BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Electronic & Information Systems
https://ojs.bilpublishing.com/index.php/jeis

ARTICLE

Ensemble Model of Attention Mechanism-Based DCGAN and Autoencoder for Noised OCR Classification

Huitao Zhang[1], Shuguang Xiong[2*], Meng Wang[3]

1 Bybit Global Digital Solutions FZE, Dubai, United Arab Emirate
2 Baidu Inc. Beijing 100085, China
3 Newmark Group, New York City, NY 10017, US

ARTICLE INFO

Article history
Received: 1 March 2022
Accepted: 24 March 2022
Published Online: 31 March 2022

Keywords:
Component
Optical character classification
Deep learning
Autoencoder
GAN

ABSTRACT

Optical Character Recognition (OCR) is a technology that converts images of text into machine-readable formats, essential for digitizing printed texts and enabling digital searches. Traditional OCR methods often struggle with variations in font styles and noise. This paper proposes an innovative approach to enhance OCR classification under challenging conditions by leveraging an ensemble model that combines an Attention Mechanism-Based Generative Adversarial Network (GAN) and an Autoencoder. The GAN generates synthetic data to mitigate the limitations of small datasets, while the autoencoder extracts robust features from noisy images. The model undergoes a two-phase training process, initially learning from the augmented dataset and then fine-tuning on a smaller, labeled dataset. Grad-CAM is used to demonstrate interpretability, highlighting the attention regions during predictions. Experimental results show significant improvements in OCR accuracy and robustness, validating the effectiveness of the proposed method in handling noise and limited training data.

1. Introduction

Optical Character Recognition (OCR) is a technology that converts different types of documents, such as scanned paper documents, PDF files, or images captured by a digital camera, into editable and searchable data [1,2]. OCR is crucial for various applications, including digitizing printed texts, enabling digital searches, automating data entry processes, and supporting accessibility for the visually impaired. By transforming images of text into machine-readable formats, OCR plays an essential role in data processing, information retrieval, and efficient document management [3].

OCR technology has its roots in traditional image processing methods, which primarily relied on handcrafted features and heuristic algorithms to identify and classify text within images. Techniques such as edge detection, contour tracing, and template matching were commonly

*Corresponding Author:
Shuguang Xiong,
Baidu Inc. Beijing 100085, China;
Email: xiongshuguang@baidu.com

DOI: https://doi.org/10.30564/jeis.v4i1.6725

used in early OCR systems [4]. These methods focused on identifying characters based on their geometric properties and pixel intensity patterns. While effective for clean, high-contrast images, traditional approaches often struggled with variations in font styles, sizes, and the presence of noise or distortions. The introduction of machine learning brought significant improvements to OCR technology. Machine learning algorithms have achieved excellent performance in many tasks [5,6]. For instance, Liu et al. proposes a framework to leverage cost-effective and robust ultra-wideband (UWB) radio technology for wireless distance sensing. It introduces a machine learning method based on the extreme gradient boosting decision tree and incorporates error mitigation techniques to enhance measurement accuracy [7]. Xiong et al. tackles the scaling issue by utilizing Distributed Data Parallel (DDP) frameworks to improve the training of deep learning models, with a particular emphasis on the generation of synthetic fingerprints [8]. These algorithms could adapt to different fonts and writing styles by learning discriminative features from the data, thus providing more flexibility and robustness compared to heuristic methods. The advent of deep learning marked a paradigm shift in OCR research and development. Deep learning models, particularly Convolutional Neural Networks (CNNs) [9–11], have demonstrated remarkable success in image recognition tasks, including OCR. CNNs can automatically learn hierarchical features from raw pixel data, making them highly effective for text recognition. These models excel at capturing spatial and temporal dependencies in text images, enabling accurate recognition of complex scripts and handwritten text. During COVID-19, OCR technology, powered by CNNs, enhanced online courses by making materials more accessible to students with disabilities and efficiently managing and organizing large volumes of course content [12,13]. Similarly, in research, OCR technology converts experimental records and notes into searchable digital formats, enhancing efficiency and aiding data integration for biomass transformation studies [14].

Despite these advancements, OCR systems still face significant challenges, particularly in extreme scenarios. Two primary issues are the presence of noise and the availability of limited training data: (1) Noise and Distortions: OCR accuracy can degrade significantly when dealing with noisy images. Noise can arise from various sources, including poor image quality, background clutter, or distortions such as blurring and compression artifacts [15,16]. Traditional denoising techniques are often insufficient, as they can remove essential text features along with the noise. (2) Limited Training Data: Deep learning models typically require large amounts of labeled data to achieve high performance [17,18]. However, obtaining a substantial volume of labeled OCR data can be challenging and costly. Small sample sizes can lead to overfitting, where the model performs well on training data but poorly on new, unseen data.

In this paper, we propose an innovative approach to enhance OCR classification under challenging conditions by leveraging an ensemble model that combines the strengths of an Attention Mechanism-Based Generative Adversarial Network (GAN) and an Autoencoder. The ensemble model operates in two key stages. First, the GAN, enhanced with an attention mechanism, generates a substantial amount of synthetic data. This data augmentation step is crucial for addressing the limitations posed by small datasets, providing the model with a diverse set of training examples that encapsulate various patterns and distortions. Once the synthetic data is generated, it is fed into an autoencoder for feature extraction. The autoencoder, trained to learn robust feature representations from noisy input images, undergoes a two-phase training process. Initially, the autoencoder is trained on the augmented dataset, allowing it to capture intricate details and patterns within the noisy data. After this pretraining phase, the autoencoder's weights are partially frozen, specifically retaining the learned representations in the earlier layers. In the final phase, only the last few layers of the autoencoder are fine-tuned using a smaller, labeled real dataset of noisy OCR images. This fine-tuning step focuses on optimizing the classification accuracy of the model, ensuring that it can effectively distinguish between different characters and texts even in the presence of significant noise.

This paper is structured as follows: Section 2 details the related works on optical character classification. Section 3 describes the workflow of the proposed method. The experimental results and their corresponding discussions are presented in Section 4. Finally, Section 5 provides a comprehensive conclusion of the study.

2. Literature review

Optical character prediction

Recent advancements in machine learning-based Optical Character Recognition (OCR) have shown significant progress across various methodologies and applications. Notably, anchor graph hashing enhances OCR by enabling efficient text data indexing and retrieval, improving processing speed and accuracy [19]. These advancements significantly boost OCR performance and usability. For instance, Memon et al. carried out a comprehensive review and outlined the evolution of handwritten OCR technologies, emphasizing the application of machine learning techniques [20]. This study reviewed numerous papers to establish the quality and relevance of research in the field, focusing on methodologies that incorporate feature extraction and classification techniques. This review highlights the increasing adoption of advanced machine learning methods in improving the accuracy and efficiency of OCR systems. Li et al. focuses on leveraging the power of pre-trained models like RoBERTa and MiniLM for OCR tasks. The approach involves a pipeline where textline images are input to extract visual features and predict wordpiece tokens, enhancing recognition capabilities for both printed and handwritten texts. The model employs pre-training on a large-scale dataset of textline images, followed by fine-tuning on specific OCR tasks, demonstrating substantial improvements in text recognition accuracy [21]. Deng et al. explores a neural network model that employs a coarse-to-fine attention mechanism for generating structured documents from images. It is particularly useful for applications requiring detailed attention to text structure and layout, such as converting mathematical formulas from images to LaTeX code [22]. This technology could also aid remote learning by enabling the accurate digital conversion of handwritten notes and complex content [23]. Additionally, advanced OCR technology enables efficient extraction and processing of financial data, supporting precise market analysis and real-time monitoring, which aids in effective economic policy [24]. These advancements can be applied to studies of molecular mechanisms in protein function and lipid metabolism, automating the extraction and analysis of experimental data to enhance research efficiency and accuracy, while also streamlining the study of enzyme activity and expression inhibition mechanisms [25–27].

Despite significant progress in OCR, the challenges posed by noise and limited training data have frequently been neglected. These issues can severely impair the performance of OCR models, especially in practical scenarios where such imperfections are common. This paper seeks to tackle these overlooked problems by developing methods that enhance the accuracy of OCR models under noisy conditions and when training data is scarce.

3. Method

As shown in **Figure 1**, we propose a novel approach to improve OCR classification under difficult conditions by combining a Generative Adversarial Network (GAN) with an attention mechanism and an Autoencoder. First, the GAN generates a large amount of synthetic data to overcome the limitations of small datasets. Then, this data is used to train an Autoencoder for robust feature extraction. The Autoencoder first learns from the synthetic data, and then its last few layers are fine-tuned with a smaller, labeled dataset to enhance classification accuracy.

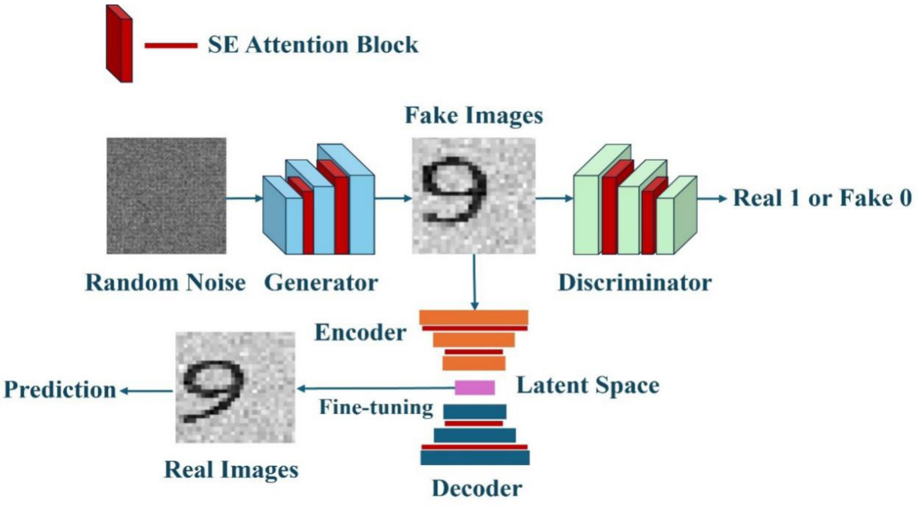

Figure 1. The general framework of Ensemble Model of Attention Mechanism-Based GAN and Autoencoder for Noised OCR Classification.

3.1 Dataset preparation

In our study, we utilized data sourced from Kaggle [28], a renowned platform for data science competitions. Kaggle allows users to find and publish datasets, explore and build models in a web-based environment, and collaborate with other data enthusiasts. The dataset was generated using 3,475 font styles available in Google Fonts, with each alphanumeric character (uppercase, lowercase, and numerals) produced in each font style and organized in a directory. This resulted in a total of 210,000 images across 62 classes. To balance performance and task complexity, we selected only 10 categories, specifically the digits 0–9.

Each image is in grayscale, and the final dataset for training consists of 31,257 images. These images were resized to 28x28 pixels and normalized. Normalization, which scales pixel values typically to a range of 0 to 1, improves model training efficiency and aids in faster convergence. The dataset was then split, with 80% used for training and 20% for testing. Sample images from the dataset are shown in **Figure 2**.

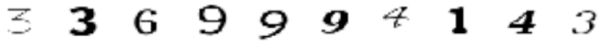

Figure 2. Sample images.

To simulate the presence of noise in our images, we introduced Gaussian noise to both the training and testing datasets. This was accomplished by adding noise with a mean of 0.0 and a standard deviation of 1.0, scaled by a noise factor of 0.1. By adjusting the noise factor, we controlled the intensity of the noise added to the images. This approach creates a more realistic scenario where the images are affected by typical disturbances, thereby training our model to be robust against such variations. Sample images with added noise are shown in **Figure 3**.

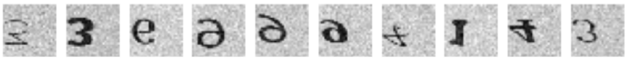

Figure 3. Noised images.

3.2 The introduction of deep learning models

GAN

Generative Adversarial Networks (GANs) are a class of artificial intelligence algorithms used in unsupervised machine learning [29–31], implemented by a system of two neural networks contesting with each other in a game (hence “adversarial”). GANs were introduced by Ian Goodfellow et al. in 2014 and have since been used to generate photorealistic images, create art, and even simulate virtual environments. DCGAN [32–34], or Deep Convolutional Generative Adversarial Network, is a variant of GAN introduced by Radford et al. that specifically uses convolutional and convolutional-transpose layers in the neural networks, making it more effective at learning spatial hierarchies of features.

In this study, we employed the architecture for DCGAN that leverages a series of specific convolutional layers in both the generator and discriminator models to enhance the generation and analysis of images. The generator model begins with a dense layer that expands a 100-dimensional random noise vector into a substantial feature map. This is followed by a sequence of four transposed convolutional layers, each playing a critical role in upscaling the image size while reducing the number of feature channels: (1) The first transposed convolutional layer increases the feature map from 256 channels of 7x7 to 128 channels of 14x14. (2) The second layer maintains the spatial dimension (14x14) but reduces the depth to 64 channels. (3) The third layer continues at 64 channels, refining details within the 14x14 spatial dimension. (4) The final transposed convolutional layer upscales the feature map to a 28x28 image with a single output channel, suitable for generating grayscale images. Each of the transposed convolutional layers uses a kernel size of 3, a stride of 2 for the upscaling layers, and a stride of 1 for layers that maintain spatial dimensions, with appropriate padding to preserve the size. The activation function between these layers is ReLU, except for the final layer, where a tanh function is applied to normalize the output values.

Conversely, the discriminator model is structured to effectively classify the generated images as real or fake, employing four convolutional layers: (1) The initial convolutional layer reduces the 28x28 input image to a 14x14 feature map with 32 channels. (2) Subsequent layers further compress and process these features, with the second layer maintaining the 14x14 dimension while increasing depth to 64 channels. (3) The third layer deepens the channel count to 128 while maintaining spatial dimensions. (4) The final convolutional layer further reduces the spatial dimension to 7x7 while increasing the feature depth to 256 channels.

In addition, we also enhanced our convolutional architecture by incorporating Squeeze-and-Excitation Networks (SE-Net) attention blocks into both the generator and discriminator models [35–37]. These attention blocks are designed to improve model performance by selectively emphasizing more informative features during image generation and discrimination. The SE-Net block operates by compressing global spatial information into a channel descriptor, effectively allowing the network to prioritize relevant features while suppressing less critical ones. This dynamic recalibration of channel-wise feature responses enables more precise control over information processing within the network. The integration of this attention mechanism significantly boosts the detail and realism of the generated images, as well as the accuracy with which the discriminator can distinguish between real and fabricated images, thereby enhancing the overall efficacy of the GAN architecture.

Autoencoder-based classification

An autoencoder is a type of neural network used to learn efficient coding of unlabeled data, typically for the purpose of dimensionality reduction or feature learning [38,39].

The network is divided into two parts: the encoder, which maps the input to a lower-dimensional representation, and the decoder, which reconstructs the input from this representation.

The goal of using this autoencoder in our research is to harness its capability to extract meaningful features from artificially generated data, which are then fine-tuned on real OCR datasets. This approach is intended to improve the predictive accuracy of OCR models by training them on rich, feature-enhanced data before applying them to real-world tasks, thereby enhancing the model's ability to generalize from synthetic to authentic text images. The encoder part comprises three convolutional layers, each followed by a max pooling layer to progressively reduce the spatial dimensions and increase the depth of feature extraction. These layers employ a stride of 1 and padding to maintain the size of the feature maps, with ReLU activations to introduce non-linearity [40,41]. Max pooling helps to distill the essential features while reducing dimensionality. The decoder reverses this process. Starting with the compressed feature map, it uses max unpooling layers to restore the spatial dimensions, paired with transposed convolutional layers that aim to reconstruct the original image from the encoded features. The unpooling layers utilize stored indices from the pooling layers of the encoder to accurately place values in the expanded map, ensuring a precise structural reconstruction. The final layer uses a sigmoid activation to normalize the output, producing a reconstructed image.

3.3 Implementation details

In this paper, both the DCGAN and the Autoencoder are trained using the Adam optimizer [42,43]. The DCGAN utilizes a loss function based on binary cross-entropy to effectively train the discriminator and generator. The Autoencoder uses Mean Squared Error (MSE) for reducing reconstruction errors. During the fine-tuning phase, the Autoencoder is trained using sparse categorical crossentropy to handle classification tasks [44,45], with a batch size of 256. The models are trained and evaluated in phases, utilizing accuracy, precision, recall, and F1 score as key metrics. All training and evaluation are performed on a 3090 GPU, ensuring high computational efficiency.

4. Results and Discussion

4.1 The performance of the GAN

Figure 4 depicts a graph showing the training curves of a DCGAN. There are two lines representing the losses for the generator (G loss, in green) and the discriminator (D loss, in blue). Both lines show a significant decrease in loss over the initial training epochs, followed by a stabilization. The generator loss levels out slightly below 0.5, whereas the discriminator loss stabilizes around 0.7, continuing through 350 epochs. This suggests that the network is learning effectively, but the generator and discriminator are still competing, which is typical for GANs. In addition, some generated sample images are provided in **Figure 5**. Although the graph indicates that many images generated during the training are of good quality, it also implies that there are still some images with suboptimal quality. This variability in output quality is common in GAN training, especially in cases where the generator has not perfectly learned to mimic the distribution of the training dataset.

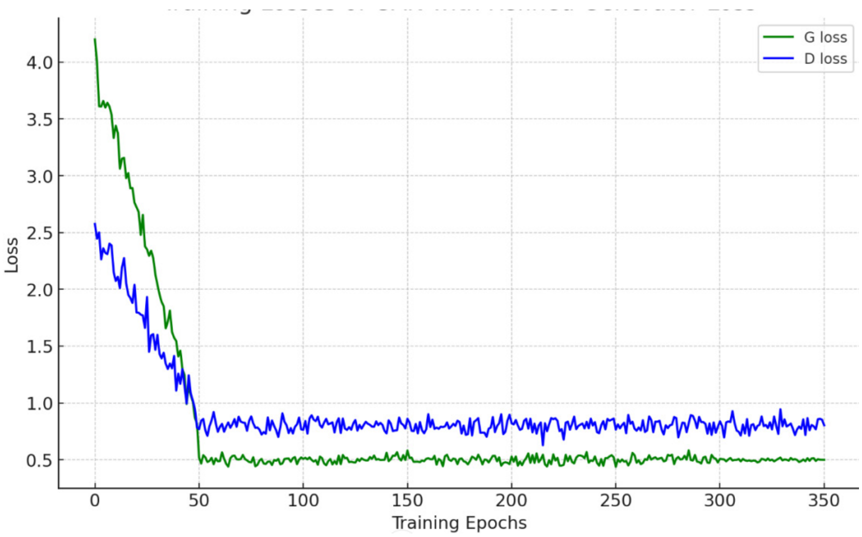

Figure 4. The training curve in terms of the loss.

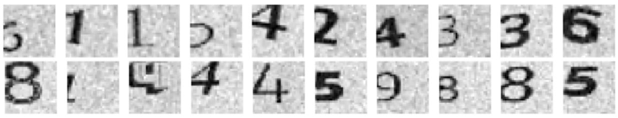

Figure 5. Generated images.

4.2 The performance of classification

In this experiment, different machine learning models were utilized to process features extracted by the encoder part of an autoencoder. The results show that the Fully Connected Layers model performed best across all metrics, followed by the Random Forest [46,47], Decision Tree [48,49], and K-Nearest Neighbors (KNN) models [50,51].

The superior performance of the Fully Connected Layers model can be attributed to its ability to learn complex non-linear relationships between features, which is crucial in handling encoded data. This model's high accuracy, precision, recall, and F1 score indicate its effectiveness in capturing and utilizing the nuances of the input features. On the other hand, the Random Forest model also showed commendable results, likely due to its ensemble approach, which helps in reducing overfitting and improving generalization over diverse data sets. However, its performance was slightly lower than the fully connected layers, possibly because it may not capture as complex patterns as neural networks can. The Decision Tree and KNN models trailed in performance, which might be due to their relatively simpler decision boundaries that struggle with the complexity and high dimensionality of encoded features. Decision trees, in particular, are prone to overfitting unless carefully tuned, while KNN's performance heavily depends on the choice of neighbors and distance metrics, which might not have been optimal in this case.

Table 1. The classification performance of the different models in testing dataset.

Model	Accuracy	Precision	Recall	F1 score
Fully connected layers	0.9107	0.9122	0.9127	0.9130
Random forest	0.8920	0.8928	0.8955	0.8898
Decision tree	0.8811	0.8702	0.8791	0.8795
KNN	0.8628	0.8604	0.8681	0.8692

Figure 6 provides the relationship between the number of images generated using a DCGAN and the accuracy of a model trained with these images. As shown, there is a clear upward trend in accuracy as more images are generated, indicating that the model benefits from a larger number of training samples. The plot visually demonstrates how accuracy increases from 0.9021 with 3000 images to 0.9107 with 9000 images.

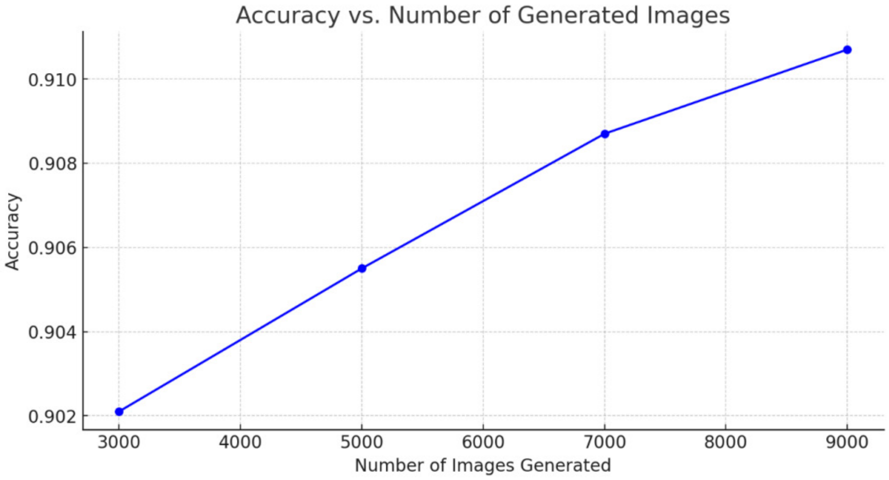

Figure 6. The relationship between the number of images generated and accuracy.

The improvement in model accuracy with an increasing number of images generated by a DCGAN can be largely attributed to the enhanced data diversity and the reduced risk of overfitting. As more images are generated, they provide a broader spectrum of data variations, which helps in training models that can generalize better to new, unseen data. This diversity not only prevents the model from learning irrelevant details—thus mitigating overfitting—but also aids the autoencoder in extracting more meaningful and discriminative features from a richer dataset. Consequently, the model trained with these features becomes more robust and performs better in predictive tasks, as it is based on a more comprehensive understanding of the data's underlying patterns.

4.3 Ablation study

Table 2 presents the results of an ablation study focusing on the impact of incorporating an attention module, specifically an attention mechanism in SENet, into a model trained using features extracted via DCGAN and autoencoder techniques. The comparison is between two configurations: one model includes the attention module, and the other does not.

The model equipped with the attention module achieves higher performance across all metrics compared to the model without the attention module. Specifically, the accuracy of the model with the attention module is 0.9107, versus 0.9021 for the model without it. Similarly, precision, recall, and F1 score are also improved with the inclusion of the attention module—precision rises from 0.9008 to 0.9122, recall from 0.9043 to 0.9127, and the F1 score from 0.9033 to 0.9130. These results underscore the advantages of integrating the SENet attention module in models involving complex feature extraction and generation tasks such as those performed by DCGANs and autoencoders. The attention module likely helps the model focus more effectively on the most informative features of the input data, thereby enhancing the model's ability to generalize and perform more accurately on diverse datasets.

Table 2. The ablation study for the attention module.

Model	Accuracy	Precision	Recall	F1 score
Model with attention module	0.9107	0.9122	0.9127	0.9130
Model without attention module	0.9021	0.9008	0.9043	0.9033

4.4 Interpretability visualization

This study utilizes Gradient-weighted Class Activation Mapping (Grad-CAM) to demonstrate the interpretability of model predictions during the fine-tuning of an autoencoder and shows the results in **Figure 7**. The heatmaps illustrate how the model focuses on specific areas to make its predictions. These visualizations confirm that the fine-tuned autoencoder has good interpretability, as the Grad-CAM technique effectively highlights the important features used by the model for its predictions.

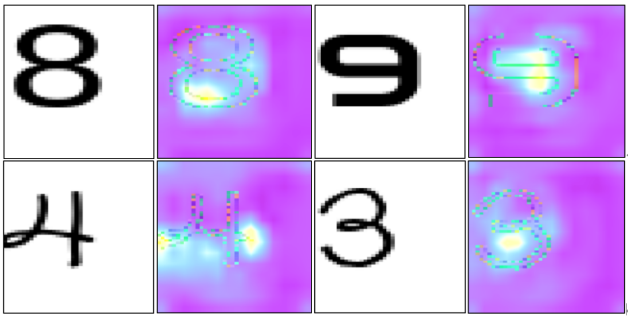

Figure 7. The interpretability of the model based on Grad-CAM.

5. Conclusion

In conclusion, this study presents a novel ensemble model combining an Attention Mechanism-Based Generative Adversarial Network (GAN) and an Autoencoder to enhance OCR classification under challenging conditions. The approach effectively addresses the limitations posed by noise and limited training data. By generating synthetic data through the GAN and extracting robust features using the autoencoder, the model significantly improves OCR accuracy and robustness. The use of Grad-CAM for interpretability further highlights the model's ability to focus on relevant features, ensuring reliable predictions.

Our experimental results demonstrate that the ensemble model outperforms traditional OCR methods, particularly in noisy environments. The inclusion of attention mechanisms within the GAN enhances the quality of generated data, contributing to better model performance. This study's findings underscore the potential of combining advanced generative models with robust feature extraction techniques to tackle real-world OCR challenges. Future work will focus on refining the model architecture, exploring additional attention mechanisms, and expanding the dataset to include more diverse and complex text patterns. The proposed approach paves the way for more accurate and reliable OCR systems, benefiting applications across various domains, from document digitization to accessibility support.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2021YFB31012003, 2021YFB31012004, 2021YFB31012005).

References

[1] A. Chaudhuri et al., 2017. Optical character recognition systems. Springer.

[2] N. Islam, Z. Islam, N. Noor, 1999. A survey on optical character recognition system. arXiv preprint arXiv:1710.05703.

[3] G. Nagy,. Nartker, T.A, Rice, S.V., 1999. Optical character recognition: An illustrated guide to the frontier. Document recognition and retrieval VII. SPIE, pp. 58–69.

[4] White, J.M., Rohrer, G.D., 1983. Image thresholding for optical character recognition and other applications requiring character image extraction. IBM Journal of research and development. 27(4), 400–411.

[5] F. Yu, J. Strobel, 2021. Work-in-Progress: Pre-college Teachers' Metaphorical Beliefs about Engineering. 2021 IEEE Global Engineering Education Conference (EDUCON). IEEE. pp. 1497–1501.

[6] Y. Qiu, 2019. Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling. Johns Hopkins University.

[7] Y. Liu, L. Liu, L. Yang, et al., 2021. Measuring distance using ultra-wideband radio technology enhanced by extreme gradient boosting decision tree (XGBoost). Automation in Construction. 126, 103678.

[8] S. Xiong, H. Zhang, M. Wang, et al., 2022. Distributed data parallel acceleration-based generative adversarial network for fingerprint generation. Innovations in Applied Engineering and Technology. 1–12.

[9] Y. Liu, Y. Bao, 2022. Review on automated condition assessment of pipelines with machine learning. Advanced Engineering Informatics. 53, 101687.

[10] S. Li, K. Singh, N. Riedel, et al., 2022. Digital learning experience design and research of a self-paced online course for risk-based inspection of food imports," Food Control, vol. 135, p. 108698, 2022.

[11] Y. Qiu, Y. Yang, Z. Lin, et al., 2020. Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV. China Communications. 17(3), 46–57.

[12] F. Yu, J. O. Milord, L. Y. Flores, et al., 2022. Work in Progress: Faculty choice and reflection on teaching strategies to improve engineering self-efficacy. 2022 ASEE Annual Conference.

[13] J. Milord, F. Yu, S. Orton, et al., 2021. Impact of COVID Transition to Remote Learning on Engineering Self-Efficacy and Outcome Expectations. 2021 ASEE Virtual Annual Conference.

[14] D. Xia, A. K. Alexander, A. Isbell, et al., 2017. Establishing a co-culture system for Clostridium cellulovorans and Clostridium aceticum for high efficiency biomass transformation. The Journal of Science and Health at the University of Alabama. 14, 8–13.

[15] E. Boros, N. K. Nguyen, G. Lejeune, et al., 2022. Assessing the impact of OCR noise on multilingual event detection over digitised documents. International Journal on Digital Libraries. 23(3), 241–266.

[16] N. Premchaiswadi, S. Yimgnagm, W. Premchaiswadi, 2010. A scheme for salt and pepper noise reduction and its application for ocr systems. WSEAS Transactions on Computers. 9(4), 351–360.

[17] J. Martínek, L. Lenc, P. Král, 2020. Building an efficient OCR system for historical documents with little training data. Neural Computing and Applications. 32,17209–17227.

[18] J. Martínek, L. Lenc, P. Král, et al., 2019. Hybrid training data for historical text OCR. 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE. pp. 565–570.

[19] G. Sun, T. Zhan, B. G. Owusu, et al., 2020. Revised reinforcement learning based on anchor graph hashing for autonomous cell activation in cloud-RANs. Future Generation Computer Systems. 104, 60–73.

[20] J. Memon, M. Sami, R. A. Khan, et al., 2020. Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). IEEE access. 8, 142642–142668.

[21] M. Li et al., 2023. Trocr: Transformer-based optical character recognition with pre-trained models,. Proceedings of the AAAI Conference on Artificial Intelligence. 37(11), 13094–13102.

[22] Y. Deng, A. Kanervisto, J. Ling, et al., 2017. Image-to-markup generation with coarse-to-fine attention. International Conference on Machine Learning. PMLR. 980–989.

[23] F. Yu, J. Milord, S. L. Orton, et al., 2022. The concerns and perceived challenges students faced when traditional in-person engineering courses suddenly transitioned to remote learning. 2022 ASEE Annual Conference

[24] Y. Qiu, 2017. Financial Deepening and Economic Growth in Select Emerging Markets with Currency Board Systems: Theory and Evidence. The Johns Hopkins Institute for Applied Economics, Global Health.

[25] Y. Shen, H.-m. Gu, L. Zhai, et al., 2022. The role of hepatic Surf4 in lipoprotein metabolism and the development of atherosclerosis in apoE−/− mice. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids. 1867(10), 159196.

[26] M. Wang et al., 2022. Identification of amino acid residues in the MT-loop of MT1-MMP critical for its ability to cleave low-density lipoprotein receptor. Frontiers in Cardiovascular Medicine. 9, 917238.

[27] J. Horne et al., 2020. Caffeine and Theophylline Inhibit β-Galactosidase Activity and Reduce Expression in Escherichia coli. ACS omega. 5(50), 32250–32255.

[28] Kaggle, 2022. OCR-dataset [Internet]. Available from: https://www.kaggle.com/datasets/harieh/ocr-dataset(cited May 1, 2024)

[29] A. Creswell, T. White, V. Dumoulin, et al., 2018.

Generative adversarial networks: An overview. IEEE signal processing magazine. 35(1), 53–65.

[30] K. Wang, C. Gou, Y. Duan, et al., 2017. Generative adversarial networks: introduction and outlook. IEEE/CAA Journal of Automatica Sinica. 4(4), 588–598.

[31] J. Gui, Z. Sun, Y. Wen, et al., 2021. A review on generative adversarial networks: Algorithms, theory, and applications. IEEE transactions on knowledge and data engineering. 35(4), 3313–3332.

[32] Q. Wu, Y. Chen, J. Meng, 2020. DCGAN-based data augmentation for tomato leaf disease identification. IEEE access. 8, 98716–98728.

[33] W. Fang, F. Zhang, V. S. Sheng, et al., 2018. A Method for Improving CNN-Based Image Recognition Using DCGAN. Computers, Materials and Continua. 57, 1.

[34] B. Liu, J. Lv, X. Fan, et al., 2022. Application of an improved dcgan for image generation. Mobile Information Systems. 2022.

[35] J. Hu, L. Shen, G. Sun, 2018. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 7132–7141.

[36] X. Jin, Y. Xie, X.-S. Wei, et al., 2022. Delving deep into spatial pooling for squeeze-and-excitation networks. Pattern Recognition. 121, 108159.

[37] Y. Qiu, J. Wang, Z. Jin, et al., 2022. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control. 72, 103323.

[38] M. Tschannen, O. Bachem, M. Lucic, 2018. Recent advances in autoencoder-based representation learning. arXiv preprint arXiv:1812.05069.

[39] Y. Zhang, 2018. A better autoencoder for image: Convolutional autoencoder [Internet]. Available from: http://users.cecs.anu.edu.au/Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58.pdf (cited Mar. 23, 2017).

[40] A. F. Agarap, 2018. Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

[41] Y. Bai, 2022. RELU-function and derived function review. SHS Web of Conferences. 144, 02006.

[42] Z. Zhang, 2018. Improved adam optimizer for deep neural networks. 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS). Ieee. 1–2.

[43] K. Bae, H. Ryu, H. Shin, 2019. Does Adam optimizer keep close to the optimal point? arXiv preprint arXiv:1911.00289.

[44] B. Chaithanya, T. Swasthika Jain, A. Usha Ruby, et al., 2021. An approach to categorize chest X-ray images using sparse categorical cross entropy. Indonesian Journal of Electrical Engineering and Computer Science. 1700–1710.

[45] J. Kakarla, B. V. Isunuri, K. S. Doppalapudi, et al., 2021. Three-class classification of brain magnetic resonance images using average-pooling convolutional neural network. International Journal of Imaging Systems and Technology. 31(3), 1731–1740.

[46] S. J. Rigatti, 2017. Random forest. Journal of Insurance Medicine. 47(1), 31–39.

[47] G. Biau and E. Scornet, 2016. A random forest guided tour. Test. 25, 197–227.

[48] Y.-Y. Song, L. Ying, 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry. 27(2), 130.

[49] S. Suthaharan, 2016. Machine learning models and algorithms for big data classification. Integr. Ser. Inf. Syst. 36, 1–12.

[50] L. E. Peterson, 2009. K-nearest neighbor. Scholarpedia. 4(2), 1883.

[51] J. Laaksonen, E. Oja, 1996. Classification with learning k-nearest neighbors. Proceedings of international conference on neural networks (ICNN'96). IEEE. pp. 1480–1483.