Journal of Advanced Marine Engineering and Technology, Vol. 47, No. 6, pp. 411~418, 2023
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2023.47.6.411

ISSN 2234-7925 (Print)
ISSN 2765-4796 (Online)
Original Paper

Study on handwritten invoice recognition system

Yoon-Sang Han[1] · Hong-Il Seo[2] · Dong-Hoan Seo[†]

(Received December 4, 2023 ; Revised December 14, 2023 ; Accepted December 23, 2023)

Abstract: In recent years, the growing preference for contactless services has resulted in an increased demand for kiosk-based reception systems. Although several kiosks have been activated for product orders, the reception services that entail more intricate procedures and require extensive information, such as those in postal and logistics services, have not been adequately embraced. Furthermore, the conventional paper-based reception systems fail to suffice the modern requirements. Therefore, this study proposes an innovative system for automatically processing handwritten invoices. The proposed system accurately extracts invoice details from parcel images and automatically digitizes the sender and receiver information by employing a visual document understanding model. In particular, this study presents the essential construction of an optimal training dataset that is required during the fine-tuning process of neural network-based models. It is anticipated that this will dramatically enhance the performance and accuracy of systems that handle handwritten invoices, paving the way for innovative advancements in the forthcoming years.

Keywords: Handwritten, Optical character recognition, Visual document understanding, Neural Network, Transformer

1. Introduction

The recent decline in the global population has resulted in an increased demand for automated mechanical consultation and reception systems to maintain service quality. A majority of the online service systems reduce manpower by using advanced chatbots to accurately comprehend user service needs and connect them to specialized agents **[1][2]**. Offline services often require manual inspection due to the changes in reception or handwritten submissions **[3][4]**. To automate this process while minimizing user discomfort, replicating these human-like actions during reception is crucial to ensure accurate verification **[5]**. The automation of postal and parcel reception systems requires a technology that can recognize and digitize the handwritten invoices captured by cameras.

The optical character recognition (OCR) technology is extensively used for identifying the text from the images of handwritten invoices **[6]**. The OCR technology is categorized into online and offline systems. The online systems utilize additional data such as the writing style, sequence, and speed to aid character recognition **[7]**. Conversely, the offline systems rely solely on images containing handwritten text, making them more challenging **[8][9]**. In recent years, there have been significant advancements in the application of neural networks; they exhibit outstanding performance when used as classification algorithms. However, as OCR can recognize only individual texts, tailored model improvements are required to suit its industrial applications.

Visual document understanding (VDU) recognizes the text within document images and comprehends the content **[10]-[12]**. This approach has been extensively studied across various aspects, such as document classification based on purpose, labeling document sequences, and classifying the attribute values of individual texts. Contrary to OCR, VDU learns both text and document structures; therefore, it is composed of vision transformers and language transformers **[13]-[17]**. Vision transformers extract visual features from document images, whereas language transformers leverage these visual features to learn text recognition and other specific objectives, thus providing information regard-

[†] Corresponding Author (ORCID: http://orcid.org/0000-0003-3610-0356): Professor, Division of Electronics & Electrical Information Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49112, Korea, E-mail: kosme@kmou.ac.kr, Tel: 051-405-1050

[1] M. S. Candidate, Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: hanysang@gmail.com, Tel: 051-410-4822

[2] Ph. D. Candidate, Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: seoluck77@gmail.com, Tel: 051-410-4822

ing the document. This technology is highly innovative for digitizing historical documents and records.

This study proposes a handwritten invoice recognition system for kiosks to automate postal and parcel reception. This system captures the invoice-related images from parcel images using a camera, and then extracts the necessary information from the cropped invoice images. Image processing algorithms are utilized in the cropping process, and a VDU model, specifically the donut model, is employed for extracting information from the cropped images. The donut was further trained to enhance the performance in the invoice domain. To address the need for a substantial volume of data when applying the model to different domains, data augmentation methods were employed in additional model training. Furthermore, this study showcases the results regarding the ratio of training between the generated and collected data to stabilize the model's performance.

The major contributions of this work are summarized as follows:

1. This paper proposes an unmanned reception-capable system for recognizing handwritten invoices.
2. VDU is applied to images captured by a camera to extract information, achieving excellent results.
3. To stabilize the VDU model, this study presents the ratio between the data collected from the applied domain and the generated data.

2. Related Work

2.1 Traditional OCR Model

The OCR technology converts text-containing images into digital formats. This technology has been extensively studied in the domain of paper document management and serves as a means of digitizing the information contained in paper documents that are hard to manage **[18]**. Recent OCR approaches primarily involve two steps: text region detection within images, and recognition of detected characters within these regions **[19]-[22]**. In the first step, convolutional neural network (CNN)-based object-detection models are used to locate text regions in images and crop these text regions for the second step. In the second step, CNN-based classification models are used to convert the text images into a digital format. This technology demonstrates the potential for automating tasks involving paper documents by recognizing and digitizing the text within document images. However, conventional OCR models only output text and coordinates, which limits their use in specific applications.

2.2 VDU Model

VDU leverages neural networks to understand the layout and content of documents. Unlike traditional OCR, VDU considers diverse text types along with the overall structure of the document, enabling precise extraction of the required data. This further enhances the document processing efficiency. Typically, VDU models operate through a pipeline involving text recognition akin to OCR, merging the extracted text features with the document image **[10]-[14]**. This approach is essential for understanding the complex layout and structure of a document and extracting important information.

In general, VDU models, such as the LayoutLM series and DocFormer series, help understand the structural context of a document by analyzing the text and location within the document **[10]-[14]**. LayoutLM identifies the structural context of text using two-dimensional positional embeddings that represent the relative positions of tokens in documents, and image embeddings for scanned token images **[10]-[11]**. DocFormer, on the contrary, combines visual features and location-related information to capture the visual context around a text by processing the multimodal data containing text, visual elements, and location information **[12]-[14]**.

In contrast, the donut model comprehensively analyzes the visual and linguistic features of document images by directly processing the text in the image along with its structure using a transformer-based architecture without embedding any location information **[17]**. This approach helps understand the document better without the need for complex pre-training or sophisticated data labeling. The donut simplifies the data collection and training processes, while exhibiting excellent performance on a variety of downstream VDU tasks. Owing to the aforementioned advantages, we chose to use the donut model in our experiments.

3. Proposed Method

3.1 Proposed Handwritten Invoice System

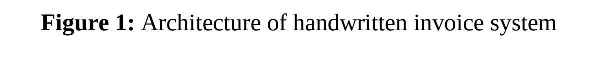

Parcel Box Image
Sender-Receiver Region Detection
Sender-Receiver Information Extraction
Output (Invoice Information)

Figure 1: Architecture of handwritten invoice system

The proposed system aims to digitize the sender and receiver information from handwritten invoice images to streamline logistics systems. **Figure 1** illustrates the structure of the handwritten invoice system comprising four steps. In the parcel-box image step, the images of parcels with attached invoices were captured using a camera. In the second step, an image-processing algorithm detects the invoices on the captured images and extracts the sender and receiver images. The sender–receiver information extraction step involves the use of a neural network-based VDU model to digitize the manually written information and extract addresses, names, and phone numbers. Finally, the system corrects the existing errors in the VDU output using an address database before providing invoice information to the user.

3.2 Parcel Box Image

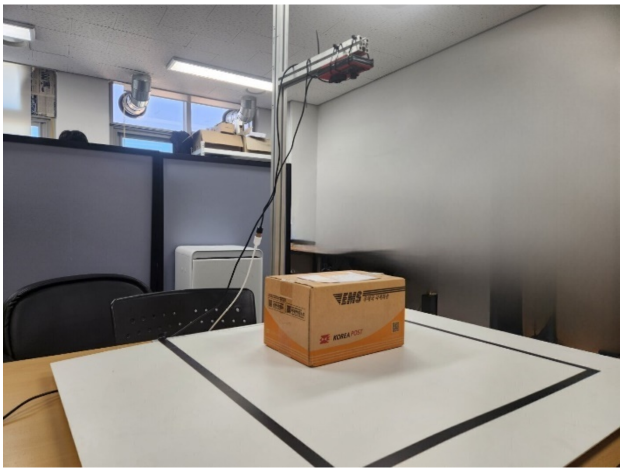


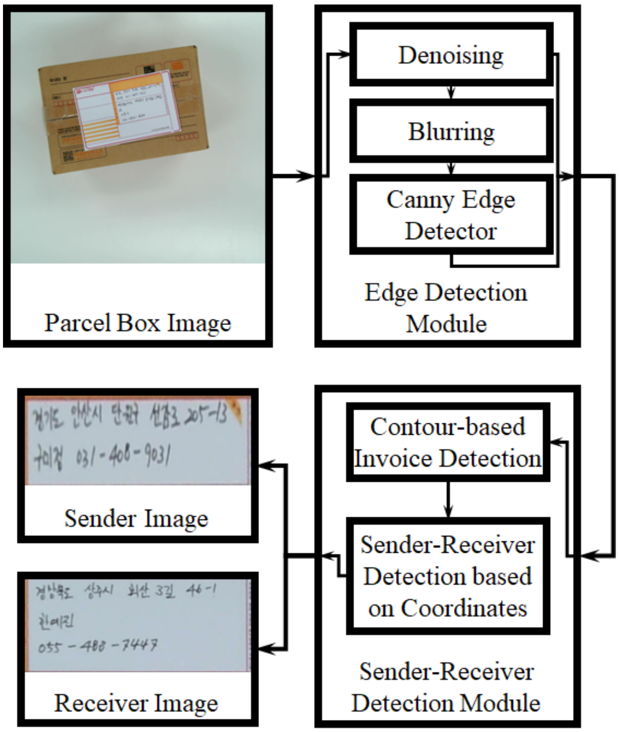
Figure 2: Environment where images are captured

(a) (b)

Figure 3: Testing image

The environment for collecting the parcel-box images is depicted in **Figure 2**. To detect invoices effectively, parcel boxes of different sizes must be encompassed within the images while preserving a minimum resolution for the invoice images. Hence, a UHD (3840 × 2160 pixels) resolution camera was utilized, and a distance of 65 cm was maintained between the camera and the ground. The images captured at this stage are then input into the subsequent step for invoice region detection, as illustrated in **Figure 3**. The results of the image acquisition environment in **Figure 2** are shown in **Figure 3(a)**,**(b)**, where the box consists of different invoices and locations within the camera range.

3.3 Sender–Receiver Region Detection



Parcel Box Image
Denoising
Blurring
Canny Edge Detector
Edge Detection Module
경기도 안산시 단원구 선감3 205-13
구미정 031-408-9031
Sender Image
경상북도 상주시 외산 3길 46-1
한예진
055-488-7447
Receiver Image
Contour-based Invoice Detection
Sender-Receiver Detection based on Coordinates
Sender-Receiver Detection Module

Figure 4: Process of sender–receiver region detection

In this stage, the objective is to detect regions containing the addresses, names, and phone numbers of both the sender and receiver on the parcel box images. As shown in **Figure 4**, the proposed method comprises two modules. The edge detection module utilized a Gaussian filter to eliminate the potential noise during image capture and applies a median filter to blur detailed information in the input image, thus enhancing the edge detection performance for invoices. Finally, a Canny edge detector was used to extract the detected edge images. The edge-detected image and noise-reduced box image were then input into the sender–receiver detection module.

The module utilized the edge and box images to detect the invoice regions and crops the area containing the sender and receiver information within the invoice. An edge image was employed to detect the coordinates of the four corners of the rectangular invoice through contour extraction, which aids the extraction of the invoice region within the box image. By leveraging the standardized format of the invoices, the images containing information regarding the sender and receiver were cropped. These two images were then input separately into the invoice understanding module.

3.4 Text Invoice Recognition

The sender–receiver information extraction involved outputting the sender and receiver information from the extracted invoice image. We used donut, which exhibits excellent performance without requiring text positional labeling for information extraction. The model was trained by adding the generated and collected data to a pretrained donor to reflect the domain of the invoices.

3.4.1 Donut model

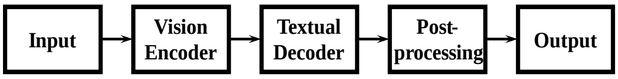


Figure 5: Operational process of donut

As depicted in **Figure 5**, the donut comprises an encoder, which is responsible for extracting detailed visual features that include text types and positions within the image, and a decoder, which is a language model that generates sentences from these feature maps. The input of the model is the sender or receiver image, which outputs addresses, names, and phone numbers. The raw output from the textual decoder undergoes post-processing for refinement, which culminates in the final results.

The donut performed three major steps. The first step involved encoding the image using a vision encoder, wherein handwritten Korean parcel invoices pass through the encoder of the Swin transformer and are converted into n image patch embeddings **[23]**. The second step involved the textual decoder, which utilizes the BART, where the output of the vision encoder and the prompt served as the inputs **[24]**. Finally, conversion was performed on the output obtained from the textual decoder. The output of the textual decoder was tagged with special tokens corresponding to the results recognized by the vision encoder.

The postprocessing handles the refinement of the JSON data output from the textual decoder. At this stage, the sole non-neural-network-based segment within the model conducted data validation using rule- and knowledge-based methods. While different rules apply to the address, name, and phone number fields, a common practice during error detection is to request alternative results from the decoder. For addresses, errors were primarily checked by using administrative rules. The phone numbers were adjusted to adhere to the standard formats consisting of numbers, spaces, and dashes.

3.4.2 Dataset

The donut, which was initially pre-trained, underwent fine-tuning using both the generated and collected datasets. The generated data consisted images wherein each character was written in different handwriting styles using a structured-conditional generative adversarial network (SC-GAN), whereas the collected data comprised images written in the same handwriting style. During the fine-tuning process, we restricted the number of invoices trained to 1000 in order to minimize the ratio of the collected data used and explored various weights for the input data ratios. The validation data ratio was set to 10%, and 500 instances of collected data were used for evaluation purposes.

3.4.3 Training detail

NVIDIA RTX 6000 served as the graphics processing unit (GPU) for training. The training framework incorporated PyTorch 1.11.0+cu113, transformers 4.28.0, tokenizers 0.13.3, Huggingface Hub 0.16.4, and sentence piece 0.1.99. An Adam optimizer with a learning rate of 0.00002 was utilized, employing a batch size of eight and training over 50 epochs. Additionally, a weight decay of 0.01 was set to prevent overfitting during the training process.

4. Experiment

In the experimental section, we created five different datasets by adjusting the ratio between the generated and collected data that were used to construct the training dataset. Subsequently, we trained each of these datasets and evaluated the training efficiency and recognition accuracy of the model both quantitatively and qualitatively. This helped assess the impact of different ratios of dataset compositions on both the learning efficiency and recognition accuracy of the model.

4.1 Quantitative Evaluation

Table 1: Data composition ratios for established model

Model	Collected Data (%)	Generated Data (%)
1	100	0
2	75	25
3	50	50
4	25	75
5	0	100

For 1000 data entries, we adjusted the ratio between the generated and previously collected data to construct the training dataset. The data composition ratios for the configured models are shown in **Table 1**. Model 1 exclusively utilized the collected data, Model 2 consisted of 75% collected data and 25% generated

Table 2: Quantitative evaluation performance of each model

Model	Address Correct	Address Error	Address Accuracy (%)	Full Name Correct	Full Name Error	Full Name Accuracy (%)	Phone Number Correct	Phone Number Error	Phone Number Accuracy (%)	Total Correct	Total Error	Total Accuracy (%)
Model 1	7915	1763	81.7834	1013	487	67.5333	5671	367	93.9218	14599	2617	84.799
Model 2	8705	973	89.9463	1321	179	88.0667	5814	224	96.2902	15840	1376	92.0074
Model 3	8823	855	91.1655	1328	172	88.5333	5824	214	96.4558	15975	1241	92.7916
Model 4	8646	1032	89.3366	1364	136	90.9333	5861	177	97.0686	15871	1345	92.1875
Model 5	8646	1032	89.3366	1364	136	90.9333	5861	177	97.0686	15871	1345	92.1875

data, Model 3 possessed an equal split between 50% collected and 50% generated data, Model 4 comprised 25% collected data and 75% generated data, and Model 5 relied solely on generated data.

Table 2 presents the accuracy evaluation based on the test dataset. The evaluation was conducted using a test dataset comprising 500 datasets. The evaluation was performed on 17,216 characters, and the accuracy for three categories were measured in addition to the overall accuracy: phone numbers, names, and addresses. The model predictions were compared with the ground truth of the actual data to verify the match for each character. In addition, the errors in the recognition and omissions from the predicted results were considered. From the results in **Table 2**, it is evident that all models exhibit high recognition rates for numerical data such as phone numbers, yet display relatively lower accuracy in handwritten recognition such as addresses and names. In particular, Model 1, which is trained solely on the generated data, demonstrated a significantly lower total accuracy of 84.799 % compared to the overall accuracies of the other models, which are 92.0074 %, 92.7916 %, 92.1875 %, and 92.1875 %. This indicates that training solely on the generated data may fail to encapsulate the complexity and diversity of real handwritten recognition. These findings underscore the significance of combining actual collected data with generated data to enhance the accuracy of the model.

4.2 Qualitative Evaluation

The model predictions for the collected data, generated data, and the data collected from real environments were qualitatively evaluated. **Table 2** illustrates the initial predictions (raw output) and final results (output) after post-processing when each dataset is provided as input. Model 1 exhibited a significantly low accuracy for certain data compared to the other models. For instance, in **Table 3**, the prediction from Model 1 appeared as '{'pnumber': '010-5413-2509', 'name': '김정치', 'address': '경정치 010-5413-2509'}.' In particular, the misinterpretation of the phone number as part of the address demonstrates the model's inability to handle the intricacies of the data properly. There were discrepancies between the initially predicted addresses and final results across all models. For example, in **Table 4**, Model 2 initially predicted '충남 태안군 뫼골길 38,' which was later refined through post-processing to '충청남도 태안군 멧골길 38.' This further demonstrates the role of post-processing in refining the model predictions to derive more accurate final results. Furthermore, minimal differences in address recognition were observed among Models 2, 3, 4, and 5. This indicates that these models can process the address data with similar levels of accuracy.

Table 3: Estimation results for testing image of each model

	Raw output	Output
Model 1	: {'pnumber': '010-5413-2509', 'name': '김정치', 'address': '경정치 010-5413-2509'}	{'pnumber': '010-5413-2509', 'name': '김정치', 'address': '경정치 010-5413-2509'}
Model 2	: {'pnumber': '010-5413-2509', 'name': '김정식', 'address': '전라남도 해남군 송지면 557'}	{'pnumber': '010-5413-2509', 'name': '김정식', 'address': '전라남도 해남군 송지면 557'}
Model 3	{'pnumber': '010-5413-2509', 'name': '김정식', 'address': '전라남도 해남군 송지면 557'	{'pnumber': '010-5413-2509', 'name': '김정식', 'address': '전라남도 해남군 송지면 557'}
Model 4	{'pnumber': '010-5413-2509', 'name': '김정식', 'address': '전라남도 해남군 송지면 557'}	{'pnumber': '010-5413-2509', 'name': '김정식', 'address': '전라남도 해남군 송지면 557'}
Model 5	{'pnumber': '010-5413-2509', 'name': '김정식', 'address': '전라남도 해남군 송지면 557'}	{'pnumber': '010-5413-2509', 'name': '김정식', 'address': '전라남도 해남군 송지면 557'}

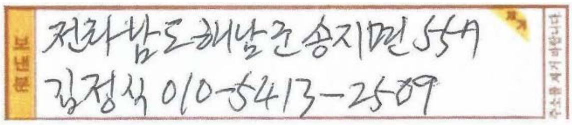


Figure 6: Testing image used as input for qualitative evaluation

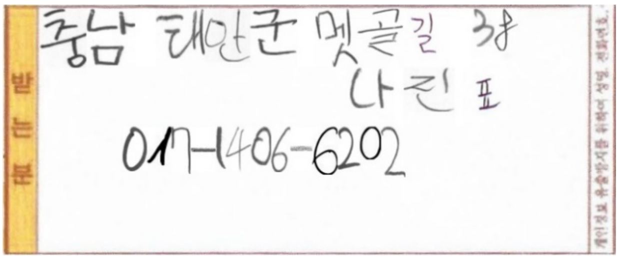


Figure 7: Training image used as input for qualitative evaluation

Table 4: Estimation results for training image of each model

	Raw output	Output
Model 1	{'pnumber': '017-1406-6202', 'name': '나진표', 'address': '충남 태안군 뫼골길 38'}	{'pnumber': '017-1406-6202', 'name': '나진표', 'address': '충청남도 태안군 멧골길 38'}
Model 2	{'pnumber': '017-1406-6202', 'name': '나진표', 'address': '충남 태안군 뫼골길 38'}	{'pnumber': '017-1406-6202', 'name': '나진표', 'address': '충청남도 태안군 멧골길 38'}
Model 3	{'pnumber': '017-1406-6202', 'name': '나진표', 'address': '충남 태안군 맷골길 38'}	{'pnumber': '017-1406-6202', 'name': '나진표', 'address': '충청남도 태안군 멧골길 38'}
Model 4	{'pnumber': '017-1406-6202', 'name': '나진표', 'address': '충남 태안군 맷골길 38'}	{'pnumber': '017-1406-6202', 'name': '나진표', 'address': '충청남도 태안군 멧골길 38'}
Model 5	{'pnumber': '017-1406-6202', 'name': '이민윤', 'address': '충청남 태안군 맷골길 38 나진포'}	{'pnumber': '017-1406-6202', 'name': '이민윤', 'address': '충청남 태안군 멧골길 38 나진포'}

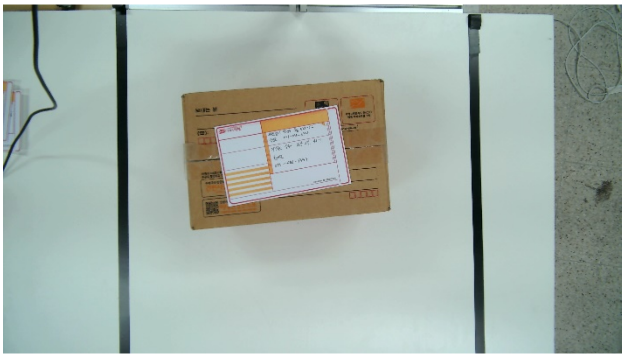


Figure 8: Box image captured by camera

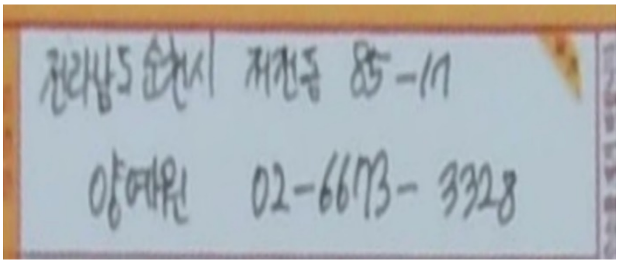


Figure 9: Captured image used as input for qualitative evaluation

Table 5: Estimation results for captured image of each model

	Raw output	Output
Model 1	{'pnumber': '02-6673-3328', 'name': '양예원', 'address': '전라남 5 순천시 정전동 85-17'}	{'pnumber': '02-6673-3328', 'name': '양예원', 'address': '전라남도 순천시 저전동 85-17'}
Model 2	{'pnumber': '02-6673-3328', 'name': '양예원', 'address': '전라남도 순천시 저전동 85-17'}	{'pnumber': '02-6673-3328', 'name': '양예원', 'address': '전라남도 순천시 저전동 85-17'}
Model 3	{'pnumber': '02-6673-3328', 'name': '양예원', 'address': '전라남도 순천시 저전동 85-17'}	{'pnumber': '02-6673-3328', 'name': '양예원', 'address': '전라남도 순천시 저전동 85-17'}
Model 4	{'pnumber': '02-6673-3328', 'name': '양예원', 'address': '전라남도 순천시 저전동 85-17'}	{'pnumber': '02-6673-3328', 'name': '양예원', 'address': '전라남도 순천시 저전동 85-17'}
Model 5	{'pnumber': '02-6673-3328', 'name': '양예원', 'address': '전라남도 순천시 저전동 85-17'}	{'pnumber': '02-6673-3328', 'name': '양예원', 'address': '전라남도 순천시 저전동 85-17'}

4.3 Discussion

Accuracy measurements were conducted using five different data combinations. Model 1 was exclusively trained using the generated data, and it yielded a significantly low accuracy. This indicates that training solely on the generated data fails to adequately capture the diversity and complexity of real-world data. The models that incorporated at least 25 % of the generated data exhibited considerable improvements in terms of accuracy. In particular, when the proportion of generated data exceeded 50%, the accuracy of the model tended to plateau. This suggests that a suitable balance between the generated and collected data may have a positive impact on the model performance.

From the prediction results for all models, the effectiveness of post-processing was confirmed through comparison of the initially predicted content with the final output. The models that included data generated by 25 % or more exhibited accurate and consistent outcomes, affirming the positive impact of a balanced combination of generated and collected data on model performance. These findings mark a crucial step toward overcoming the limitations inherent in conventional approaches based solely on generated datasets, particularly in applications such as handwriting recognition, by enhancing dataset diversity and real-world applicability. A balanced amalgamation of generated and collected data enables models to effectively handle various scenarios that may arise in real-world settings, thereby broadening their applicability. Finally, this study substantiates the impact of dataset composition on the performance of artificial intelligence models, which highlights the significance of blending generated and collected data to enhance model accuracy and generalization capabilities. This underscores the importance of these insights as critical guidelines for future model development and optimization strategies.

5. Conclusion

This paper delves into the research on a donut-based handwritten invoice-recognition system. The proposed system integrates image-processing algorithms to accurately crop areas containing information within images and further trains the donut using both the generated and collected data, enabling the digitization of images captured using a camera. In particular, by presenting the results trained with various ratios of generated and collected data weights, this study demonstrates the effectiveness of incorporating collected data during fine-tuning and offers guidelines regarding the proportion of collected data. The future works may involve experiments on diverse datasets of varying scales and ratios to explore methods for enhancing the performance during fine-tuning.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00092, Development of core technology for delivery and parcel reception automation system using cutting-edge DNA and interaction technology)

Author Contributions

Conceptualization, Y. S. Han and H. I. Seo; Methodology, Y. S. Han; Software, H. I. Seo; Validation, Y. S. Han and H. I. Seo; Formal Analysis, Y. S. Han; Data Curation, H. I. Seo; Writing—Original Draft Preparation, Y. S. Han; Writing—Review & Editing, H. I. Seo; Visualization, D. H. Seo; Supervision, Corresponding Author; Project Administration, D. H. Seo.

References

[1] C. W. Chen, Y. H. Cheng, T. Y. Lee, S. W. Chuang, P. Y. Hsu, S. P. Tseng, and J. F. Wang, “Automatic telephone interview survey system,” 2021 9th International Conference on Orange Technology (ICOT), IEEE, pp. 1-4, 2021.

[2] S. Saiki, N. Fukuyasu, K. Ichikawa, T. Kanda, M. Nakamura, S. Matsumoto, S. Yoshida and S. Kusumoto, “A study of practical education program on ai, big data, and cloud computing through development of automatic ordering system,” In 2018 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD), IEEE, pp. 31-36, 2018.

[3] F. Gao and X. Su, “Omnichannel service operations with online and offline self-order technologies,” Management Science, vol. 64, no. 8, pp. 3595-3608, 2018.

[4] M. Chung, “Smart ordering application for assigning sequence numbers to customers at offline sites,” In ITM Web of Conferences, EDP Sciences, vol. 24, 2019.

[5] T. Y. Wen and M. I. P. Mohamed, “Human and technology interaction: Consumer perception toward the touch screen ordering kiosk in fast food restaurant,” Research in Management of Technology and Business, vol. 3, no. 2, pp. 328-343, 2022.

[6] J. Memon, M. Sami, R. A. Khan, and M. Uddin, Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR), IEEE Access, vol. 8, pp. 142642-142668, 2020.

[7] S. D. Connell and A. K. Jain, “Template-based online character recognition,” Pattern Recognition, vol. 34, no. 1, pp. 1-14, 2001.

[8] I. K. Pathan, A. A. Ali, and R. J. Ramteke, “Recognition of offline handwritten isolated Urdu character,” Advances in Computational Research, vol. 4, no. 1, pp. 117-121, 2012.

[9] M. T. Parvez and S. A. Mahmoud, “Offline Arabic handwritten text recognition: A survey,” ACM Computing Surveys (CSUR), vol. 45, no. 2, p. 1-35, 2013.

[10] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, “LayoutLM: Pre-training of text and layout for document image understanding,” In Proceedings of the 26th ACM

SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1192-1200, 2020.

[11] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, and L. Zhou, “LayoutLMv2: Multi-modal pre-training for visually-rich document understanding,” Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 2579-2591, 2020.

[12] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, “LayoutLMv3: Pre-training for document AI with unified text and image masking,” In Proceedings of the 30th ACM International Conference on Multimedia, pp. 4083-4091, 2022.

[13] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, “Docformer: End-to-end transformer for document understanding,” In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 993-983, 2021.

[14] S. Appalaraju, P. Tang, Q. Dong, N. Sankaran, Y. Zhou, and R. Manmatha, “DocFormerv2: Local features for document understanding,” arXiv preprint arXiv:2306.01733, 2023.

[15] K. Lee, M. Joshi, I. R. Turc, H. Hu, F. Liu, J. M. Eisenschlos and K. Toutanova, “Pix2struct: Screenshot parsing as pretraining for visual language understanding,” In International Conference on Machine Learning, pp. 18893-18912, 2023.

[16] Z. Tang, Z. Yang, G. Wang, Y. Fang, Y. Liu, C. Zhu, and M. Bansal, “Unifying vision, text, and layout for universal document processing, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19254-19264, 2023.

[17] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, and *et al.*, “OCR-free document understanding transformer,” In European Conference on Computer Vision, pp. 498-517, 2022.

[18] Y. B. Hamdan and A. Sathesh, “Construction of statistical SVM based recognition model for handwritten character recognition,” Journal of Information Technology and Digital World, vol. 3, no. 2, pp. 92-107, 2021.

[19] A. Gupta, A. Vedaldi, and A. Zisserman, “Synthetic data for text localisation in natural images,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2315-2324, 2016.

[20] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, “Recognizing text with perspective distortion in natural scenes,” In Proceedings of the IEEE International Conference on Computer Vision, pp. 569-576, 2013.

[21] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, “Multi-oriented text detection with fully convolutional networks,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4159-4167, 2016.

[22] F. Borisyuk, A. Gordo, and V. Sivakumar, “Rosetta: Large scale system for text detection and recognition in images,” In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 71-79, 2018.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9992-10002, 2021.

[24] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871-7880, 2020.