neural Network

The training and validation dataset were split according to the algorithm proposed in Section 4. The base networks comprise of neural networks with one hidden layer. The lower dimensional outputs are then passed into a concatenation layer that subtracts these feature maps hence producing an interpretable difference function that is then passed to the final sigmoid unit that learns to predict if the texts used as input were from the same or different authors. The training and validation results hence obtained are displayed below.

Figure 8: Training accuracy for the Siamese Network

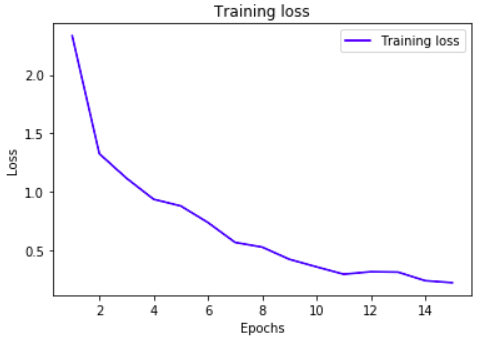

Figure 9: Training loss for the Siamese Network

Method used	**Training accuracy**	**Validation accuracy**	**F1 score**
Neural Network	100.00	50.00	0.33
Convolutional Neural Network	100.0	75.0	0.56
Siamese Neural Network	90.00	81.25	0.66

6 Conclusion and Future Scope

In this report, we present three methods to assign authors to texts and compare their results. While the neural network and the CNN show promising results on the dataset, the dearth of training and validation datapoints limit the scalability of such architectures.

On the other hand, a differential style of interpreting textual information like the Siamese network leads to an architecture where we can compare the test example with the classes already trained on to ascertain the difference between them and hence determine the author to reasonable confidence.

Another plausible architecture that could learn to generalise well in this setting would be a Siamese Convolutional Network (Siamese Network with Convolutional

base networks) that would calculate the distance function specified in Section 5 along with spatial invariance properties. The output of the networks proposed in this report could be used for generative modelling i.e a setting where we could generate textual data containing an authors signature. This could prove to be useful in few-shot learning settings where generative modelling could be used to generate new datapoints to train and test on.

As discussed in Section 4, a major drawback of authorship assignment corpora is the lack of training data to classify new documents. However, we may use these networks to circumvent the problem by performing unsupervised learning in the form of clustering the texts and hence extracting meaningful information about authorship.

References

[1] Frederick Mosteller and David L Wallace. *Applied Bayesian and classical inference: the case of the Federalist papers*. Springer Science & Business Media, 2012.

[2] DAVID I. HOLMES. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 09 1998.

[3] Joseph Rudman. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4):351–365, 1997.

[4] Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 9 2005.

[5] Carole Chaski. Who's at the keyboard? authorship attribution in digital evidence investigations. *IJDE*, 4, 01 2005.

[6] Tim Grant. Quantifying evidence in forensic authorship analysis. *International Journal of Speech Language and the Law*, 14, 09 2007.

[7] Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas. Effective identification of source code authors using byte-level information. In *Proceedings of the 28th International Conference on Software Engineering*, ICSE '06, pages 893–896, New York, NY, USA, 2006. ACM.

[8] Burrows J. F. 'delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17:267–287, 09 2002.

[9] D L. Hoover. Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing*, 17:157–180, 06 2002.

[10] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214, May 2001.

[11] X. FRANK ZHANG. Information uncertainty and stock returns. *The Journal of Finance*, 61(1):105–137.

[12] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

[13] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature Verification Using a "Siamese" Time Delay Neural Network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, pages 737–744, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc. event-place: Denver, Colorado.

[14] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1, June 2005.

[15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015. arXiv: 1503.03832.

[16] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, June 2014.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] Rie Johnson and Tong Zhang. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*, 2014.

[20] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.