ViTs and More: Exploring Signature Forgery Verification

Aribah Aizaz
KAIST
Daejeon, South Korea
aaribah0709@kaist.ac.kr

Mahnoor Shafiq
KAIST
Daejeon, South Korea
mahnoorshafiq13@kaist.ac.kr

Maida Aizaz
KAIST
Daejeon, South Korea
maidaa25@kaist.ac.kr

Muhammad Faizan Zahid
KAIST
Daejeon, South Korea
faizan.zahyd@kaist.ac.kr

Uzair Ahmed
KAIST
Daejeon, South Korea
uziahmd@kaist.ac.kr

Abstract

Offline signature forgery has been creating challenges for people since the dawn of time, making the detection of such forgery imperative. In recent years, much research has been conducted with countless benchmarks formed. In 2022, students from Zhejiang University created a novel Chinese document offline signature forgery detection benchmark, namely ChiSig. This paper recreates the ChiSig benchmark by improving the verification tasks. Our main improvement is through our usage of a vision transformer (ViT). The ChiSig benchmark was used as the created dataset was available for further experimentation and replication.

1. Introduction

Signature forgery detection, a critical process in both automated and manual realms, aims to verify the authenticity of signatures. This process is essential in combating fraud and validating documents. Signatures hold significant importance in documentation, symbolizing acknowledgment and agreement [1]. In today's digital age, the practice of forging signatures, a serious ethical breach, poses a heightened risk. Forgeries on critical documents like legal contracts or financial transactions can lead to dire consequences, including significant financial losses and legal issues. This escalating concern necessitates ongoing research and development in signature verification techniques.

Our study builds upon the foundational work conducted by students from Zhejiang University [2]. We aim to advance the field of signature verification by employing a novel approach: the integration of a Vision Transformer (ViT). This strategy enhances the existing signature verification methods, offering more robust and efficient solutions for identifying and countering signature forgeries. By adopting and improving upon these established methodologies, our research contributes to the critical task of safeguarding the authenticity and integrity of signatures in our increasingly digitized world.

2. Related Works

2.1. Signature Verification

In the field of biometrics it is often difficult to ascertain whether a pair of signatures are genuine or forged. There are primarily two methods used for this purpose: writer-dependent and writer-independent. [3]

The writer-dependent method, while effective, is limited as it cannot accommodate new users. On the other hand, the writer-independent method offers greater robustness and versatility, making it a focal point of current research. With recent advancements in technology, deep learning has become a prevalent approach in tackling signature verification challenges, especially those building off the writer-independent method. For instance, SigNet employs Siamese convolutional networks to extract features and learn signature embeddings [4]. Another innovative approach is the Inverse Discriminative Network, which utilizes inverse supervision and a multi-path attention mechanism to address the issue of sparse signature information [5]. However, most such approaches do not make their datasets public, thereby making it difficult to replicate the results – hence our choice of the reference paper. Contrary to many existing datasets, the ChiSig

dataset has 10,242 signature images comprising 500 unique names.

2.2. Offline Signature Dataset

There are vast resources publicly available for signature verification, such as CEDAR [6], BHSig260 [7], etc. However, many are limited in terms of the number of signers, as well as the number of signatures.

Innovations in offline signature verification have seen unique approaches, particularly in signature duplication to model spatial intrapersonal variability. Studies like those of Galbally et al. [8] and Ferrer et al. [6] have explored various methods of duplicating signatures, from introducing distortions to employing cognitive models. These methods, while improving performance in some aspects, often did not address the dynamic properties of signatures, leading to a reliance on offline signature verifiers.

3. The Dataset

3.1. Data Acquisition

We have made use of the ChiSig dataset which was readily available for use, making it the primary dataset for our research. The ChiSig dataset is a comprehensive collection of signature images, specifically designed for the study and analysis of signature verification and forgery detection. This dataset encompasses a total of 10,242 signature images, showcasing a diverse rang6

e of signature styles across 500 different signed names.

The naming convention adopted in this dataset is both systematic and informative: each image file is named following the format "name-id-number.jpg". In this scheme, 'name' corresponds to the signed name by the volunteer, 'id' serves as a unique file identifier within the dataset, and 'number' represents the sequence or count of the signature in the dataset.

One of the key features of the ChiSig dataset is its detailed categorization of forgeries. It includes skilled forgeries, identified by an ID number greater than 100. To determine the original signature corresponding to a skilled forgery, one simply subtracts 100 from the ID of the skilled forgery. For instance, if we consider an original signature with the file name "name-100-5.jpg", a skilled forgery of this signature would be named "name-101-5.jpg". This indicates that it is a skilled forgery of the original signature associated with the name "name-1". Additionally, the dataset includes random forgeries, characterized by ID numbers less than 100, such as "name-1-5.jpg". This comprehensive structure makes the ChiSig dataset an invaluable resource for research in signature verification and forgery detection.

3.2. Data Manipulation

In our methodology, we prepared the dataset for optimal processing and analysis. Initially, we divided the dataset into three distinct subsets: 70% for training, 15% for testing, and the remaining 15% for validation purposes. This distribution ensures a comprehensive training of the models while retaining adequate data for robust testing and validation.

To ensure reproducibility and consistency in our experiments, we fixed the random seed at 42. This step is crucial as it guarantees that the splitting of the dataset into training, testing, and validation subsets is deterministic, allowing for consistent results across different runs of the experiment.

Normalization of the images was a key part of our data manipulation process. This technique involves adjusting the pixel intensity values across all images to a common scale. Normalization is vital as it reduces disparities in lighting and contrast between different images, thereby facilitating more accurate and consistent analysis by the machine learning models.

Moreover, we standardized the size of all images in the dataset to 224x224 pixels. This resizing is essential for two reasons: firstly, it ensures that all images fed into the models are of a uniform dimension, which is a prerequisite for many deep learning architectures. Secondly, this uniformity in image size helps in reducing computational complexity and expedites the training process of the models.

4. Experiment

We employ five different embedding methods: ResNet50, InceptionResnet, ResNeXt50, VGG16, and vision transformer (ViT). Once we obtain the embeddings for a pair of signatures, we assess whether they are made by the same individual by calculating the cosine similarity between these two embeddings. This similarity measure helps in estimating the likelihood of both signatures being authored by the same person.

4.1. Evaluation Metrics

To evaluate the effectiveness of our system, we focus on three crucial metrics: Accuracy (Acc), Equal Error Rate (EER), and True Acceptance Rate (TAR) at a specified False Acceptance Rate (FAR) of 0.1% (1e−3). Each of these metrics provides a unique perspective on the system's performance in signature verification.

Accuracy (Acc) This metric gauges the overall precision of our system. It measures the percentage of predictions that are correct, encompassing both true positives (correctly identifying valid signatures) and true negatives (correctly identifying forgeries). A higher accuracy rate indicates a more reliable system in

distinguishing between genuine and forged signatures.

Equal Error Rate (EER) EER is a critical measure in biometric systems, representing the point where the rate of false acceptances (incorrectly identifying a forged signature as authentic) equals the rate of false rejections (incorrectly rejecting a genuine signature). This balance point is a key indicator of the system's overall reliability, as it reflects its ability to equally manage both types of potential errors.

True Acceptance Rate (TAR) at a specific False Acceptance Rate (FAR) of 1e−3 TAR, especially at a low FAR like 0.1%, shows how effectively the system authenticates genuine signatures. It measures the proportion of actual valid signatures that the system correctly identifies as authentic, under the condition that the likelihood of mistakenly accepting a forged signature as genuine (FAR) is set to a stringent threshold of 0.1%. This metric is crucial for assessing the system's ability to accurately verify signatures without being overly permissive in accepting forgeries.

The calculations are as follows:

$$FAR = \frac{Number\ of\ false\ accepted}{Number\ of\ forged} \quad (1)$$

$$FPR = \frac{Number\ of\ false\ rejected}{Number\ of\ genuine} \quad (2)$$

$$TAR = 1 - FPR \quad (3)$$

4.2. Training Loss Function

The training loss function we employed was the better of the two in the reference paper, i.e., cross-entropy loss. Depending on the performance and capability of the embedding model in question, we used either sigmoid cross-entropy loss or softmax cross-entropy loss. Both these activation functions are known to perform well for binary classification problems similar to the one we have at hand. The formula can be aptly described as in Figure 1 below.

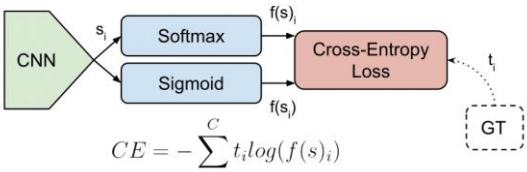

Figure 1: Our implementation of the cross-entropy loss [14]

Here, s_i stands for the logits for class i before passing through the softmax or sigmoid activation function, t_i stands for the true label for class i, and C' stands for the number of classes, which in our case is 2.

4.3. ResNet50

ResNet-50 is a convolutional neural network (CNN) architecture that is a part of the ResNet (Residual Network) family, introduced by He et al [9]. It stands out for its deep network structure of 50 layers, primarily composed of residual blocks. These blocks feature skip connections, or shortcuts, that jump over one or more layers. The primary function of these shortcuts is to address the vanishing gradient problem, allowing the network to be deeper without suffering from training difficulties. ResNet-50 is particularly noted for its efficiency in terms of computational resource usage, while still maintaining high accuracy in various image recognition and classification tasks. This architecture has been widely adopted in the field of deep learning for its effectiveness in training deeper neural networks without a significant increase in the complexity of the model.

4.4. InceptionResnet

InceptionResnet is a fusion of two powerful neural network architectures: the Inception network and ResNets [10]. This architecture combines the strengths of both networks to enhance feature extraction and recognition capabilities, particularly in image processing tasks like signature verification. The InceptionResnet model benefits from the depth and width of the Inception network and the residual connections of ResNets, which help in avoiding the vanishing gradient problem.

For the purpose of signature verification in the ChiSig benchmark, the InceptionResnet model is utilized for its advanced capabilities in handling complex image data. Given the intricate and varied nature of signatures, this model is particularly adept at extracting nuanced features that are crucial for differentiating between genuine and forged signatures.

4.5. ResNeXt50

ResNeXt-50 is a CNN architecture that represents an evolution of the original ResNet design, introduced by Xie et al [11]. It is characterized by its innovative use of "cardinality" - the size of the set of transformations, which is considered a new dimension alongside depth and width in neural network architectures. ResNeXt-50 enhances the ResNet model by incorporating groups of convolutions, allowing it to learn more complex features with a reduced number of parameters. This approach provides an efficient way to increase the accuracy of the network without significantly increasing computational complexity. Known for its balance of efficiency and performance, ResNeXt-50 has become a popular choice for image recognition and classification tasks.

4.6. VGG16

The VGG16 model, designed by Simonyan and Zisserman, is renowned for its depth and efficacy in large-scale image recognition tasks. It consists of 13 convolutional layers and 3 fully-connected layers, making it one of the deeper architectures in image processing and feature extraction [12]. This depth allows VGG16 to capture intricate details and patterns in images, which is crucial in signature verification. In the context of signature verification, VGG16's depth offers a significant advantage. Each layer captures different aspects of the signature, such as stroke curvature, pressure variations, and line thickness. These features are critical in distinguishing between genuine and forged signatures.

4.7. ViT

Vision transformer (ViT) reshapes neural network architectures for image processing. ViT adopts the transformer framework, initially developed for natural language tasks, and applies it to images by treating them as sequences of patches. This innovation enables ViT to capture long-range dependencies and global context in images, making it highly effective in recognizing complex patterns and structures, a crucial aspect of tasks like image recognition [13]. In the domain of signature verification, ViT's ability to comprehend the entire signature as a sequence of patches using self-attention mechanisms holds promise for enhancing accuracy and robustness in verification systems, making it a compelling choice for this application.

Figure 2 below depicts how a ViT performs on signature data.

Figure 2: ViT's conversion of a signature into patches

5. Results

Our results can be summarized as in Tables 1 and 2 below, followed by visualized misclassified samples (out of a total of 1547 test samples) for each of the embedding networks.

Table 1. Results for our embedding models

Model	EER	TAR	Acc
ResNet50	0.3545	0.2649	78.80
InceptionResnet	0.1920	0.1319	85.29
ResNeXt50	0.2178	0.0764	83.08
VGG16	0.4707	0.0193	80.35
ViT	0.4500	0.0001	81.26

Table 2. Misclassification results, out of 1547 samples

Model	Misclassified Samples
ResNet50	328
InceptionResnet	218
ResNeXt50	263
VGG16	304
ViT	273

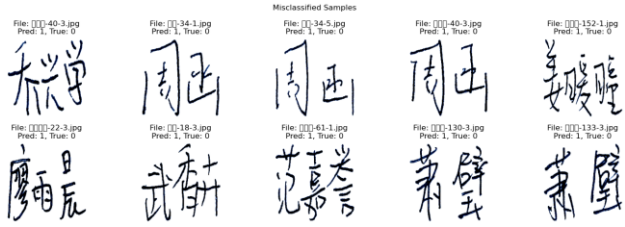

Figure 3: Misclassified samples of ResNet50

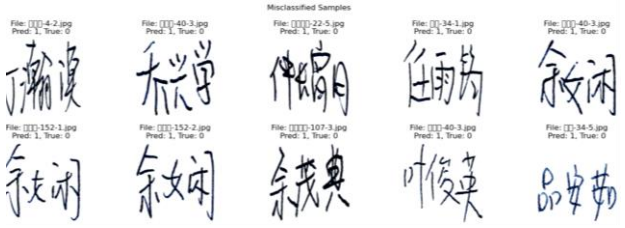

Figure 4: Misclassified samples of InceptionResnet

Figure 5: Misclassified samples of ResNeXt50

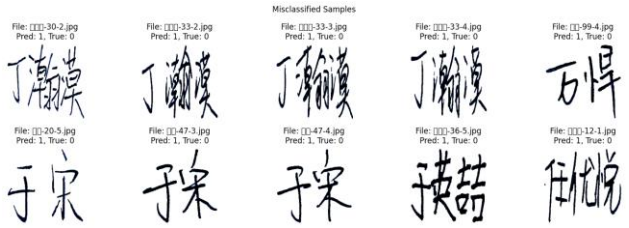

Figure 6: Misclassified samples of VGG16

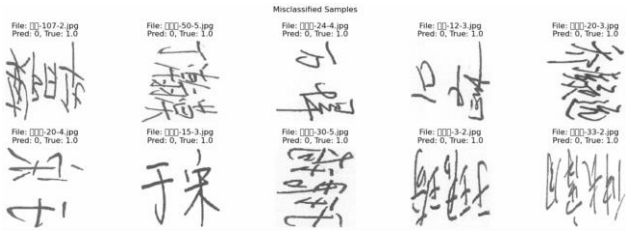

Figure 7: Misclassified samples of ViT

6. Comparative Analysis

In our Comparative Analysis, the deployment of the Vision Transformer (ViT) from scratch represents a substantial advancement in signature verification technology. Our deliberate choice to custom-develop ViT for our dataset allowed us to capitalize on its unique capabilities, particularly in processing images as sequences of patches. This method is a significant departure from conventional CNN-based approaches and offers a fresh perspective in image analysis, especially in discerning intricate patterns in signatures.

When we compare ViT's performance with models like ResNet50, InceptionResnet, and ResNeXt50, its unique strengths become apparent. ViT exhibited notable proficiency in detecting professional forgeries. In terms of numerical performance, ViT achieved an accuracy rate of 81.26%. While this rate may appear marginally lower compared to the 85.29% accuracy of InceptionResnet, it highlights the transformative potential of transformer-based models in complex image recognition tasks.

In error analysis, ViT's Equal Error Rate (EER) was registered at 0.4500, and its True Acceptance Rate (TAR) at a False Acceptance Rate (FAR) of 0.1% (1e−3) was measured at 0.0001. Considering that out of 1547 test samples, ViT misclassified 273, these figures underscore ViT’s capability in enhancing the precision of signature verification systems. This performance is especially noteworthy given the inherent complexities and nuanced variations in human signatures.

The ViT's application in our study marks a notable improvement over the methodologies utilized in the base paper. It introduces an innovative approach to the field of signature verification. The transformer model's ability to understand the global context and long-range dependencies in images suggests a significant potential for its application in not only signature verification but also in other complex image-processing tasks.

Towards the end of our analysis, we observed that the class imbalance in our dataset presented a challenge, slightly skewing the learning process of ViT. While this did impact the overall performance metrics, it's a common hurdle in machine learning and deep learning applications, particularly in scenarios with real-world data. This aspect, while a point of consideration, does not diminish the overall potential and breakthroughs offered by the ViT model in our study.

7. Discussion

In our study, we have attempted to replicate and enhance the benchmarks established in the ChiSig paper, utilizing their dataset of Chinese signatures. The key innovation in our approach lies in the introduction of a Vision Transformer (ViT) to the realm of signature verification. This paper presents the capabilities of ViT in distinguishing between forged and original signatures, a critical aspect of document security and authenticity verification.

However, our findings revealed that the performance of the ViT did not meet our initial expectations. A significant challenge encountered was the class imbalance within our dataset. Specifically, there was a disproportionate number of negative images (forgeries) compared to positive ones (originals) [15]. This imbalance posed a considerable challenge for our ViT, as it was not optimally configured to handle such skewed data distribution. This observation underlines the need for more refined models that can adapt to and effectively process datasets with significant class imbalances.

Another notable aspect of our study was the difference in the data split compared to the baseline paper. The original paper did not specify the parameters for splitting the data, which led to discrepancies in our results. This variation is a crucial reminder of the dependence of machine learning outcomes on the specific nature and division of the dataset used. It emphasizes that exact replication of results is often challenging due to the dynamic nature of data-driven learning processes.

We hope that our research, despite the challenges and variations encountered, will inspire further exploration in the field. Specifically, we see a promising avenue for future research in developing adaptive Vision Transformers that can more effectively handle class imbalances and other dataset-specific challenges in signature verification. Our experience underscores the importance of continual evolution and adaptation in machine learning methodologies to meet the ever-changing demands of real-world applications.

References

[1] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 63–84, Jan. 2000, doi: 10.1109/34.824821.

[2] K. Yan et al., "Signature Detection, Restoration, and Verification: A Novel Chinese Document Signature Forgery Detection Benchmark," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA: IEEE, Jun. 2022, pp. 5159–5168. doi: 10.1109/CVPRW56347.2022.00564.

[3] M. M. Hameed, R. Ahmad, M. L. M. Kiah, and G. Murtaza, "Machine learning-based offline signature verification systems: A systematic review," Signal Processing: Image Communication, vol. 93, p. 116139, Apr. 2021, doi: 10.1016/j.image.2021.116139.

[4] S. Dey, A. Dutta, J. I. Toledo, S. K. Ghosh, J. Llados, and U. Pal, "SigNet: Convolutional Siamese Network for Writer Independent Offline Signature Verification." arXiv, Sep. 30, 2017. doi: 10.48550/arXiv.1707.02131.

[5] P. Wei, H. Li, and P. Hu, "Inverse Discriminative Networks for Handwritten Signature Verification," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5764–5772. Accessed: Dec. 17, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Wei_Inverse_Discriminative_Networks_for_Handwritten_Signature_Verification_CVPR_2019_paper.html

[6] M. A. Ferrer, M. Diaz-Cabrera, and A. Morales, "Static Signature Synthesis: A Neuromotor Inspired Approach for Biometrics," IEEE Trans Pattern Anal Mach Intell, vol. 37, no. 3, pp. 667–680, Mar. 2015, doi: 10.1109/TPAMI.2014.2343981.

[7] S. Pal, A. Alaei, U. Pal, and M. Blumenstein, "Performance of an Off-Line Signature Verification Method Based on Texture Features on a Large Indic-Script Signature Dataset," in 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Apr. 2016, pp. 72–77. doi: 10.1109/DAS.2016.48.

[8] J. Galbally, J. Fierrez, M. Martinez-Diaz, and J. Ortega-Garcia, "Improving the Enrollment in Dynamic Signature Verification with Synthetic Samples," in 2009 10th International Conference on Document Analysis and Recognition, Jul. 2009, pp. 1295–1299. doi: 10.1109/ICDAR.2009.38.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[10] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning." arXiv, Aug. 23, 2016. doi: 10.48550/arXiv.1602.07261.

[11] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks." arXiv, Apr. 10, 2017. doi: 10.48550/arXiv.1611.05431.

[12] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv, Apr. 10, 2015. doi: 10.48550/arXiv.1409.1556.

[13] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021. doi: 10.48550/arXiv.2010.11929.

[14] 'Understanding Categorical Cross-Entropy Loss, Binary Cross-Entropy Loss, Softmax Loss, Logistic Loss, Focal Loss and all those confusing names'. Accessed: Dec. 17, 2023. [Online]. Available: https://gombru.github.io/2018/05/23/cross_entropy_loss/

[15] T.-Y. Lin et al., 'Microsoft COCO: Common Objects in Context'. arXiv, Feb. 20, 2015. doi: 10.48550/arXiv.1405.0312.