Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(10): 18630–18649.
DOI: 10.3934/mbe.2023826
Received: 28 June 2023
Revised: 17 September 2023
Accepted: 19 September 2023
Published: 07 October 2023

Research article

A visual transformer-based smart textual extraction method for financial invoices

Tao Wang[1,*] **and Min Qiu**[2]

[1] School of Innovation and Entrepreneurship, Zhengzhou University of Science and Technology, Zhengzhou 450064, China

[2] Institute of Business Administration, Zhengzhou University of Science and Technology, Zhengzhou 450064, China

* **Correspondence:** Email: wangtao@zit.edu.cn.

Abstract: In era of big data, the computer vision-assisted textual extraction techniques for financial invoices have been a major concern. Currently, such tasks are mainly implemented via traditional image processing techniques. However, they highly rely on manual feature extraction and are mainly developed for specific financial invoice scenes. The general applicability and robustness are the major challenges faced by them. As consequence, deep learning can adaptively learn feature representation for different scenes and be utilized to deal with the above issue. As a consequence, this work introduces a classic pre-training model named visual transformer to construct a lightweight recognition model for this purpose. First, we use image processing technology to preprocess the bill image. Then, we use a sequence transduction model to extract information. The sequence transduction model uses a visual transformer structure. In the stage target location, the horizontal-vertical projection method is used to segment the individual characters, and the template matching is used to normalize the characters. In the stage of feature extraction, the transformer structure is adopted to capture relationship among fine-grained features through multi-head attention mechanism. On this basis, a text classification procedure is designed to output detection results. Finally, experiments on a real-world dataset are carried out to evaluate performance of the proposal and the obtained results well show the superiority of it. Experimental results show that this method has high accuracy and robustness in extracting financial bill information.

Keywords: computer vision; textual extraction; visual transformer; smart recognition

1. Introduction

In the era of big data, every individual is surrounded by a huge amount of financial information [1]. From small individuals to make accurate judgments with the help of financial news when investing [2].

For large enterprises, when managing internal financial data, they need to be able to efficiently extract the information in bills and store and maintain them [3]. How to effectively use and manage this financial information has become an important issue in front of us [4]. In the real financial reimbursement work, due to the increasing number of invoices to be reimbursed, the average number of invoices generated by each enterprise is nearly 10,000 per day. The workload of the finance department without using any automatic identification system for reimbursement has increased sharply, spending a lot of time every day on the work of organizing and filing invoices [5]. In recent years, deep learning has made certain achievements in various industries and has also made certain breakthroughs in the fields of text extraction. With the increasing computing power of computers, the scale of data processed by deep learning has been expanding [6].

Intuitively, deep learning can be utilized to quickly extract texts from financial invoices instead of human labors [7]. This will greatly reduce enterprise labor costs, shorten invoice processing time and improve the accuracy of identification results, greatly enhancing the maintainability of financial systems and safeguarding the security of enterprise funds. Image recognition, as the original intention of the development of computer vision, is one of the most fundamental and longest-developed applications in the field [8]. Traditional image recognition aims at a coarse-grained classification of the objects present in a picture and wants the results to be as close as possible to human recognition accuracy [9]. Unlike traditional image recognition, fine-grained image recognition aims at distinguishing different subclasses of a major class of objects to be recognized [10]. To solve these problems, an efficient automatic bill recognition processing system is urgently needed to connect the above processes of bill recognition to structured data in series. Then, intermediate processing links are optimized and streamlined as much as possible and fault-tolerant verification of key data are performed to achieve the production-ready requirements [11].

This paper proposes an automatic extraction method of financial bill information based on visual transformer. The method combines image processing techniques and sequential transduction models to achieve high-accuracy and robust bill information extraction. In this approach, we use a large-scale dataset with different input data types and process multiple text fields. The relevant quantitative parameters and comparison results are listed in the table below. By comparing with other methods, our method achieves excellent performance in terms of accuracy. Additionally, our dataset is relatively large in size and able to handle multiple text fields. The above results show that our proposed visual transformer-based automatic extraction method of financial bill information has broad potential in applications. This paper intends to use current outstanding hardware devices as the source of image data ingestion, which can be high-speed scanners or equally inexpensive cell phone handheld devices. This paper nextly focuses on the design and implementation of software algorithms and architectures, aiming at forming an overall common processing framework through effective algorithm design for the complex and diverse bill processing processes of current enterprises. In other words, the whole processing process is abstracted into a standard process of converting optical character recognition (OCR) to structured data storage. New or additional ticket templates can be flexibly configured to re-read the training set and train the model [12].

The attention mechanism is a neural network layer that aggregates information from the entire input sequence. Transformer introduces a self-attention layer that scans each element of the sequence and updates these elements by aggregating information from the entire sequence. Transformer is proved to be very effective in sequence processing, where the key lies in modeling the relationships between

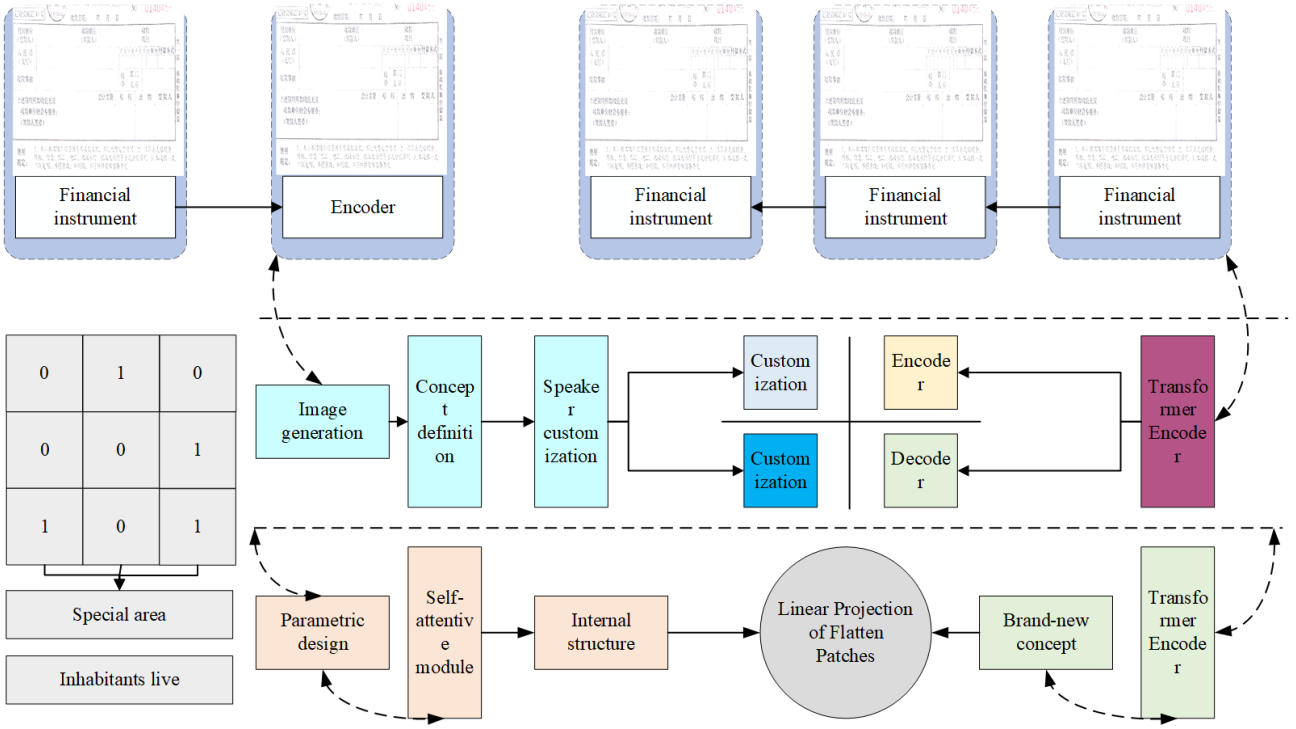


Figure 1. The detailed structure diagram of transformer used in this article.

input words. In this paper, we hypothesize that the capability of the transformer can be adapted to small-sample learning because such relationships between sparse data can significantly help the network to perform classification. The attention of long experience accumulated in the memory of network parameters in a transformer can better improve the network performance. To improve the accuracy of few-sample image classification and to explore the wider application of transformer in computer vision, this paper proposes to introduce transformer into few-sample learning with results.

2. Related works

The main technical means of automatic processing of bills are image classification, image preprocessing, image region of interest positioning, OCR and other technologies. By combining the above technologies in a reasonable order, the problem of automatic bill recognition processing can be effectively solved [13]. Overall, the layout of a bill can be considered as two parts: the fixed information field and the variable information field [14]. Most of the bills to be identified belong to the form-based bills, where the valid information areas are circled form box lines and the information is organized variably, either in a left-right structured information correspondence or in a top-down structured information correspondence [15].

A few other tickets are of the simple key-value pair type, which is generally banded, and very few card-type tickets, which have a large degree of typographical freedom and no explicit box-line constraints [16]. In addition to the above characteristic information, the note layout has additional auxiliary information that is helpful for information extraction and type identification [17]. The 2D

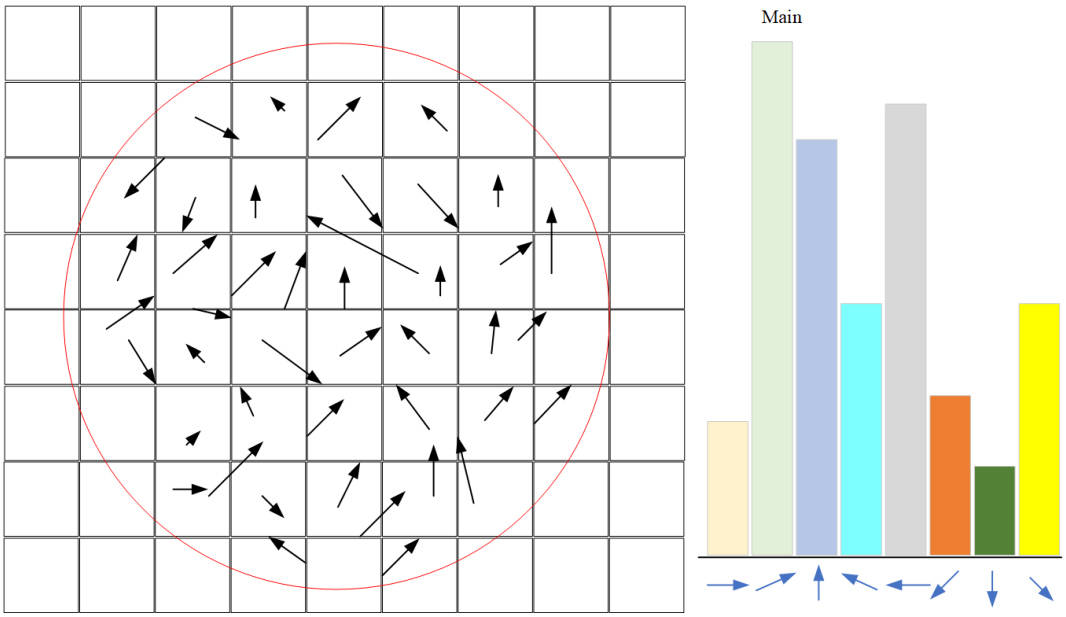


Figure 2. Schematic diagram of the main direction statistics of features.

code and barcode carry information by themselves and generally the code on the layout of the bill with a relatively short issuance time can still identify a lot of useful information. On the contrary, if the issuance time is relatively long, optical character recognition is required for a lot of layout content information to extract it [18]. In most cases, the code-like pattern can be used as an auxiliary tool for calibration, for example, we can infer the current posture information of the ticket based on the position of the code on the ticket layout so that the ticket can be targeted directly by certain morphological transformation means [19].

Non-amount data, such as transaction details, are usually not overly demanding and can allow for a small number of typographical errors if they do not affect the overall understanding and recognition [20]. If there is an imbalance between assets and liabilities, it is necessary to re-check and enter the relevant amount transaction fields to achieve a balance between debits and credits. Therefore, the identification of the amount fields requires extra attention [21]. An effective fine-grained stream distillation method is then explored, which can compute images, patches and randomly selected streams in both teacher and student models. A novel visual transformer (ViT) model compression framework is proposed that jointly reduces the redundancy of attention head, neuron and sequence dimensions. The authors propose a pruning criterion based on statistical dependencies that can be generalized to different dimensions to identify harmful components [22].

In addition, the authors use multidimensional compression as an optimization to learn the best pruning strategy across three dimensions to maximize the accuracy of the compression model under the computational budget. An efficient super-resolution transformer is proposed, which is a hybrid transformer consisting of a lightweight convolutional neural network backbone and a lightweight transformer backbone [23]. Where the lightweight convolutional neural network backbone advances the deep super-resolution features at a lower computational cost by dynamically resizing the feature

maps. And the lightweight transformer backbone occupies very little memory and proposes an internal feature segmentation module to split long sequences into subsegments, which can significantly reduce the memory occupation.

This subsection first introduces the current state of development of transformer-based image feature extraction methods and summarizes them in a well-developed manner. Then, current work on accelerating the transformer is presented and representative methods are described. Most of these accelerated models need to introduce additional parameters to determine which units are redundant or require a more complex design to achieve a lightweight transformer. Second, part of the transformer-based image feature extraction methods lack a systematic understanding of the nature of attention and are not comprehensive enough. Based on the E-Attention proposed in this paper, we further combine depth convolution and null convolution to introduce the inductive bias lacking in the transformer model from three perspectives: translation invariance, localization and scale invariance. Then, a lightweight convolution module is used to change the way the traditional transformer model processes the input images to speed up the convergence speed and improve the stability. The result is an efficient transformer image feature extraction network (CEFormer) combined with convolution. Experiments show that CEFormer achieves good results between performance and computing speed.

3. Methodology

3.1. Data preparation

It is required to build a data set containing large-scale financial bill images and corresponding annotations. These images can be scans, photos or electronic documents. The annotation information covers the location and content of each field in the bill, such as date, amount, payee, etc. Preprocessing of financial instrument images using computer vision techniques. This may include operations such as image enhancement, size normalization and denoising to improve the accuracy and robustness of subsequent information extraction. The processed image is converted into a feature vector representation using a convolutional neural network (CNN). CNN models can be trained to semantically encode images and capture important features related to bill information. The transformer architecture is used as the sequence transduction model and the feature vector of the image is used as input. The goal of this model is to generate sequence labels for each field in the ticket. Use labeled data sets for model training and optimization. The model parameters are adjusted by minimizing the loss function so that it can accurately predict the location and content of the note fields. In a practical application, the trained model is applied to new financial note images. By inputting images into the model, label sequences for corresponding fields can be generated, allowing automatic extraction of key information.

3.2. Feature extraction

Transformer is a deep neural network based on a self-attentive mechanism while being able to process data in parallel [24]. Originally a concept in the field of natural language processing, the transformer has been introduced to the field of computer vision in the last year or two. Based on this, this subsection first introduces the attention mechanism, then describes the structure of the transformer in the original natural processing domain and finally describes the transformer for processing vision tasks. Taking a neural network trained for classification as an example, a picture is an input to the network, and then the

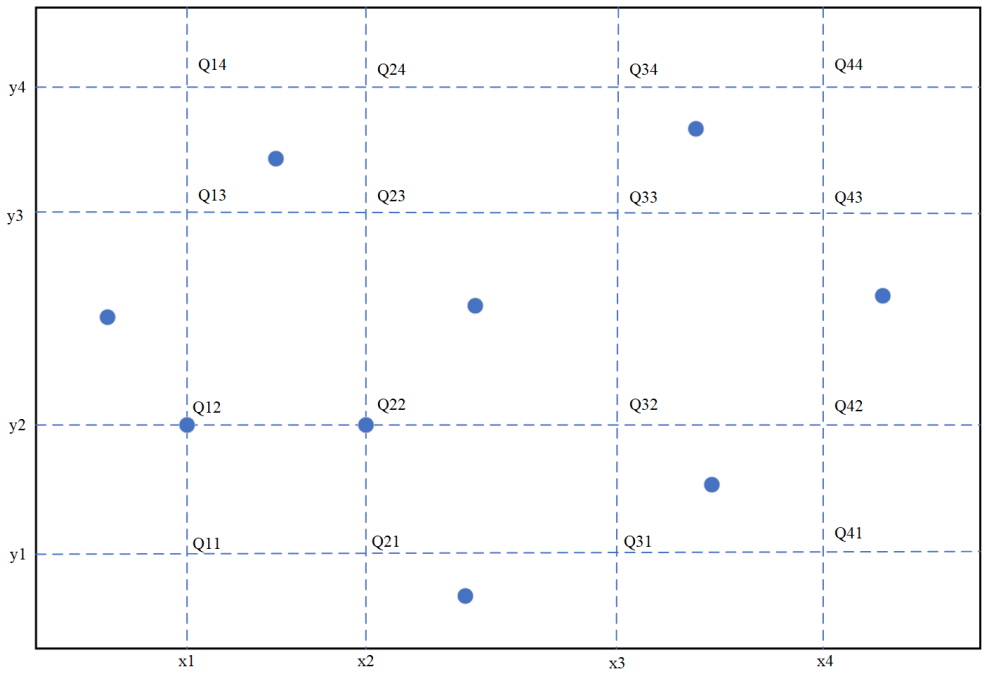


Figure 3. A relatively simple diagram of bilinear interpolation.

weights W and input X in the network are computed by attention to extract features in the input that are beneficial to the classification of the neural network, and the extracted features can be used as the basis for the final network to determine the category. The weights W and input X are both matrices, and to achieve the purpose of using W to reweight X, it can be seen as the first point multiplying W and X, calculating the similarity between them and then converting them into weight probability distributions and finally acting on X. The calculation process is shown as:

$$Att(Q, K, V) = Softmax(QK^T / \sqrt{d_k})V \tag{3.1}$$

The Q in Eq (3.1) is the trained W matrix in the above classification network and K is the image input X, V and K are equal. Equation (3.1) can be interpreted as first calculating the similarity between the query matrix Q and matrix K, then converting them into probability distributions using the Softmax operator and then right multiplying the obtained probability distributions by the matrix V to achieve the weighting of the matrix V using the attention weight distribution [25]. Transformer is proposed to solve machine translation tasks. Machine translation is understood as a sequence-to-sequence problem, i.e., a seq2seq structure, for which an encoder-decoder structure is generally used to solve, and the transformer follows the encoder-decoder structure shown as Figure 1. The transformer model mainly consists of an encoder and a decoder, where the encoder includes a self-attentive module and a feed-forward neural network, while the decoder has an internal structure like the encoder but with an additional cross-attentive module that interacts with both the encoder and the decoder. In general, the standard transformer model has 6 encoders and decoders arranged serially.

First, the general form of the transformer is described mathematically, given that the input is represented in the embedding space as $X \in R^{(n \times d)}$, and the transformation that the input undergoes into

the transformer module is denoted as T. The transformation T is thus defined as:

$$T(X) = F(Att(X) - X) \tag{3.2}$$

where F is the feedforward neural network containing the residual connections; Att is the function used to compute the self-attention matrix. The self-attention mechanism in the current transformer model is scaled dot product attention. For the convenience of the narrative, the scaling factor of Att is hidden and defined as:

$$Att(Q, K, V) = Softmax(QK^T)V \tag{3.3}$$

where $Q \in R^{(n\times d_q)}$,$Q \in R^{(n\times d_k)}$,$Q \in R^{(n\times d_v)}$, respectively, can be calculated from the input.

In-depth analysis of above formula, we can find that the performance of the scaled dot product attention mechanism, so that its complexity is quadratic level is the Softmax operator in the definition of above formula. Therefore, it is necessary to first calculate (QK^T). This step is an n * n matrix, and thus the complexity is $O(n^2)$level, if there is no Softmax operator, then it is $(QK^TV)V$. If there is no Softmax operator, then it is QK^T three matrices multiplied together, and by using the combination law of matrix multiplication, we can calculate the product of the last two matrices to get a $d_q \times d_v$ matrix, and then let the matrix Q go to the left multiplication because in the actual case, d_k and d_v are much smaller than n , so the overall complexity can be seen as $O(n^2)$ level:

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{C=1}^{C} e^{2z_C}} \tag{3.4}$$

where z_i is the output value of the i-th node and C is the number of output nodes, i.e., the number of categories classified. The Softmax algorithm converts the output values of multiple classifications into a probability distribution ranging from 0 to 1 and sums to 1.

$$Att(Q, K, V) = \frac{\sum_{i=1}^{n} e^{2z_C q_i^T k_j} V_j}{\sum_{i=1}^{n} e^{q_i^T k_j c}} \tag{3.5}$$

where q_i is the i-th column vector of Q , jk is the j-th column vector of K and V_j is the j-th column vector of V. According to the above formula, the generalized form of the attention mechanism is proposed by replacing the exponential form in the above formula with the general function $S(\cdot)$ on q_i and V_j , which is defined as:

$$Att(Q, K, V) = \frac{\sum_{i=1}^{n} S(q, k) e^{2z_C q_i^T k_j} V_j}{\sum_{i=1}^{n} S(q, k) e^{q_i^T k_j c}} \tag{3.6}$$

The input matrix is the matrix of all image slice vectors after the input embedding and position encoding, in which "num patches" is the row dimension of the above matrix, representing the number of image slices; the number of columns d is the dimension of each image slice vector, which is also the dimension of the is the dimension of the post-order transformer model [26]. The input matrix is multiplied with the weight matrix WQ to obtain the linearly varied matrix Q. The matrix V is generated in the same way as the matrix K. The difference is that the linear matrices used are WK and WWV and the specific calculation of Q , K and V can be expressed by the following equation:

$$Q = XW_Q \tag{3.7}$$

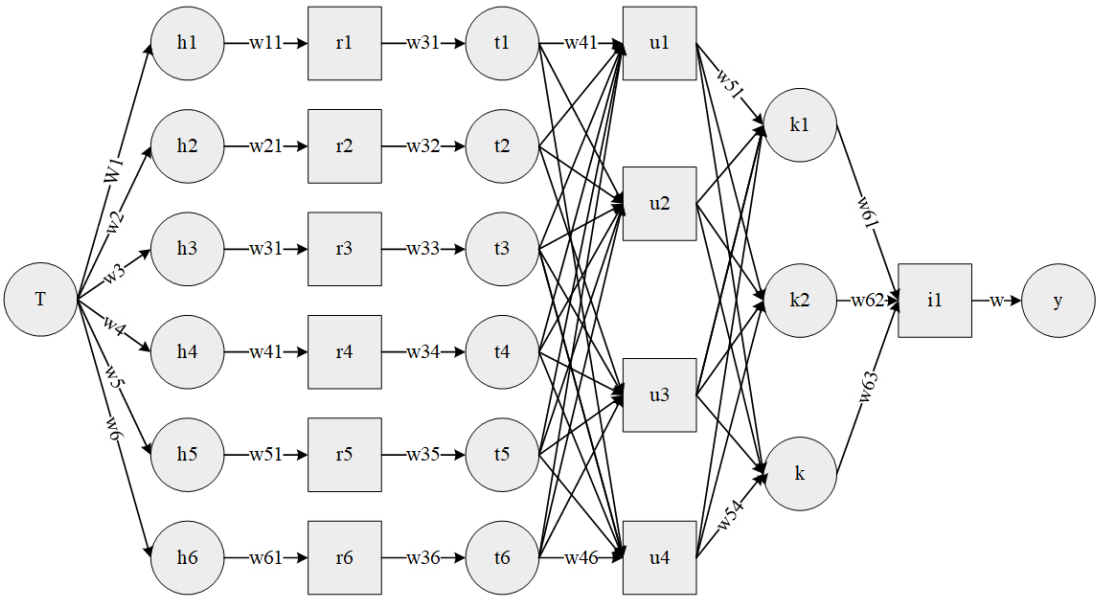


Figure 4. Detailed block diagram for important parts of the neural architecture.

$$K = XW_K \tag{3.8}$$

$$V = XW_V \tag{3.9}$$

where X is the input matrix. After linear variation of the input sequence, the output of the self-attentive mechanism can be calculated by the following equation:

$$Att(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.10}$$

The softmax-activated QKV matrix is equivalent to the attention map of the correlation between image slices, from which each row represents the weight of the relationship between the image slice at the current position and the other image slices in the sequence and the weights add up to 1. Multiplying this weight matrix with V finally obtains the feature expression of the sequence the weight matrix is multiplied by V to obtain the weighted feature expression of the image slice in the sequence, which not only extracts the semantic and spatial features of the current position image but also incorporates the information of other image slices through the self-attentiveness mechanism [27]. After understanding how the above self-attentive mechanism works, the multi-headed self-attentive mechanism is a network of several identical self-attentive mechanism modules in parallel and the output of each self-attentive mechanism module is finally combined as the output of this multi-headed self-attentive mechanism layer.

$$MultiHead(Q, K, V) = Concat(head_1, head_2, ..., head_n)W^T \tag{3.11}$$

Schematic diagram of the main direction statistics of features is shown as Figure 2. Once the magnitude and angle of the feature point neighborhood are calculated, the next step is to determine which direction is the main direction of the feature point. The method used here is histogram statistics, in which all pixels about the Gaussian image are placed in a coordinate system with the direction as the

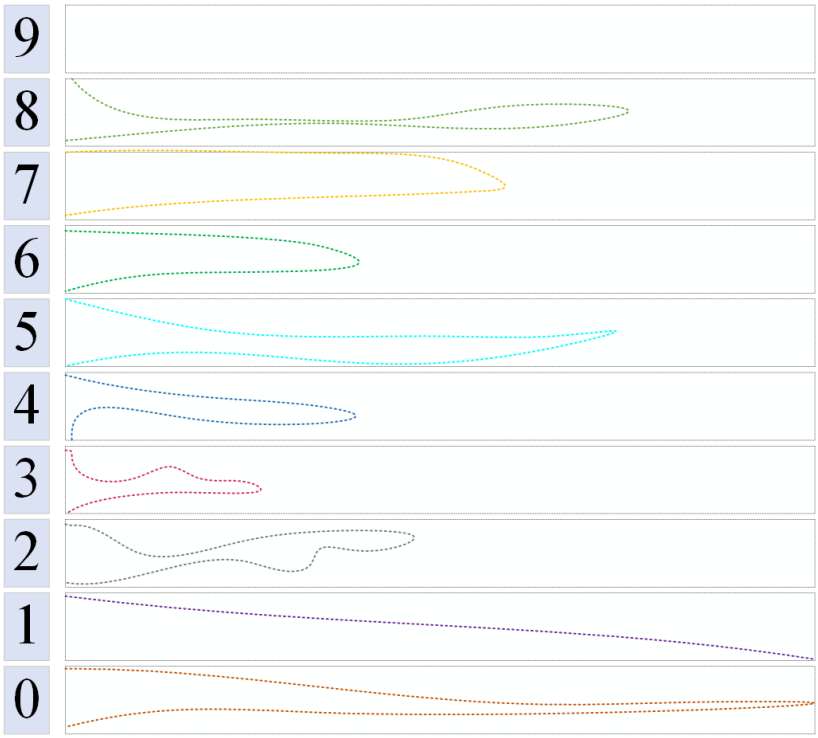


Figure 5. Vertical projection of the number of images.

horizontal coordinate and the amplitude accumulation as the vertical coordinate and finally the direction with the largest amplitude is the dominant direction.

$$\theta(x, y, z) = arctan\frac{L(x, y+1) - L(x, y) + L(x, z+1)}{L(x, y+1) + L(x, z+1)} \tag{3.12}$$

In this stage, we follow a meta-training paradigm [28] to fine-tune the backbone. The approach in this paper uses transformer to fine-tune the backbone to uncover potential relationships among samples from different categories [29]. Meta-learning [28] requires learning the characteristics of each task by working on different small tasks and then generalizing the commonality of all tasks. So, we construct different minis from the dataset at pre-training and construct a new loss function used to optimize the network. In each task, all images are first input into the pre-trained feature extractor, then the feature vectors from the obtained support set are input into the transformer structure and new feature vectors are obtained. After that, the feature vectors from the query set are used to compare with them and the corresponding loss functions are generated to optimize the network.

3.3. Textual recognition

Text detection, the first step of transformer technology, is used to analyze the pixels in an image to get the area containing text and display it visually or crop the text area into a separate image containing only text lines. Since text detection is an indispensable step for text recognition, an excellent text detection tool can largely affect the accuracy of subsequent text recognition [30]. The focus of this paper is on the subsequent text detection and text classification stages. So, the parameters are partially modified and

optimized based on existing detection methods to achieve optimal detection results in the text detection stage.

The dataset consists of 1740 images of real tickets, including food and beverage tickets, shopping tickets, bus tickets, etc. Only one ticket exists in each image and all the information of the ticket is captured completely and the language of the ticket is English. A total of 68,975 text boxes are provided, with an average of 39 text boxes per image. These text boxes are first labeled with the location information in the original image. Then, each text box is labeled with the corresponding text content. Finally, all text boxes are labeled into 25 different categories. The original 1740 images in the wild receipt dataset are divided into two non-intersecting copies, 1268 and 472 images, for training and testing respectively to ensure that the images in the test set are not used in training. The robustness of the model training results can be tested by differentiating the image file names and image retrieval methods to ensure that the ticket image layouts in the training and test sets are largely non-crossover, i.e., the layouts that appear in the training set are largely absent in the test set.

It is expected to gather a representative dataset of financial invoices that need to be processed for reconstruction. The dataset should include a diverse range of invoice formats, layouts and styles. Clean and preprocess the invoices to ensure consistent formatting and remove any noise or irrelevant information. This may involve tasks such as image normalization, resizing and text extraction. The proposed model is trained using the preprocessed dataset. It is required to define appropriate labels for the textual information, such as vendor names, invoice numbers, dates and amounts. Fine-tuning the model on the specific reconstruction domain can further enhance its performance. Integrate the trained model into the existing reconstruction system. This may involve developing APIs or incorporating the model into the system's backend infrastructure.

It is expected to ensure compatibility and seamless communication between the model and other components of the system. It is expected to utilize the integrated model to automatically extract relevant textual information from incoming financial invoices. The model can identify and extract critical fields, such as vendor details, invoice numbers, item descriptions and amount. This automation can significantly reduce manual data entry and minimize errors. Implement checks and validation mechanisms to ensure the accuracy of the extracted information. The model's outputs are compared with ground truth data or perform cross-validation to measure its performance and identify any discrepancies or errors. User feedback and monitor the model's performance are continuously gathered in real-world scenarios. This feedback is incorporated to iterate on the model, fine-tune it and address any limitations or errors observed during operation. Measures are implemented to handle edge cases, exceptions and pot entail security vulnerabilities. The model can handle variations in invoice formats, handle missing or incomplete information gracefully and protect sensitive financial data.

Image scaling is the resizing of the original image to make it meet the requirements of subsequent model training. The scaling algorithm used in this paper is a bilinear interpolation algorithm. The bilinear interpolation algorithm can solve the problem of image distortion due to inaccurate interpolated coordinates caused by rounding of floating-point coordinates using the nearest neighbor interpolation algorithm [31]. The bilinear interpolation algorithm calculates the pixel value of a coordinate by taking the current coordinate as the center and calculating the pixel value of the new coordinate according to the pixel value of the four points adjacent to the horizontal and vertical coordinates and their distance from the current position as a percentage, as shown in Figure 3.

To find the value of the location of point P, the values of the four nearby points $Q11$, $Q12$, $Q21$ and

$Q22$ are calculated as:

$$f(P) \approx frac f(Q_{11})L(x, y+1)L(x, y) - \frac{f(Q_{21})}{L(x, y+1)}L(x, y) + \frac{f(Q_{12})}{L(x, y+1)}L(x, y) - \frac{f(Q_{22})}{L(x, y+1)}L(x, y) \tag{3.13}$$

where $f(Q)$ denotes the pixel value at the Q-coordinate position. x and y denote the horizontal and vertical coordinates respectively. To cope with the fact that the images are not always horizontal due to the shooting angle, a random angle rotation is added to the training data to effectively improve the diversity of the data and the detection model can also handle the detection of text at various angles well. Here, the center of rotation is defined as the center of the image, i.e., $(w/2, h/2)$. If the rotation angle of the point $P_0(x_0, y_0)$ clockwise around the rotation center is θ, the coordinates of the rotated point can be obtained as:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} sin\theta & 0 & -sin\theta \\ 0 & sin\theta & 0 \\ -sin\theta & 0 & sin\theta \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} \tag{3.14}$$

The top-down approach is to generate $M2$, $M3$, $M4$, and $M5$ by up-sampling the higher-level feature maps and fusing them with the adjacent feature maps in a lateral connection. Finally, a 3×3-convolutional layer is used to convolve the summed and fused features to obtain the final feature maps $P2$, $P3$, $P4$ and $P5$, which correspond to $C2$, $C3$, $C4$ and $C5$ respectively and have the same feature size. Since the fusion between features in each layer of the feature pyramid needs to use the same feature dimension (number of channels), the number of convolution kernels of the corresponding convolution layers is set fixed to 256 in this paper.

$$IOU = inter(gt_i, det_j)/union(gt_i, det_j) \tag{3.15}$$

The detailed block diagram for important parts of the neural architecture is shown in Figure 4.

The Hough transformation algorithm can theoretically obtain the exact tilt angle but it is computationally intensive, time-consuming and requires more storage space. The projection-based method is to project the image to different angles, store the projection values obtained from different angles in a matrix, find the largest projection value in the matrix and then the tilt angle corresponding to this maximum value is the tilt angle of the image [32]. The projection-based method is computationally intensive and the maximum value found is often not necessarily the exact tilt angle. It is necessary to find the minimum value point by analyzing the set energy function and to find a reasonable tilt angle step by step from coarse to fine, as shown in Figure 5.

After the first step of roughly locating the number area, the next step is to locate the exact position of the ticket number. For the top and bottom of the numbered area, we scan from top to bottom, stop scanning as soon as we reach a black pixel in a certain line and write down the position X1. Then, we scan from the bottom upward, stop scanning as soon as we reach a black pixel and write down the position X2. Then, the roughly positioned image is intercepted according to the vertical coordinates of X1 and X2 and the top and bottom of the numbered area are intercepted [33]. For the positioning of the left and right sides of the number area, the vertical projection method is used because the left and right sides of the number sequence have an interval of background pixel points. Based on this feature, the number of black pixels on each column of the image of the previous step is counted vertically to obtain

the vertical projection map of the number image. By analyzing the vertical projection map, the starting positions of the numbers on the left and right sides of the image can be obtained and the left and right widths of the number sequence can be determined.

Electronic documents are stored in a structured manner in a database or file system, and provide functions for classifying and organizing according to folders, tags, attributes, etc. Fast document retrieval function is provided, allowing users to find the required documents through keywords, attributes or custom conditions. At the same time, permissions can be managed on documents to control the access permissions of different users or user groups. Supports version management of documents, records modification history of documents and avoids conflicts and loss. At the same time, it provides collaboration functions so that multiple users can edit and comment on documents at the same time to promote teamwork. Describe and manage document metadata, such as document name, author, creation date, associated projects, etc. This facilitates document classification, archiving and retrieval. Ensure document security, set up permission controls, encryption and prevent unauthorized access. At the same time, perform regular backups and disaster recovery to prevent document loss or damage. Supports the definition and management of document-related workflows, such as approval processes, sign-off processes, etc., to improve work efficiency and organizational collaboration capabilities.

4. Analysis of results

4.1. Performance evaluation analysis

This subsection compares the proposed methods in this paper with the existing excellent methods on a dataset collected from real-world financial scenes. It contains about 700 bill images, in which the textual information includes vendor names, invoice numbers, dates and amounts. In this paper, we first use 400 bill images as learning samples, and then send them to the pattern classifier for training by extracting forty-dimensional features through pre-processing, segmentation and normalization steps. Then, 300 bills are used as the samples to be tested and sent to the pattern classifier for recognition through the same preprocessing and feature extraction steps.

Figure 6 shows the experimental results of all compared methods and the results show that the Masked-AN method proposed in this chapter achieves the best results in all metrics. The three masked-AN algorithms are listed in Figure 6 and it can be observed that the sparse structure model masked-AN (Sparse) has better results among the two proposed individual models because the sparse structure graph is a semantic structure graph constructed based on the dependency between words, which enables the model to capture the association information between words more accurately. The information exchange with neighboring nodes makes the features more discriminative. The dense structure model masked-AN(dense) considers the information transfer process among all nodes in the graph structure, which is easy to cause ambiguity. In addition, the integrated model masked-AN(sparse+dense) has the best experimental results due to the complementary nature of the two separate models. Compared with the BFAN algorithm, the masked-AN algorithm improves the Recall@1 metric by 8.9% and 7.9% in the directions of picture query text and text query picture, respectively. The bidirectional focal attention network (BFAN) algorithm [34] also adopts the idea of suppressing irrelevant local information representation and enhancing relevant local information representation.

The graph structured network (GSMN) [35] and similarity reasoning and filtration (SGRAF) [36] algorithms use graph convolutional networks and graph attention envelopes for inference, respectively.

In addition, in the SGRAF algorithm, the overall similarity between the two modalities is calculated in addition to all the similarities between all local features. The results in the table show that the algorithm proposed in this chapter still outperforms the other algorithms in terms of retrieval accuracy index and verifies the effectiveness of similarity inference using the transformer structure, which is better than the two algorithms using the graph structure in terms of accuracy. The reasons for this are, on the one hand, the structural design of the transformer itself and, on the other hand, the use of an attention mechanism that allows the network model to learn more important similarity vector information.

The transformer model shuffle transformer was proposed because the global self-attentive mechanism used by traditional transformer models has a quadratic computational complexity concerning the number of input tokens. This makes ViT difficult to be applied to intensive prediction tasks such as semantic segmentation and object detection, which require high-resolution image input. In this paper, a new transformer structure is proposed, which combines the spatial shuffle and the window-based self-attentive mechanism to effectively establish the cross-window connection and enhance the expressiveness of the model. In addition, a deeply separated convolutional layer with residual connections is inserted between the window self-attentive module and the feedforward network and its convolutional kernel size is the same as the window size. This operator enhances the information flow between adjacent windows. The four properties are named $C1$, $C2$, $C3$ and $C4$ respectively and then the four properties are ablated separately and one of the properties of the ablation model is denoted as w/o. Full means that all properties are retained and the other experimental settings are unchanged.

In general, deep learning models for textual extraction can be compared based on their performance metrics such as precision, recall, F1 score and processing time. In terms of accuracy, state-of-the-art deep learning models for textual extraction have achieved high precision, recall and F1 scores on various datasets. For example, recent research has shown that models based on attention mechanisms, graph convolutional networks and transformer architectures can achieve state-of-the-art performance for textual extraction tasks such as named entity recognition (NER) and information extraction (IE). In terms of efficiency, the processing time of deep learning models for textual extraction is dependent on factors such as the size of the input data, the complexity of the model architecture, and the computational resources available. While deeper and more complex models may achieve better accuracy, they often require more computational resources and longer processing times. Therefore, when comparing the proposed model with other state-of-the-art deep learning models for textual extraction, it is important to consider both accuracy and efficiency metrics. It is also important to note that the performance of deep learning models can vary depending on the specific task and dataset being used.

As can be seen from Figure 7, the accuracy decreases to a certain extent when any of the characteristics is ablated, among which the accuracy decreases the most after ablating $C4$, i.e., removing stability, which shows that this characteristic increases the model the most. Next, the accuracy decreases by the same amount after ablating $C2$ or $C3$, i.e., after removing localization or scale invariance, which means that the increase of these two characteristics is the same. Finally, ablating $C1$, i.e., removing translational invariance, results in the smallest decrease in accuracy, indicating that the translational invariance is the least gaining of the four characteristics. In addition, it can be seen that the computation of the forward inference of the model decreases to a certain extent after ablating any of the characteristics, indicating that the model is faster because ablating any of the characteristics indicates that the introduction of a certain convolution into the model is reduced so the model speeds up and the computation of the forward inference decreases. And the ablation $C4$, i.e., after removing stability, reduces the computation

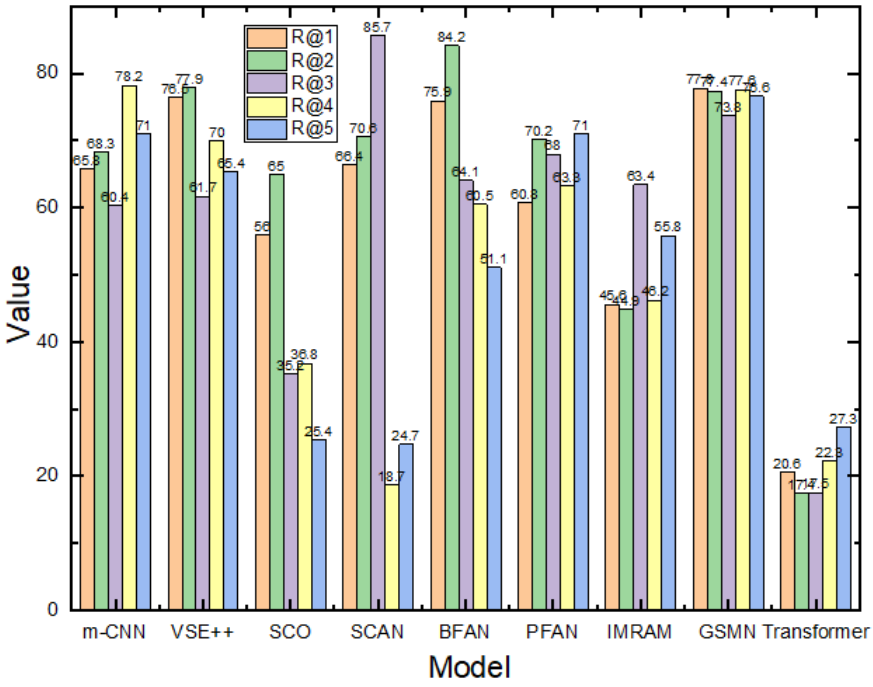


Figure 6. Comparison of the results of each algorithm on the dataset.

of forward inference of the model the most, which shows that this feature is the most burdensome to the model, but its corresponding accuracy improvement is also the largest. Thus, the two complement each other.

The first way to optimize the performance of the model is through transfer learning. One can use pre-trained models like VGG-16 or ResNet-50, which have been trained on large datasets and proven to yield good results on various image classification tasks. The pre-trained model can be fine-tuned for the specific task of information extraction from financial invoices. Data augmentation is another way to improve the model's performance. One can use techniques like rotation, flipping, zooming and cropping to generate new data points from the existing dataset. This increases the diversity of the dataset and enables the model to learn more robust features. Hyperparameters like the learning rate, batch size and optimizer settings have a significant impact on the performance of the model. By tuning these hyperparameters, one can find the optimal settings that enable the model to converge faster and achieve better results. Ensemble learning involves combining multiple models to improve performance. One can train several visual transformer-based models with different settings and merge their predictions to obtain an overall improved result. Attention mechanisms can help the model focus on important regions of the input image. By adding attention mechanisms to the visual transformer-based model, one can improve the ability of the model to identify and extract the relevant textual information from financial invoices.

4.2. Analysis of the results of financial instrument information extraction

To examine the generalization ability of the model, an additional 200 other bills are selected here as confusion samples, mainly to test the accuracy of the classification model in three important aspects. In

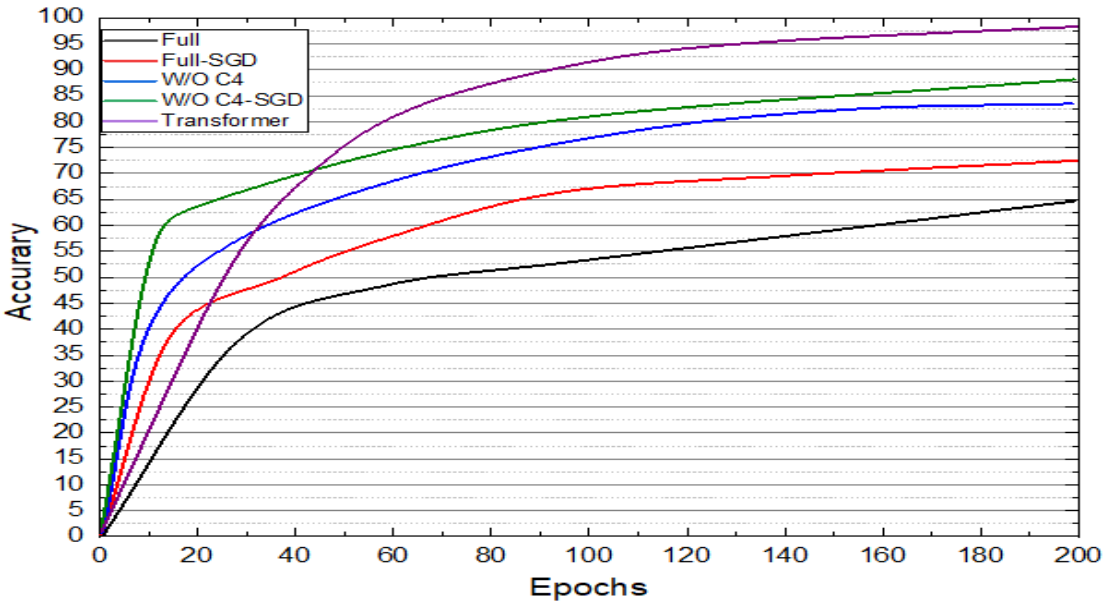


Figure 7. Experimental results of optimizer changes.

addition, the main types of bills tested are the more representative VAT invoices, whose information of interest includes basic invoice information, information on the buyer, seller, goods, taxable services and amount total information. Under the premise of testing 300 positive invoice samples, ignoring the time of system startup, a total of about 245 seconds is consumed and the average processing time of a single invoice is about 0.92 seconds and its various recognition accuracies are shown in Figure 8 below.

The character recognition accuracy rate reflects the ratio of the number of correctly recognized characters to the total number of recognized characters, which can reflect the situation of wrong recognition and multiple recognition but cannot reflect the situation of missing recognition; the character recognition recall rate reflects the ratio of the number of correctly recognized characters to the actual number of characters, which can reflect the situation of wrong recognition and missing recognition but cannot reflect the situation of multiple recognition. The average edit distance can reflect the situation of wrong recognition, missing recognition and multiple recognition at the same time. From the results, the comprehensive accuracy rate of the system for VAT invoice recognition is about 91.1%, and the same accuracy rate can be achieved for other types of invoices, which can meet the basic requirements for enterprise invoice recognition.

According to the above algorithm, the neuron that has the largest ranking number for a class of samples is the output neuron that can distinguish this class of samples. It is important to note that the number of linkage weights between the neuron and the neuron with the largest ranking does not exceed the number of classifiable categories. For neurons with a larger ranking, we only need to select the linkage weights between the neuron with the next largest ranking, i.e., the neuron with the smallest ranking among the neurons with a larger ranking. This acts as a mask for the neuron with a larger ordinal number. Adjusting each un from the basis of increasing the number of separable samples and increasing the distance of segmentation. The weights of the sorted neurons. From Figure 9, the text boxes overlap significantly and the text boxes are dense in areas. The reason for this phenomenon is that the amount of data processed by the text localization module is too large and it will be interfered

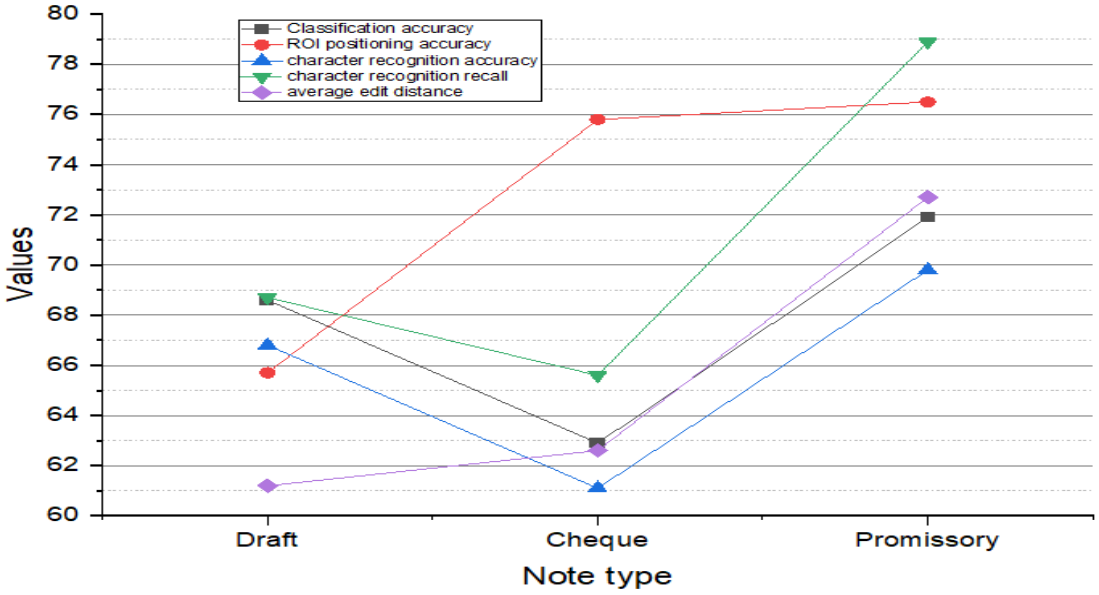


Figure 8. Chart for testing verification results of financial invoice.

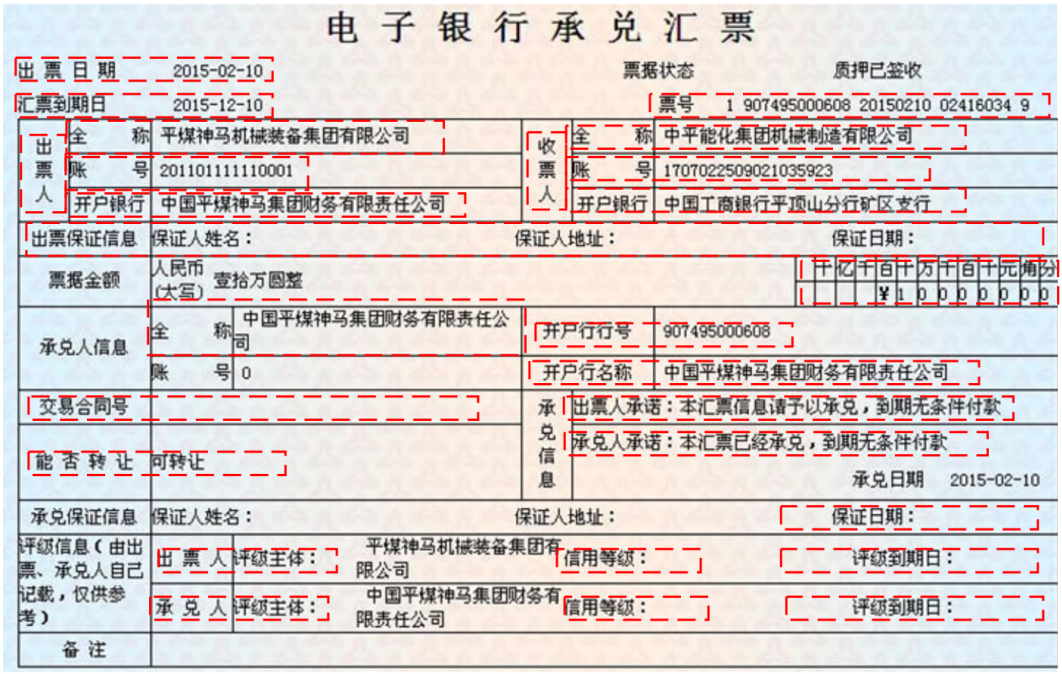
with by the surrounding data in the training process. If the text detection results are directly sent to the recognition module without any processing, the accuracy of the recognition results will be affected.

Because of this, this paper introduces the idea of partitioning and dividing the VAT invoice into several different regions by taking advantage of the characteristics of the VAT invoice layout itself, and then locating and recognizing each part after the division. The partitioning is mainly based on the a priori information and layout rules of each part. By dividing the regions, the text positioning module can reduce the amount of data to be processed simultaneously and avoid the interference of the surrounding irrelevant data. This paper also makes improvement to the text line post-processing algorithm for merging multiple small text boxes so that the text localization algorithm can accurately detect text lines with a skewed distribution.

5. Conclusions

In this paper, we apply deep learning to financial information extraction and prediction and do research on VAT invoice information extraction and mining the sentiment information behind financial news. For VAT invoice information extraction, this paper first does pre-processing on images to facilitate the use of subsequent sessions. Then, a text detection model with an improved text line processing algorithm is proposed based on the text localization model visual transformer. Next, text recognition algorithms based on introduction of a visual transformer are proposed and each method is compared on the dataset. Finally, a VAT invoice recognition interface is developed, integrating each module for easy visualization by users.

One of the innovations is that instead of detecting the text of the whole invoice layout, the invoice image is first divided into different parts according to the corresponding areas and the text is detected and recognized on the segmented parts. The improvement of the text line post-processing algorithm


电 子 银 行 承 兑 汇 票

出票日期 2015-02-10　　票据状态　　质押已签收

汇票到期日 2015-12-10　　票号 1 907495000608 20150210 02416034 9

出票人	全　称	平煤神马机械装备集团有限公司	收票人	全　称	中平能化集团机械制造有限公司
	账　号	201101111110001		账　号	1707022509021035923
	开户银行	中国平煤神马集团财务有限责任公司		开户银行	中国工商银行平顶山分行矿区支行
出票保证信息	保证人姓名：		保证人地址：		保证日期：
票据金额	人民币(大写)	壹拾万圆整			千 亿 千 百 十 万 千 百 十 元 角 分 ¥ 1 0 0 0 0 0 0 0 0
承兑人信息	全　称	中国平煤神马集团财务有限责任公司	开户行行号		907495000608
	账　号	0	开户行名称		中国平煤神马集团财务有限责任公司
交易合同号			承兑信息	出票人承诺：本汇票信息请予以承兑，到期无条件付款	
能否转让	可转让			承兑人承诺：本汇票已经承兑，到期无条件付款 承兑日期 2015-02-10	
承兑保证信息	保证人姓名：		保证人地址：		保证日期：
评级信息（由出票、承兑人自己记载，仅供参考）	出票人	评级主体： 平煤神马机械装备集团有限公司	信用等级：		评级到期日：
	承兑人	评级主体： 中国平煤神马集团财务有限责任公司	信用等级：		评级到期日：
备注					

Figure 9. The effect of text detection of the whole invoice.

also makes text detection applicable to the skewed case. Meanwhile, the trained recognition model is finetuned on another dataset, which strengthens the robustness of the system. There are still shortcomings in this paper on the feature extraction of printing numbers, the number of dimensions of the feature extraction is small, which leads to the situation that sometimes a part of numbers cannot be recognized. Therefore, future research can extract the features of the high dimension of the number to make this algorithm more perfect.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgement

This work was supported by 2022 Henan Provincial Key R&D and Promotion Special (Soft Science Research) Project under grant 222400410636.

We also would like to express sincere thanks to Zhengzhou University of Science and Technology, China. Because the university provides us good experimental conditions and platform support for our research works.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Y. Chen, C. Liu, W. Huang, S. Cheng, R. Arcucci, Z. Xiong, Generative text-guided 3d vision-language pretraining for unified medical image segmentation, preprint, arXiv:2306.04811. https://doi.org/10.48550/arXiv.2306.04811
2. Z. Wan, C. Liu, M. Zhang, J. Fu, B. Wang, S. Cheng, et al., Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias, preprint, arXiv:2305.19894. https://doi.org/10.48550/arXiv.2305.19894
3. C. Liu, S. Cheng, C. Chen, M. Qiao, W. Zhang, A. Shah, et al., M-FLAG: medical vision-language pre-training with frozen language models and latent space geometry optimization, preprint, arXiv:2307.08347. https://doi.org/10.48550/arXiv.2307.08347
4. Z. Guo, K. Yu, N. Kumar, W. Wei, S. Mumtaz, M. Guizani, Deep distributed learning-based poi recommendation under mobile edge networks, *IEEE Internet Things J.*, **10** (2023), 303–317. https://doi.org/10.1109/JIOT.2022.3202628
5. Y. Jin, L. Hou, Y. Chen, A time series transformer based method for the rotating machinery fault diagnosis, *Neurocomputing*, **494** (2022), 379–395. https://doi.org/10.1016/j.neucom.2022.04.111
6. Q. Li, L. Liu, Z. Guo, P. Vijayakumar, F. Taghizadeh-Hesary, K. Yu, Smart assessment and forecasting framework for healthy development index in urban cities, *Cities*, **131** (2022), 103971. https://doi.org/10.1016/j.cities.2022.103971
7. J. Zhang, X. Liu, W. Liao, X. Li, Deep-learning generation of poi data with scene images, *ISPRS J. Photogramm. Remote Sens.*, **188** (2022), 201–219. https://doi.org/10.1016/j.isprsjprs.2022.04.004
8. Z. Guo, Y. Shen, S. Wan, W. Shang, K. Yu, Hybrid intelligence-driven medical image recognition for remote patient diagnosis in internet of medical things, *IEEE J. Biomed. Health. Inf.*, **26** (2022), 5817–5828. https://doi.org/10.1109/JBHI.2021.3139541
9. D. Zhang, X. Gao, A digital twin dosing system for iron reverse flotation, *J. Manuf. Syst.*, **63** (2022), 238–249. https://doi.org/10.1016/j.jmsy.2022.03.006
10. Z. Guo, Q. Zhang, F. Ding, X. Zhu, K. Yu, A novel fake news detection model for context of mixed languages through multiscale transformer, *IEEE Trans. Comput. Social Syst.*, 2023. https://doi.org/10.1109/TCSS.2023.3298480
11. X. Sun, Y. Zou, S. Wang, H. Su, B. Guan, A parallel network utilizing local features and global representations for segmentation of surgical instruments, *Int. J. Comput. Assisted Radiol. Surg.*, **17** (2022), 1903–1913. https://doi.org/10.1007/s11548-022-02687-z
12. Z. Chen, J. Chen, S. Liu, Y. Feng, S. He, E. Xu, Multi-channel calibrated transformer with shifted windows for few-shot fault diagnosis under sharp speed variation, *ISA Trans.*, **131** (2022), 501–515. https://doi.org/10.1016/j.isatra.2022.04.043
13. M. Sun, L. Xu, R. Luo, Y. Lu, W. Jia, Ghformer-net: Towards more accurate small green apple/begonia fruit detection in the nighttime, *J. King Saud Univ.-Comput. Inf. Sci.*, **34** (2022), 4421–4432. https://doi.org/10.1016/j.jksuci.2022.05.005
14. D. Chen, J. Zheng, G. Wei, F. Pan, Extracting predictive representations from hundreds of millions of molecules, *J. Phys. Chem. Lett.*, **12** (2021), 10793–10801.

15. N. P. Tigga, S. Garg, Efficacy of novel attention-based gated recurrent units transformer for depression detection using electroencephalogram signals, *Health Inf. Sci. Syst.*, **11** (2023). https://doi.org/10.1007/s13755-022-00205-8

16. B. Wang, Q. Li, Z. You, Self-supervised learning based transformer and convolution hybrid network for one-shot organ segmentation, *Neurocomputing*, **527** (2023). https://doi.org/10.1016/j.neucom.2022.12.028

17. S. Xiao, S. Wang, Z. Huang, Y. Wang, H. Jiang, Two-stream transformer network for sensor-based human activity recognition, *Neurocomputing*, **512** (2022), 253–268. https://doi.org/10.1016/j.neucom.2022.09.099

18. M. Mao, R. Zhang, H. Zheng, T. Ma, Y. Peng, E. Ding, et al., Dual-stream network for visual recognition, *Adv. Neural Inf. Process. Syst.*, **34** (2021), 25346–25358.

19. R. Kozik, M. Pawlicki, M. Choraś, A new method of hybrid time window embedding with transformer-based traffic data classification in iot-networked environment, *Pattern Anal. Appl.*, **24** (2021), 1441–1449. https://doi.org/10.1007/s10044-021-00980-2

20. A. A. Khan, R. Jahangir, R. Alroobaea, S. Y. Alyahyan, A. H. Almulhi, M. Alsafyani, et al., An efficient text-independent speaker identification using feature fusion and transformer model, *Comput. Mater. Contin.*, **75** (2023), 4085–4100.

21. D. Li, B. Li, S. Long, H. Feng, T. Xi, S. Kang, et al., Rice seedling row detection based on morphological anchor points of rice stems, *Biosyst. Eng.*, **226** (2023), 71–85. https://doi.org/10.1016/j.biosystemseng.2022.12.012

22. Y. Yang, J. Yu, H. Jiang, W. Han, J. Zhang, W. Jiang, A contrastive triplet network for automatic chest x-ray reporting, *Neurocomputing*, **502** (2022), 71–83. https://doi.org/10.1016/j.neucom.2022.06.063

23. B. Zhang, J. Abbing, A. Ghanem, D. Fer, J. Barker, R. Abukhalil, et al., Towards accurate surgical workflow recognition with convolutional networks and transformers, *Comput. Methods Biomech. Biomed. Eng.: Imaging Visualization* , **10** (2022), 349–356. https://doi.org/10.1080/21681163.2021.2002191

24. X. Pan, X. Gao, H. Wang, W. Zhang, Y. Mu, X. He, Temporal-based swin transformer network for workflow recognition of surgical video, *Int. J. Comput. Assisted Radiol. Surg.*, **18** (2023), 139–147.

25. Y. J. Shin, S. B. Jeong, H. I. Seo, W. Y. Kim, D. H. Seo, A study on handwritten parcel delivery invoice understanding model, *J. Adv. Mar. Eng. Technol. (JAMET)*, **46** (2022), 430–438. https://doi.org/10.5916/jamet.2022.46.6.430

26. Y. Liu, T. Bai, Y. Tian, Y. Wang, J. Wang, X. Wang, et al., Segdq: Segmentation assisted multi-object tracking with dynamic query-based transformers, *Neurocomputing*, **481** (2022), 91–101. https://doi.org/10.1016/j.neucom.2022.01.073

27. L. Tang, X. Xiang, H. Zhang, M. Gong, J. Ma, Divfusion: Darkness-free infrared and visible image fusion, *Inf. Fusion*, **91** (2023), 477–493. https://doi.org/10.1016/j.inffus.2022.10.034

28. H. Jiang, M. Gao, H. Li, R. Jin, H. Miao, J. Liu, Multi-learner based deep meta-learning for few-shot medical image classification, *IEEE J. Biomed. Health Inf.*, **27** (2023), 17–28. https://doi.org/10.1109/JBHI.2022.3215147

29. M. Luo, H. Wu, H. Huang, W. He, R. He, Memory-modulated transformer network for heterogeneous face recognition, *IEEE Trans. Inf. Forensics Secur.*, **17** (2022), 2095–2109. https://doi.org/10.1109/TIFS.2022.3177960

30. J. Izquierdo-Domenech, J. Linares-Pellicer, J. Orta-Lopez, Towards achieving a high degree of situational awareness and multimodal interaction with ar and semantic ai in industrial applications, *Multimedia Tools Appl.*, **82** (2023), 15875–15901. https://doi.org/10.1007/s11042-022-13803-1

31. Z. Yu, Y. Shen, J. Shi, H. Zhao, Y. Cui, J. Zhang, et al., Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer, *Int. J. Comput. Vision*, **131** (2023), 1307–1330.

32. H. Ji, X. Cui, W. Ren, L. Liu, W. Wang, Visual inspection for transformer insulation defects by a patrol robot fish based on deep learning, *IET Sci. Meas. Technol.*, **15** (2021), 606–618. https://doi.org/10.1049/smt2.12062

33. Y. Wu, K. Liao, J. Chen, J. Wang, D. Z. Chen, H. Gao, et al., D-former: A u-shaped dilated transformer for 3d medical image segmentation, *Neural Comput. Appl.*, **35** (2023), 1931–1944. https://doi.org/10.1007/s00521-022-07859-1

34. C. Liu, Z. Mao, A. Liu, T. Zhang, B. Wang, Y. Zhang, Focus your attention: A bidirectional focal attention network for image-text matching, in *Proceedings of the 27th ACM International Conference on Multimedia*, ACM, (2019), 3–11. https://doi.org/10.1145/3343031.3350869

35. C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, Graph structured network for image-text matching, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 10918–10927. https://doi.org/10.1109/CVPR42600.2020.01093

36. H. Diao, Y. Zhang, L. Ma, H. Lu, Similarity reasoning and filtration for image-text matching, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 1218–1226. https://doi.org/10.1609/aaai.v35i2.16209

AIMS Press