USING DEEP NEURAL NETWORKS IN WRITER IDENTIFICATION & ANALYSIS

Aditya A. Majithia,* **Arthur Paul Pedersen & Michael D. Grossberg**
CITY UNIVERSITY OF NEW YORK
amajith000@citymail.cuny.edu {apedersen,grossberg}@cs.ccny.cuny.edu

ABSTRACT

Historical writer identification is a challenging task. Documents of interest are typically written by authors who are no longer alive, posing a serious challenge to verification. Document digitization has created opportunities for developing automated methods that lower costs, improve time efficiency, reduce variability, and increase accuracy. While the preponderance of work on automated writer identification for modern and historical handwritten documents has focused on entire document pages, the integrity of historical documents encountered in practice is frequently compromised in multivarious ways, imposing a significant limitation on automated methods that require for their operation entire document pages. This paper focuses on developing automated methods for historical writer identification that are capable of drawing on individual words. Previous studies have used convolutional and recurrent neural network models to achieve significant accuracy in identifying authors from modern handwritten word images. Recent advancements in deep learning indicate that transformer models outperform CNNs and RNNs in vision tasks. Introduced in this paper is CONVOLUTIONAL TRANSFORMER ENCODER (CTE), a deep learning model designed to use individual words for writer identification. We examine the effectiveness of transformers for identifying authors from modern handwritten words and then apply the findings to classify historical authors based solely on individual words.

1 INTRODUCTION

Author identification from historical handwritten manuscripts poses a critical challenge for forensic analysts and historians. Traditionally, author identification has been performed by experts, which is expensive, time-consuming, and largely unreliable. While digitization has opened the door for developing automated methods that study handwriting, technologies such as optical character recognition (OCR) that are built into common software products like Adobe Acrobat fail to answer to the unique challenges posed by handwritten content. For handwriting recognition, software like Transkribus have been developed (see, e.g., Memon et al. (2020)). Notwithstanding these technological strides, the problem of author identification from handwritten content is host to its own distinct goals and requirements.

Prior work on automated writer identification has confined itself to handwritten signatures in determining authorship (Bilski et al., 2023). This research presumes signatures to be unique identifiers for its signatories, imposing a significant restriction on the scope of evidence admitted for resolving the question of authorship. Yet nothing in principle prevents automated methods from being applied to analyze handwriting across environments or conditions in order to uncover clues, sometimes subtle or subtextual, of demonstrable probative value in determining authorship or discerning characteristics attributable to writers, including schooling, birthplace, gender, age, and emotional disposition. Such automated methods, should they prove to be practicable, would lay the groundwork for significant advancements in historical inquiry and forensic science as well as confer unprecedented means for preserving cultural heritage and achieving historical justice.

*Corresponding author.

While prior research has tried to identify writers on the basis statistical features of handwriting extracted using machine learning classifiers (Rehman et al., 2019), modern approaches apply deep learning methods capable of automatically learning and extracting features from raw pixels. This paper adapts a modern approach to study the problem of writer identification. In contrast with prior work, we study methods for extracting features of handwriting from individual words. We introduce CONVOLUTIONAL TRANSFORMER ENCODER (CTE), a model inspired by a high-performing design for the CERUG-EN dataset (He & Schomaker, 2021). We show that CTE surpasses state-of-the-art results, achieving 89.7% accuracy on the CERUG-EN test set, significantly outperforming the 1% random chance baseline.

For historical writer identification, we used the International Conference on Document Analysis and Recognition (ICDAR) 2017 competition dataset (Fiel et al., 2017). This dataset contains 3,600 manuscript pages from 720 authors, spanning from the 13th to the 20th centuries. Using a pre-trained neural network, we extracted words from manuscript images to create a dataset of 699 authors. Our model achieved a validation set accuracy of 50%, significantly outperforming the 0.1% random chance baseline in spite of falling short of the 89.7% accuracy achieved on the modern handwritten words dataset.

The paper is organized as follows: after a brief background on related work, we review the datasets used and the methods applied. We then evaluate the effectiveness of these methods, followed by our conclusions.

An Appendix reviews alternative techniques that we tested.

2 BACKGROUND

Previous research has gained significant accuracy on the ICDAR 2017 Historical Writer Identification dataset. Christlein et al. (2017) approach writer identification for the ICDAR 2017 competition using SIFT keypoints for sampling 32×32 image patches from the entire document images as depicted in Figure 1. These patches were grouped to obtain a new set of surrogate classes that represent a set of visual attributes of the different patches. By training a deep residual network (ResNet) using these surrogate classes, the model learnt to associate unique combinations of visual attribute sets with specific writers. In addition, VLAD encoding and Exemplar Support Vector Machines were used for further refining feature representations. The model achieved an accuracy of 88.9%.

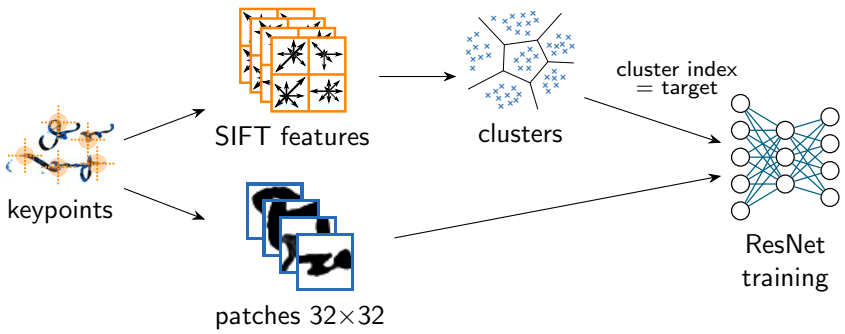

Figure 1 Overview of unsupervised feature learning (Christlein et al., 2017). SIFT descriptors and image patches are extracted at SIFT keypoint locations. Clustered SIFT descriptors and the corresponding patches are used as inputs for the ResNet model.

Lai et al. (2020) utilize the ICDAR 2017 writer identification dataset where whole document images were used to extract pathlet features to characterize handwriting contours beyond slant and curvature, along with unidirectional SIFT features for describing corner and junctions. These features were encoded using a bagged VLAD method and cosine similarity was used to perform writer identification. This methodology achieved an accuracy of 90.1%.

He & Schomaker (2021) subsequently developed a deep neural network architecture to perform writer identification based on word or text block images which approximately contain one word. The CERUG-EN dataset (He et al., 2015) was used which consists of online handwritten word images.

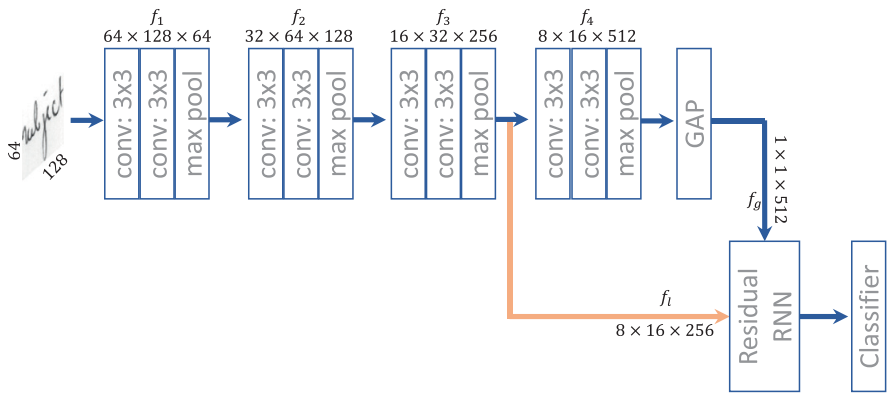

Figure 2 Illustration of GR-RNN model (He & Schomaker, 2021). It contains four blocks f_1, f_2, f_3 and f_4. Each block contains two convolutional layers and one max-pooling layer. f_g is the global context extracted by the global average pooling (GAP) and f_l is the local feature tensor extracted after the third block f_3. The size format of each tensor is denoted by height × width × depth.

As illustrated in Figure 2, the network architecture comprises four blocks, each containing two convolutional layers and one max-pooling layer. The input image size was set to 64 × 128 × 1, where 64 is the height, 128 is the width and 1 is the number of channels. Each convolutional layer was followed by batch normalization and a ReLU activation function. Max-pooling layers were used to reduce spatial resolution. At the end of the convolutional layers, global average pooling was applied to extract global context features. Additionally, local feature maps were extracted after the max-pooling layer of the third block as these maps contain high-level abstract features with good spatial resolution. The local feature map interacts with the global context in a residual recurrent neural network (RNN) block. Finally, a fully connected layer with a softmax activation function was used for writer identification. This architecture achieved an accuracy of 83.2%. The architecture of CTE is inspired by Sheng He and Lambert Schomaker's Global-context residual recurrent neural network (GR-RNN) architecture.

3 DATA

The CERUG dataset comprises handwritten documents from 105 Chinese subjects, primarily students from China, with some residing in China and others studying in the Netherlands. Each participant was asked to produce four different A4 pages. Page 1 required participants to copy a text consisting of two paragraphs in Chinese. On page 2, the subjects wrote about topics of their choice in Chinese. Page 3 featured English text copied from two paragraphs, divided into two sub-pages, each containing one paragraph. Thus, each writer provided four handwritten samples: two in Chinese and two in English. All documents were scanned at 300 dpi, 8 bits per pixel, and in grayscale. For our analysis, we used only the documents written in English containing about 50 word images per writer thereby creating a balanced dataset.

Another dataset used in our analysis comes from the ICDAR 2017 competition on historical document writer identification. The initial dataset comes from the electronic library of the Universitätsbibliothek Basel and includes 140,000 images, which are publicly available under the Public Domain Mark. This collection features not just document images but also drawings, music scores, photographs, blank pages, envelopes, small pieces of handwritten pages, and technical drawings. The document images predominantly consist of correspondences, along with some notes and books, written in various languages, mainly German and French, with some Arabic handwriting as well.

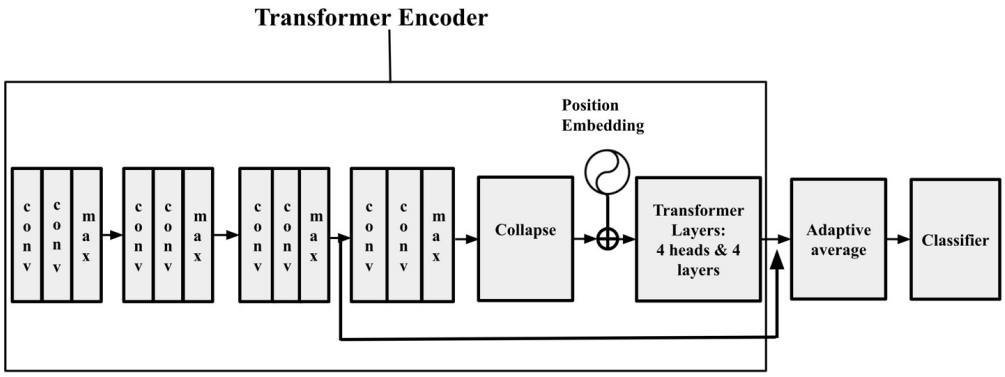

Figure 3 Illustration of CTE. The transformer encoder comprises four convolutional blocks, each containing two convolutional layers and a max pooling layer. The output is flattened in the collapse layer, followed by the addition of position embeddings. The output then passes through the transformer layers, followed by adaptive average pooling and the classifier layers.

The filtering process for the dataset is largely automated. Firstly, the Metadata Encoding and Transmission Standard (METS) files accompanying the images are analyzed to check for author names. Images without an author name are excluded. For authors with known birth and death years, all the dates and names are matched to confirm the writer's identity. The dataset is further refined by filtering images based on text presence. To estimate text regions, several heuristics, such as the distribution of SIFT features, are used.

The final competition dataset comprised 3600 handwritten pages from 720 writers. All images have a quality of 300 dpi and are stored in jpeg format. This dataset provides a solid foundation for developing and evaluating writer identification methods, as it encompasses writing styles, language and document contexts.

4 METHODS

In the ICDAR 2017 competition, the task was to identify authors based on a reasonably large document. Since we are interested in author identification for operational documents like logs, receipts or ledgers, we are going to assume we have small snippets of the text on which we must perform classification, rather than whole documents. Logs, receipts and ledgers could consist of different rows or entries written by different writers and signed by various parties. In such a document, every entry or snippet may be the responsibility of a different person. Hence, analyzing small snippets makes practical sense to identify authors accurately.

Inspired by the GR-RNN model (He & Schomaker, 2021), we developed a model for writer identification on the CERUG-EN dataset, which includes word images of modern handwriting. As shown in Figure 3, we employed a transformer architecture. The initial part of our transformer encoder comprises four convolution blocks based on the GR-RNN model. Hence, we refer to our model as the CONVOLUTIONAL TRANSFORMER ENCODER (CTE). We use a collapse layer to flatten the output from the final convolution block, allowing it to be effectively processed by the transformer layers. We added positional encoding before the transformer layers to maintain the order of the output sequence.

The transformer layers are configured with 256 dimensions, 4 attention heads, and 4 layers. This transformer encoder design is inspired by (Barrere et al., 2022), who used a similar approach for handwritten text recognition. The output from the third convolutional block is horizontally and vertically sliced, followed by concatenation. This processed output is subsequently combined with the output generated by the transformer encoder. Adaptive average pooling and classifier layers are applied following the transformer encoder.

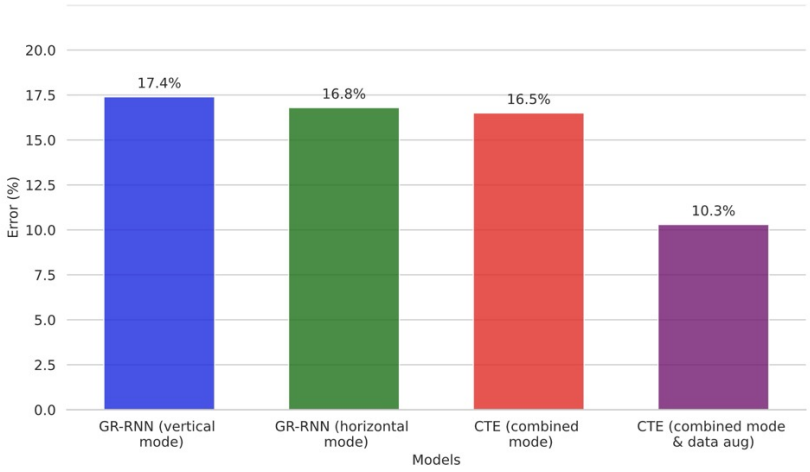

Figure 4 Bar plots (lower is better) of errors for the GR-RNN and CTEs on the CERUG-EN dataset (He et al., 2015). The errors of the GR-RNN model, both in vertical and horizontal modes, and the CTE model without data augmentation, are not significantly different. The error drops significantly to 10.8% when the CTE model is trained with data augmentation.

This architecture has 7 million parameters. The model is implemented using the PyTorch framework. It is trained using the Adam optimizer, with a weight decay of 0.0001 and a batch size of 16. The initial learning rate is set to 0.0001 that gets cut in half after every 10 epochs. We train the model for 70 epochs.

Transformer models are often large, typically containing around 100 million parameters, and handwritten datasets usually have too few training samples for such models to perform well (Kang et al., 2022). To address this, we employed data augmentation to generate additional training samples. We applied elastic and perspective transformations to each image, similar to the methods used by Barrere et al. (2022), effectively tripling the size of our training data.

We also applied CTE to the ICDAR 2017 competition dataset, which includes historical document images. This dataset was divided into two pages per writer, from which we extracted samples for training. We used one page for validation and two pages as a holdout set for testing. We employed a neural network pre-trained on the IAM dataset (Manmatha & Srimal, 1999) of handwritten documents to extract words from the document images. Due to variations in handwriting quality and style, the number of words detected per writer ranged from 5 to 250. For instance, the model detected more words in documents with larger handwriting and fewer words in those with smaller handwriting. Consequently, the dataset was reduced to 699 writers by excluding those for whom the word detection network failed to extract any words. The final dataset used contained 59,667 words for training, 30,057 words for validation, and 63,793 words for testing.

5 EVALUATION

CTE achieves state-of-the-art results with an accuracy of 83.5% without data augmentation. When data augmentation is applied, the accuracy improves to 89.7%, surpassing the state-of-the-art and significantly outperforming the baseline of 1%. Figure 4 shows a chart comparing the errors between the GR-RNN and the CTE models.

Since the classes in the CERUG-EN dataset are balanced, we calculated the macro averages of the precision and recall scores for the CTE model, treating each class equally. Figure 5 presents the False Discovery Rates (FDR) and False Negative Rates (FNR) of the model trained with and without data augmentation. Data augmentation results in lower FDR and FNR, demonstrating improved precision and reliability in identifying writers from modern handwritten words.

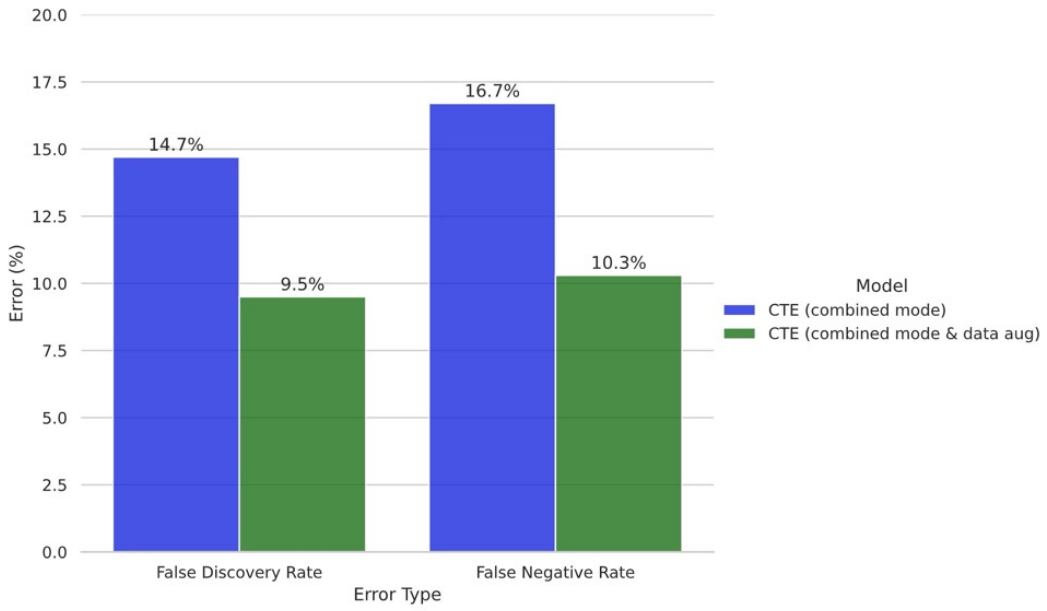

Figure 5 Bar plots (lower is better) of evaluation metrics for the CTE model with and without data augmentation. Data augmentation results in lower False Discovery Rates and False Negative Rates, demonstrating improved precision and reliability in identifying writers from modern handwritten words (CERUG-EN).

Figure 6 and Figure 7 show the loss plots for the CTE models during training. The training curve decreases smoothly, while the testing curve is a bit unstable at first but smooths out towards the end. Both curves plateau after decreasing, showing no signs of overfitting. Hence, our model trains well.

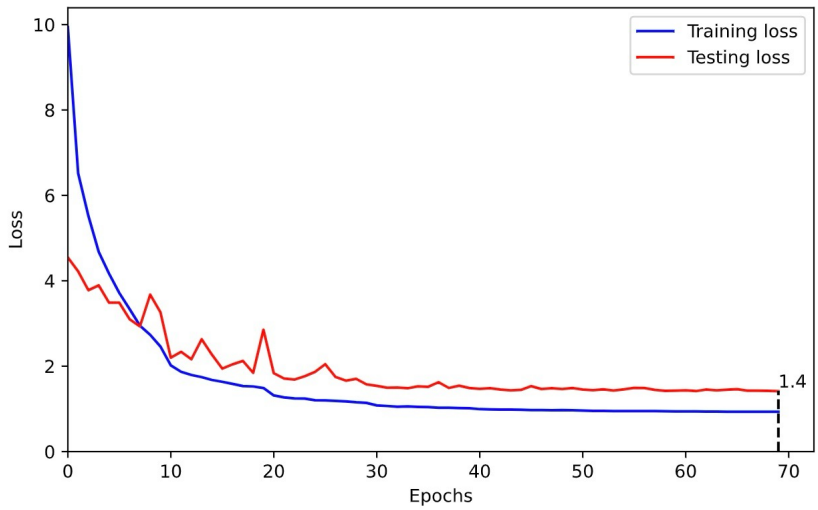

Figure 6 Loss plot of the CTE model without data augmentation. The testing loss is initially unstable but stabilizes towards the end. The curve flattens around 1.4, showing no signs of overfitting.

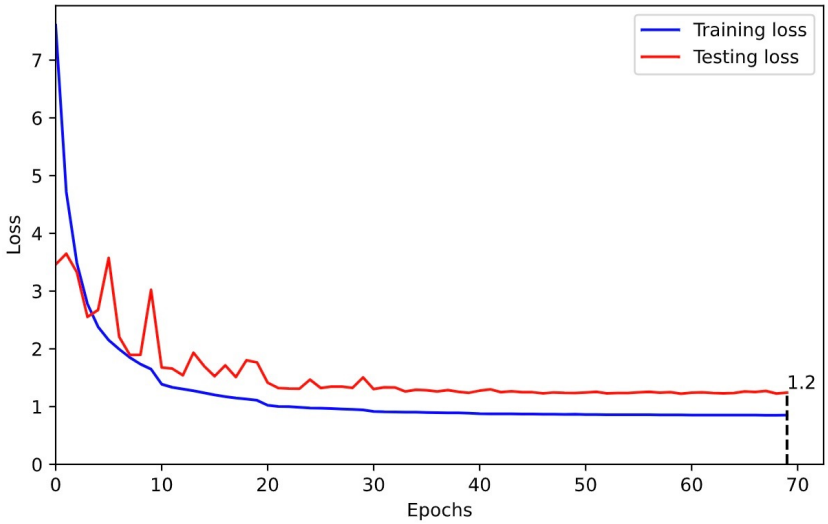

Figure 7 Loss plot of the CTE model with data augmentation. The testing loss is again initially unstable but stabilizes towards the end. The curve flattens around 1.2, showing no signs of overfitting.

The model achieved a 50% accuracy on the validation set of the historical dataset significantly outperforming the 0.1% random chance baseline for 699 classes. This is promising since we achieved this result with minor optimization. However, this performance is much lower than the 89.7% accuracy achieved on the modern handwriting dataset (CERUG-EN) with 105 classes.

The loss plot in Figure 8 shows that the training loss gradually decreases after each epoch and plateaus at 1, the validation loss stays at about 3.4. This discrepancy indicates that the model is learning specific information in the training data that does not generalize well in the validation data.

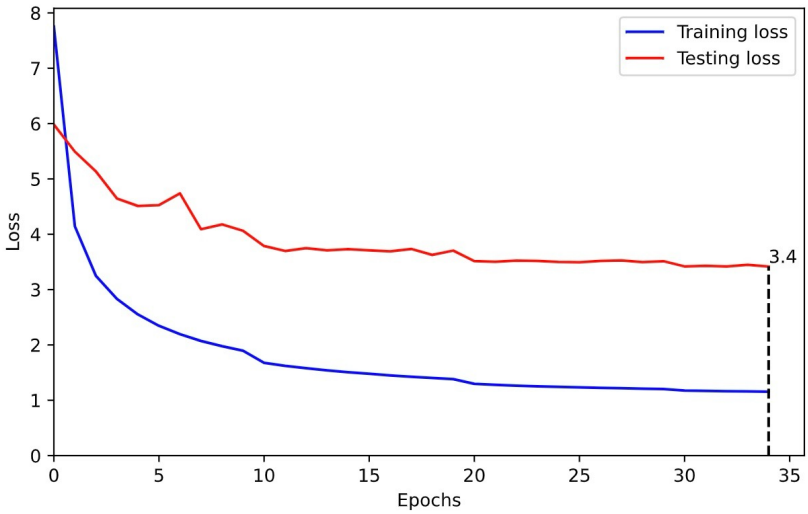

Figure 8 Loss plot of the CTE model on the ICDAR17 historical words dataset (Fiel et al., 2017). The training loss decreases and plateaus around 1, while the validation loss remains at 3.4, indicating that the model learns training-specific information that does not generalize well.

Figure 9 Interpretation of the CTE model for Writer ID 477 from the ICDAR17 dataset (Fiel et al., 2017). The image on the left is from the training set, and the one on the right is from the validation set, written by the same writer. The colormap scale ranges from 0 to 1, showing the weight the model assigned to the pixel in its decision-making. The heat map highlights that the model focused on the shape of "u" during both training and validation to classify the writer.

Figure 10 Interpretation of the CTE model for two different writers from the ICDAR17 dataset (Fiel et al., 2017), one sample from each. The image on the left shows the letter "u" written by Writer ID 477 that the model focuses on, while the image on the right shows the same letter written by Writer ID 488, with a more italicized style.

To gain insight into the model's decision-making process, we examined its focal points during classification. We created heat maps to visualize the pixels that the model considered most important. The colormap scale ranges from 0 to 1, representing the weight the model assigned to the pixel in its decision-making process. The heat map in Figure 9 shows that the model focused on the features of "u" (highlighted in yellow) to classify the writer. This behavior persisted during validation, indicating that the model is focusing on meaningful handwriting features for classification.

Similarly, Figure 10 demonstrates that the model used the shape of "u" to categorize another writer with a different handwriting style (more italicized) for the same letter. This indicates that the model can distinguish various handwriting styles for the same feature, like the shape of "u."

When interpreting the model trained on modern handwritten word dataset (CERUG-EN), we observed a similar focus on handwriting style of specific letter shapes for classification. Figure 11 demonstrates that the model focused on the shape of the letters highlighted in yellow as crucial features for classification during training and testing.

The primary reason for the drop in accuracy with the historical dataset is image quality. Poor quality images cause the model to shift its focus on noise, like the texture of the paper and random ink traces, instead of the handwriting features. This negatively affects performance. Figure 12 shows that the model's focus is shifting to noise unrelated to the handwriting features when making predictions.

6 CONCLUSION

In this paper, we addressed the challenge of identifying historical writers based solely on individual words from the ICDAR17 dataset. We employed the transformer architecture, a state-of-the-art deep learning model, and also evaluated it on the modern handwritten words dataset, CERUG-

Figure 11 Interpretation of the CTE model for Writer ID 8181 from the modern words dataset, CERUG-EN (He and Schomaker 2021) The heat map illustrates the model's focus on specific letter shapes during training and validation for the writer. This behavior of the model is similar to when it is applied to historical words dataset.

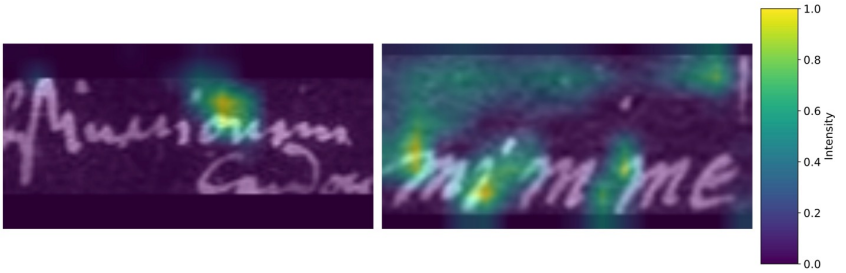

Figure 12 Failure points of the CTE model on the ICDAR17 historical words dataset (Fiel et al. 2017), demonstrating model sensitivity to image quality. The heat map shows the model's focus shifting to pixels above the actual word, that is noise and negatively impacts performance.

EN. Identifying authors from individual words addresses issues related to damaged and fragmented historical documents, as well as operational documents like logs, receipts, and ledgers that may contain small snippets of text from various writers.

We reviewed previous work and demonstrated that the transformer architecture outperforms traditional CNN and RNN architectures for writer identification from modern handwritten words. CTE achieved an accuracy of 83.5% on the CERUG-EN dataset, which is comparable to the accuracy of the state-of-the-art model. With data augmentation, the model surpassed the state-of-the-art, achieving an accuracy of 89.7%, significantly outperforming the baseline of 1%.

When applied to the ICDAR 2017 historical dataset containing only word images, CTE achieved an accuracy of 50% on the validation set, which is significantly higher than the baseline of 0.1%. This is promising since we achieved this result with minor optimization. We also showed that the model effectively focuses on specific letter shapes for classification. However, the poor quality of the historical word images causes the model to focus on noise instead of meaningful handwriting features.

Based on these results, further research should be developed to address challenges such as poor image quality in historical handwritten dataset. Exploring the development of a model with more parameters could also improve performance.

REFERENCES

Killian Barrere, Yann Soullard, Aurélie Lemaitre, and Bertrand Coüasnon. A Light Transformer-Based Architecture for Handwritten Text Recognition. pp. 275–290. 2022. doi: 10.1007/978-3-031-06555-2{_}19.

Piotr Bilski, Jacek Olejnik, and Mieczysław Goc. Automated Verification of the Signatures' Authenticity Using Artificial Intelligence Methods. In *2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, pp. 1252–1257. IEEE, 9 2023. ISBN 979-8-3503-5805-6. doi: 10.1109/IDAACS58523.2023.10348789.

Vincent Christlein, Martin Gropp, Stefan Fiel, and Andreas Maier. Unsupervised Feature Learning for Writer Identification and Writer Retrieval. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 991–997. IEEE, 11 2017. ISBN 978-1-5386-3586-5. doi: 10.1109/ICDAR.2017.165.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

Stefan Fiel, Florian Kleber, Markus Diem, Vincent Christlein, Georgios Louloudis, Stamatopoulos Nikos, and Basilis Gatos. ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1377–1382. IEEE, 11 2017. ISBN 978-1-5386-3586-5. doi: 10.1109/ICDAR.2017.225.

Sheng He and Lambert Schomaker. GR-RNN: Global-context residual recurrent neural networks for writer identification. *Pattern Recognition*, 117:107975, 9 2021. ISSN 00313203. doi: 10.1016/j.patcog.2021.107975.

Sheng He, Marco Wiering, and Lambert Schomaker. Junction detection in handwritten documents and its application to writer identification. *Pattern Recognition*, 48(12):4036–4048, 12 2015. ISSN 00313203. doi: 10.1016/j.patcog.2015.05.022.

Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. Pay attention to what you read: Non-recurrent handwritten text-Line recognition. *Pattern Recognition*, 129:108766, 9 2022. ISSN 00313203. doi: 10.1016/j.patcog.2022.108766.

Songxuan Lai, Yecheng Zhu, and Lianwen Jin. Encoding Pathlet and SIFT Features With Bagged VLAD for Historical Writer Identification. *IEEE Transactions on Information Forensics and Security*, 15:3553–3566, 2020. ISSN 1556-6013. doi: 10.1109/TIFS.2020.2991880.

R. Manmatha and Nitin Srimal. Scale Space Technique for Word Segmentation in Handwritten Documents. pp. 22–33. 1999. doi: 10.1007/3-540-48236-9{_}3.

Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *IEEE Access*, 8:142642–142668, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.3012542.

Arshia Rehman, Saeeda Naz, and Muhammad Imran Razzak. Writer identification using machine learning approaches: a comprehensive review. *Multimedia Tools and Applications*, 78(8):10889–10931, 4 2019. ISSN 1380-7501. doi: 10.1007/s11042-018-6577-1.

Leslie N. Smith. Cyclical Learning Rates for Training Neural Networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472. IEEE, 3 2017. ISBN 978-1-5090-4822-9. doi: 10.1109/WACV.2017.58.

APPENDIX

This appendix reviews techniques we tested but decided not to pursue further because their results were not very promising.

A HYPERPARAMETER OPTIMIZATION OF THE GR-RNN MODEL

We tried tuning the hyperparameters to assess if the performance of the GR-RNN model (He & Schomaker, 2021) could be improved further. The model in horizontal mode, trained with a cyclical learning rate (Smith, 2017), achieved slightly better results on the CERUG-EN test set than the state-of-the-art, reaching an accuracy of 84.9%. Cyclical learning rates help the model avoid local minima and achieve faster convergence. The learning rate fluctuated between $1e-5$ and $1.2e-4$ in a cyclical manner. We determined this optimal range by observing the testing loss with different learning rates. Figure 13 shows the loss plot that demonstrates this cyclical pattern during training and testing.

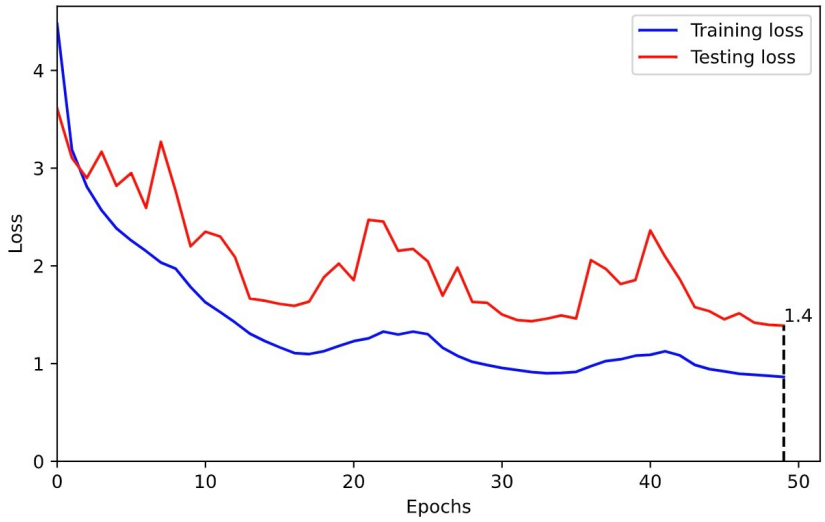

Figure 13 Loss plot of the GR-RNN model trained with Cyclical Learning Rate. The cyclical pattern is observed in both training and testing losses. The technique reduced the loss to 1.4 and achieved an accuracy of 84.9%, showing a slight improvement in performance.

B VISION TRANSFORMER (VIT) ARCHITECTURE

We developed a transformer architecture inspired by the Vision Transformer (Dosovitskiy et al., 2021), using a sliding window approach where tokens are $8x8$ patches of the word image overlapping at the 4th pixel. The model trained very quickly but only achieved an accuracy of 29.7% on the CERUG-EN test set. As shown in Figure 14, the loss curves indicate that the training and testing losses decreased much more smoothly compared to other models, but the accuracy remained low. To improve performance, further preprocessing or feature extraction methods could be explored, as using convolutional blocks, like the model for CTE, has been shown to boost performance.

C REPLACING THE RNN BLOCK OF GR-RNN WITH A TRANSFORMER BLOCK

We also developed a hybrid architecture based on the GR-RNN model, replacing the RNN block with a transformer block featuring a 256-dimensional embedding, 4 multi-head attention layers, and

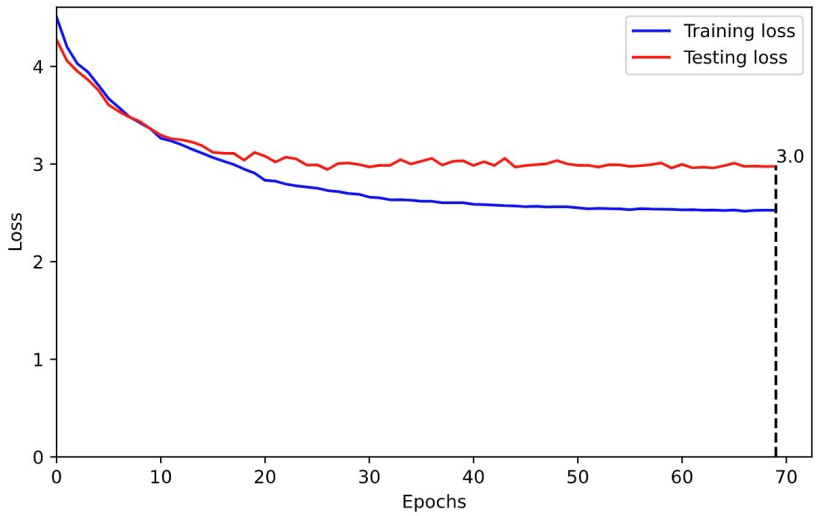

Figure 14 Loss plot of the Vision Transformer model trained on the CERUG-EN dataset (He et al., 2015). The loss curves show that both the training and testing losses decreased smoothly, but the loss remained high at 3.0, indicating that the model still needs significant improvement.

4 transformer layers resulting in 4.5 million parameters. Figure 15 shows the architecture of the hybrid model. It achieved 84% accuracy on the CERUG-EN test set without data augmentation. With data augmentation, the model's accuracy improved to 89.3%, surpassing the state-of-the-art like our other transformer model. However, it trained 3.5 times slower than our CTE model. Therefore, design of CTE with convolutional blocks within the transformer encoder is a better choice as it trains much faster achieving similar accuracy.

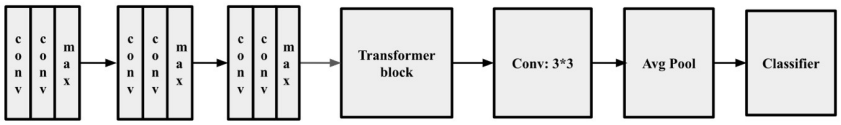

Figure 15 Illustration of the Hybrid architecture with the transformer block replacing the RNN block of the GR-RNN model (He & Schomaker, 2021). The first three convolution blocks contain two convolution layers and a max pooling layer. The output from these convolution blocks is passed to the transformer block, followed by a convolution layer, an average pooling layer, and a classifier layer.

D CLASS BALANCING OF THE HISTORICAL WORDS DATASET

Since the historical word dataset is unbalanced, with the number of word images per writer ranging from 5 to 250, we tried balancing it to see if performance would improve. We used data augmentation techniques like elastic distortion and perspective change with added noise, and applied bootstrapping to the underrepresented classes. We also reduced the dataset to 255 classes with 100 images per class to examine the effect of class balance without data augmentation and bootstrapping. The CTE model still behaved the same during training and only achieved 55% accuracy. This suggests that class imbalance is not the main reason for the lower accuracy.

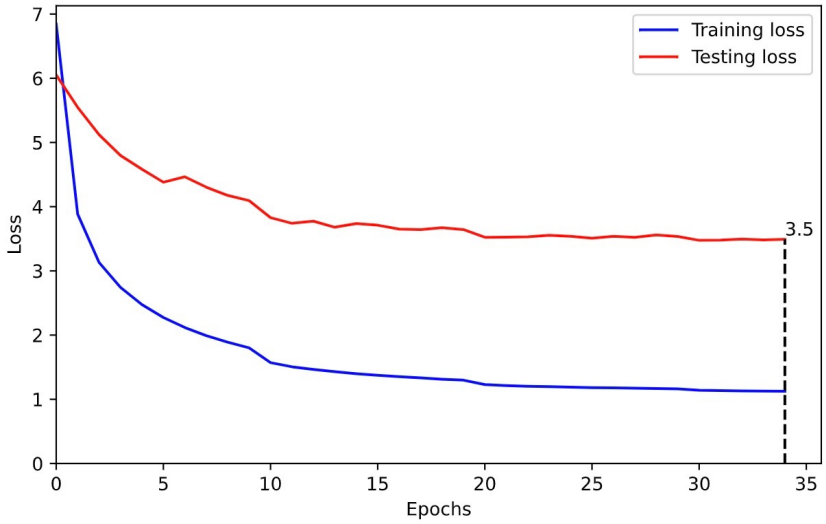

Figure 16 Loss plot of the CTE model after data augmentation and class balancing using the historical words from the ICDAR17 dataset (Fiel et al. 2017). Both training and testing losses follow similar patterns as shown in Figure 7, indicating that the model does not generalize well, and the performance did not improve.