A literature review of offline text-independent writer verification and identification by learning the global feature vectors via triplet CNN

Davit Kvartskhava
dkvart17@earlham.edu
Earlham College
Richmond, Indiana

1 INTRODUCTION

“Handwriting is a kind of behavioral biometrics [12].” Every person has a somewhat unique handwriting style, which makes it possible to verify or identify a person based on their handwriting [1]. Manual forensic handwriting analysis is used by law enforcement agencies to identify the writer, and it plays a huge role in some investigations [6]. However, identifying a writer based solely on their handwriting requires a lot of human expertise and experience in addition to being very time-consuming. Hence, automating this process is a research topic of interest. The emergence of Convolutional Neural Networks has brought hope that machines will surpass the baseline set by human experts. The research on automating writer identification methods has also become relevant to analyzing historical documents as more digitized data is now available.

Signature verification could be viewed as another specific application of writer identification problem. However, in the case of signature verification, the problem space is different, as the main concentration is on distinguishing between forged and genuine signatures [5]. It has to be noted that because we do not have a similar handwriting database, a limitation of this study is that it will most likely fail in case of a skilled forgery.

The research in the area of writer identification is usually divided into two sub-categories – on-line and off-line writer identification. In on-line writer identification, the dynamic information about the procedure of writing is preserved using specialized devices. In off-line writer identification, such information is not available and the only source of input is the handwritten text itself.

The approaches for solving the problem of writer identification can also be divided into two categories – text-independent and text-dependent methods. The text-dependent method requires the input to contain the same text as the target handwriting (or at least the same set of characters), while the text-independent method tries to solve the problem regardless of the content of handwriting.

In the last decade, Convolutional Neural Networks have become a popular choice for analyzing the visual documents [7]. The groundbreaking work on object detection, OCR, face verification and many other successful applications of CNNs have revolutionized the field [8]. CNNs have also been successfully used in the writer identification problem [3, 11, 12]. Such neural networks have set the state-of-the-art baseline in terms of accuracy of identifying the writers based on their handwriting [3, 11, 12].

However, the use of CNNs is not a simple recipe for guaranteed success - different preprocessing steps, the type of the input, the loss function and different formulation of the same problem often leads to different results. In this paper, I will review the methods that have been used in the last decade to tackle the problem of writer identification with Convolutional Neural Networks as well as present the topic of my research below.

This literature review concentrates on off-line text-independent writer identification using CNNs. First, I will present the idea for my research. Then I will review the methods that have been used in the last decade and finally, I will conclude by pointing out the main trends in this field and the future of such research.

2 MY RESEARCH IDEA

My approach is to train CNN to directly optimize the embeddings of the images of handwritten text. I will be using the method of unified embeddings that has been successfully used in face recognition and other visual analysis tasks [9]. Similar research has already been done by Keglevich et al. [6], but I will combine the previous research ideas that I believe will improve the accuracy. The motivation for my approach is that the retrieval of local features from the small patches of the images of handwritten text leads to the loss of information that might be key to identify the author with high accuracy. The aggregation methods that combine local descriptors to the global ones are not perfect and the spatial information of the location of the patches is lost. Instead, I will be feeding a CNN with larger patches of the images of handwriting. The downside of this method is that it requires more data for the CNN to become accurate, so I’m also employing data augmentation technique inspired by Tang and Wu [11].

3 METHODS

This section describes different methods that have been used to solve writer identification problem using CNNs. Two main methods are pointed out below: (1) directly training a CNN to classify the handwriting samples and (2) learning the feature vectors via CNN. In addition to that, section 3.3 reviews the methods that have been used to address the lack of training data.

3.1 Classification task

Convolutional Neural Networks have been used in two distinct ways to identify writers based on their handwriting. The first approach treats the problem as a classification task, and the CNN is trained through the softmax loss function where the number of output nodes correspond to the number of users in the database.

Conference’17, July 2017, Washington, DC, USA
2020. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The output of each node signifies the probability that each user is the author of the handwriting, given the handwritten text. The shortcoming of this approach is that it is not scalable and the network needs to be retrained every time a new writer is registered in the database.

Xing and Qiao [12] have taken the approach mentioned above of directly training a classifier. Such a CNN outputs a vector of probabilities of a handwriting sample belonging to a specific writer in the database. Xing and Qiao extracted the patches from the lines of handwritten text. They used a specific architecture (multi-stream structure) of a neural network that comprised of two dependent CNNs that share the features in some layers. The reason for using such architecture was to take advantage of the spatial relationship between different square patches. The input for this network was a pair of two adjacent patches. They conducted their experiments on IAM and HWDB1.1 datasets. They achieved the maximum accuracy of 99.01% on IAM and 93.85% on HWDB1.1.

3.2 Methods for obtaining encodings

Another method for learning to identify writers is to produce the feature vectors or encodings associated with each input image. This approach deals with the issue of scalability of the basic classifiers. The encodings are supposed to capture the unique features of the handwriting, so that the encodings themselves are enough to differentiate between two writers. This way, a feature vector can be produced for the images of handwritten text whose author was not in the training dataset. After the feature vectors have been generated, the task of identification becomes trivial. All that is left to do is to find another handwriting from the labeled examples so that some metric of similarity between the encodings is minimized.

3.2.1 *Encodings produced through classification.*

There are different methods for obtaining the encodings. A more outdated approach starts out by training a CNN with a classification layer with a task to learn to classify the handwriting samples into the writer classes [9, 10]. The second step is to extract the penultimate layer of the network. This second to the last layer of the network allegedly contains the features specific enough to a writer so that different feature vectors can be used to distinguish between different handwriting samples [10].

Fiel and Sablatnig [4] were first to propose to extract feature vectors from handwriting for the writer identification task. They trained a CNN on a classification task and then cut off the last layer of the network to get a network that outputs a feature vector. They also used data augmentation techniques to enlarge the database. The specific technique that they used comprised of tilting the patches obtained by sliding window method. An encoding for an entire image of handwriting was obtained by averaging the encodings generated by the small patches. These encodings were later compared using 2 distance. They conducted the experiments on three datasets:the ICDAR 2011 and 2013 writer identification contests, and CVL datasets.

Christlein and Maier [3] took a similar approach to extract local feature vectors - they took the penultimate layer of a CNN as an encoding. For identification, they used cosine distance between global descriptors. They combined local feature vectors using different algorithms in order to produce a global descriptor for each handwriting sample. They compared the VLAD encoding to triangulation embedding. They also compared max pooling to sum pooling in the writer identification task. The input for the CNN was the small 32x32 patches that were randomly drawn from inside of the contours of the handwriting image.

3.2.2 *Directly optimized encodings.*

Another method for obtaining encodings was devised in 2015 [9], and it improved the benchmark on face recognition/verification task. This method was also applied to writer identification problem but it did not improve the baseline set by other approaches. Below, I will review both of these papers.

Schroff et al. [9] published a paper in 2015 on learning unified embeddings for face recognition. The method that they proposed produced an algorithm with 30% less error rate than other known approaches. They started training a CNN with the direct aim to optimize the encodings themselves, instead of treating the problem as a classification task. They mention that the downside of the older approach "are its indirectness and its inefficiency". The algorithm starts by picking three examples from the data - anchor, positive example and negative example. Then the triplet loss function is used to maximize the distance between the encodings of anchor and negative example, while at the same time minimizing the distance between anchor and positive example. This way the network learns how to encode the images so that the resulting feature vector accurately represents the unique features of a face specific to different individuals. In this paper, they also talk about the importance of choosing the best triplets for training and propose a specific algorithm for choosing such triplets.

This recent version of obtaining the encodings was used in the research for writer identification led by Manuel Keglevic [6]. Again, the objective was to learn the encodings of the handwriting samples where the square distance (L2 measure) between encodings obtained from two different classes is maximized and the same measurement for the identical classes is minimized. In this paper they incorporated an interesting algorithm for extracting the patches. They retrieved the patches around the SIFT keypoints. As they claim, based on previous research, SIFT points are such that there is enough information around them for the network to learn useful encodings. After feeding the CNN with these patches, they aggregated the vectors from different patches into one encoding. For this process of creating one feature vector per entire image of handwriting, they use VLAD encodings. This approach was tested on ICDAR 13 database and the authors report near the-state-of-the-art results.

3.3 Methods addressing the lack of data

Tang and Wu [11] proposed a novel data augmentation technique because of the necessity of large amounts of data to train a CNN. For the feature vector retrieval, they used the method of training with classification objective and extracting the last layer. They also proposed the use of joint Bayesian technique instead of square distance for the identification task. All the previous research that has been done in this area was concentrated on training the CNN on small image patches; however, the problem of this approach, as the authors of this paper point out, is that local features extracted from patches don't contain enough information about a person's writing style. However, learning the global features requires a lot more data, so they first extracted the words from the images of handwritten texts and then randomly permuted each word in a line. As a result, they were able to accumulate thousands of handwriting images for each writer in the dataset. They reported the best results on CVL dataset and near state-of-the-art on ICDAR 13.

Chen et al. [2] also pointed out that CNNs need a lot of training data to achieve the satisfying accuracy in real world applications. The data augmentation techniques do generate more data but the downside of using such techniques is the risk of overfitting to the repeated data. Instead they proposed a semi-supervised feature deep learning algorithm that learns to extract the features of the writing style from the mixture of labeled and unlabeled data. The patches are extracted from the original images and VLAD encodings are used to produce global descriptors from the local feature vectors. ResNet-50 is used a baseline with WLSR method for regularization.

4 CONCLUSION

Writer identification has historically been done by human experts. The automation of this task has attracted many computer scientists because of its impact on crime investigation. A new wave of interest in this research area has been brought about by the recent success of CNNs in the visual task analysis. However, neural networks need a vast amount of labeled data in order to be applicable for solving this problem in real-world settings [2, 11]. In the absence of such data, the researchers have been incorporating the recent refinements in deep learning algorithms to achieve new baselines in the writer identification task. The methods that have involved CNNs either focused on directly learning handwriting to writer mapping through classifiers, or learning the feature vectors that described the style of writing. The latter has proved to be more useful as it is a more scalable approach. In addition to that, there are numerous ways to extract global feature vectors from a handwriting image. Some researchers have tried to train the network on extracting the local features and then combined them with different aggregation methods, while others took an end-to-end approach of learning the global feature vectors directly from the image of handwriting. Triplet CNNs have emerged in 2015 and improved the CNN's ability to learn useful features. To my knowledge, there is no published research on using the Triplet CNNs to directly learn the global descriptors. Such an approach requires a large amount of data, so the data augmentation methods have to be incorporated.

ACKNOWLEDGMENTS

Thanks to me, myself and I.

REFERENCES

[1] Marius Bulacu and Lambert Schomaker. 2007. Text-independent writer identification and verification using textural and allographic features. *IEEE transactions on pattern analysis and machine intelligence* 29, 4 (2007), 701–717.

[2] Shiming Chen, Yisong Wang, Chin-Teng Lin, Weiping Ding, and Zehong Cao. 2019. Semi-supervised feature learning for improving writer identification. *Information Sciences* 482 (2019), 156–170.

[3] Vincent Christlein and Andreas Maier. 2018. Encoding CNN activations for writer recognition. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 169–174.

[4] Stefan Fiel and Robert Sablatnig. 2015. Writer identification and retrieval using a convolutional neural network. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 26–37.

[5] Luiz G Hafemann, Robert Sabourin, and Luiz S Oliveira. 2017. Learning features for offline handwritten signature verification using deep convolutional neural networks. *Pattern Recognition* 70 (2017), 163–176.

[6] Manuel Keglevic, Stefan Fiel, and Robert Sablatnig. 2018. Learning features for writer retrieval and identification using triplet CNNs. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 211–216.

[7] YD Li, ZB Hao, and Hang Lei. 2016. Survey of convolutional neural network. *Journal of Computer Applications* 36, 9 (2016), 2508–2515.

[8] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. 2017. A survey of deep neural network architectures and their applications. *Neurocomputing* 234 (2017), 11–26.

[9] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[10] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2015. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2892–2900.

[11] Youbao Tang and Xiangqian Wu. 2016. Text-independent writer identification via CNN features and joint Bayesian. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 566–571.

[12] Linjie Xing and Yu Qiao. 2016. Deepwriter: A multi-stream deep CNN for text-independent writer identification. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 584–589.