


Received June 17, 2020, accepted June 20, 2020, date of publication July 6, 2020, date of current version July 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007528

Classification of Financial Tickets Using Weakly Supervised Fine-Grained Networks

HANNING ZHANG^{1,2}, BO DONG^{3,4} , (Member, IEEE), BOQIN FENG¹, FANG YANG⁵, AND BO XU⁵

¹School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China

²Shaanxi Province Key Laboratory of Satellite and Terrestrial Network Technology Research and Development, Xi'an Jiaotong University, Xi'an 710049, China

³School of Continuing Education, Xi'an Jiaotong University, Xi'an 710049, China

⁴National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University, Xi'an 710049, China

⁵Xi'an Network Computing Data Technology Company, Ltd., Xi'an 710049, China

Corresponding author: Bo Dong (dong.bo@xjtu.edu.cn)

This work was supported in part by the Fundamental Theory and Applications of Big Data with Knowledge Engineering under the National Key Research and Development Program of China under Grant 2018YFB1004500, in part by the MOE Innovation Research Team under Grant IRT_17R86, in part by the National Science Foundation of China under Grant 61721002 and Grant 61532015, in part by the Project of China Knowledge Center for Engineering Science and Technology, and in part by the Project of XJTU-SERVYOU Joint AI Innovation Center.


ABSTRACT Facing the rapid growth in the issuance of financial tickets, traditional manual invoice reimbursement methods are imposing an increasing burden on financial accountants and consuming excessive manpower. There are too many categories of financial ticket that need to be classified with high accuracy. Therefore, we propose a Financial Ticket Classification (FTC) network based on weakly supervised fine-grained classification discriminative filter learning networks, which greatly improves the work efficiency of financial accountants. The FTC network adopts an end-to-end network structure and uses a deep convolution network to extract highly descriptive features. By using a fully convolutional network (FCN), this method reduces the depth and width of the whole network and avoids the over duplication of features and the overconsumption of system memory. To obtain more accurate classification results, we use the large-margin softmax (L-softmax) loss function, which can make the features learned in the class more compact, make it easier to separate subclasses, and effectively prevent overfitting. Experimental results show that the proposed FTC network achieves both high accuracy (up to 99.36%) and high processing speed, which perfectly meets the requirements of accurate and real-time classification for financial accounting applications.

INDEX TERMS Ticket classification, weakly supervised learning, fine-grained networks, deep learning.

I. INTRODUCTION

In recent years, with the rapid development of computer hardware, computer vision and other technologies, deep learning is being adopted by an ever-widening group of fields [1]–[5]. Finance-and-tax is an important field that implements deep learning applications [6], [7]. Traditionally, accounting is usually performed manually as follows. First, the different types of financial tickets, such as value-added tax (VAT) invoices (common invoices, electronic invoices, and special invoices), bank tickets, toll tickets (highway passenger tickets, vehicle occupation fees, highway tolls,) are manually sorted. Second, the basic information of these financial tickets

is manually input into the financial software to produce accounting vouchers for the corresponding category. Then, each financial ticket is sequentially attached to the accounting voucher for the corresponding category. Finally, the accountant must repeatedly check whether the ordering of the tickets is correct and whether there are any missing tickets. However, this approach is obviously slowed by the lack of automation. Due to the large number and variety of financial tickets, the process results in massive classification workloads, time consumption, and labor effort on the part of the accounting staff, leading to high labor costs and low work efficiency. The accuracy of input information is also greatly affected. Therefore, in order to make accounting more accurate, more efficient and highly automated, optical character recognition (OCR) technology has been gradually applied to the field

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo .

of financial ticket recognition [7]–[9]. The ticket information identification system can not only reduce work tasks and pressure and improve work efficiency but also resolve contradictions caused by rising labor costs and labor shortages. Additionally, it can promote the digitalization, maintenance and intelligent accounting and storage of accounting information, making it more convenient for accountants to review.

The highly accurate and efficient classification of tickets is a very important step in the ticket recognition system. The correct classification of tickets can result in more accurate OCR and structured information extraction. Through this method, the work efficiency of financial accountants can be greatly improved.

To meet the practicality of use requirements for financial accounting systems, our solution should implement the following four improvements: 1. Reduce the cost of manual labeling in the model training process; 2. Improve the classification accuracy of the system as much as possible; 3. Support more financial ticket types, that is, the number of categories; and 4. Increase the classification processing efficiency.

To implement the improvements mentioned above, we proposed the Financial Ticket Classification (FTC) network based on weakly supervised fine-grained classification discriminative filter learning networks. This paper proposes 1. Using DenseNet as the basic network, which has a strong feature description capability and can effectively improve the classification accuracy; 2. A fully convolutional network (FCN) to replace the fully connected (FC) structure, which can effectively reduce the depth and width of the entire network, the number of training parameters and system resource consumption; and 3. The large-margin softmax (L-softmax) loss function, which can effectively improve the classification accuracy by using small intraclass variance and large interclass variance. Compared with existing financial ticket classification methods, the proposed method can support more financial ticket types and achieve higher classification accuracy and processing efficiency.

The structure of this manuscript is as follows: Section I introduces the latest research. Section II describes related work. Section III presents the overall framework of the proposed system, including detailed information about the dataset, data preprocessing, and FTC network structure, and the layer configuration of the model used in the experiment. Section IV describes the verification of the FTC network proposed in this work and analyzes the experimental results. Section V summarizes overall research.

II. RELATE WORKS

Currently, there are two main automatic financial ticket classification methods. One method is based on combining artificially designed features (such as SIFT and HOG) with machine learning classifiers (such as SVM and KNN) for classification [10]–[12]. The other method is based on a deep neural network, which extracts discriminative features for

classification [13], [14]. The artificially designed features depend on the layout of the tickets, such as frame lines, headers, text areas, and other discerning feature information [15]. The descriptiveness of the features extracted using this method is very limited. Indeed, this method is intended mainly for a certain type of ticket and has poor adaptability to tickets with new and different structures. The features extracted with methods based on deep neural networks are relatively highly descriptive and do not need to be manually designated, so such methods have been widely used. Due to the low similarity between major categories of the financial ticket, there is a large interclass variance. Tickets within the same major category (subclass) have high similarity and small variance. Therefore, we use the fine-grained classification technology of convolutional neural networks to classify financial tickets. At present, fine-grained classification is mainly divided into strongly supervised fine-grained classification and weakly supervised fine-grained classification. Strongly supervised fine-grained classification methods such as Part RCNN [16] provide image category labels, labeled boxes and local area locations for the training samples, but the manual labeling cost is very large. Weakly supervised fine-grained classification methods, such as B-CNN [17] and DFL-CNN [18], only need to label the classified categories.

Ahmad S. Tarawneh *et al.* [19] used AlexNet to extract convolutional features, and then separately used Random Forests, K-nearest neighbors (KNN), and Naive Bayes to classify the tickets. In their method, the invoices are divided into three categories: handwritten, machine-printed and receipts. Among the classification methods, the KNN classification performed best, with the classification accuracy reaching 98.4%. However, this method only supports three categories, which is unable to support subsequent OCR and structured information extraction for tickets.

JIE YANG *et al.* [20] proposed an intelligent reimbursement system. The invoice classification module in this system divides invoices into three categories: VAT (value-added tax) invoices, common printed invoices and train tickets. The invoice image is obtained with a scanner. The smallest size tickets are directly judged as train tickets. Since VAT invoices and common printed invoices have similar sizes, they are distinguished by extracting keywords from the invoice title. However, this method has highly restrictive requirements on the source of the data, as it must be generated by the same type of scanner, and there are few classification categories.

YINGYI SUN *et al.* [21] classified invoices into three categories: VAT (value-added tax) invoice, train tickets, and taxi invoices. This method proposes an optimized network based on the VGG-16 network. The optimized network model includes 8 convolutional layers and 3 fully connected layers. To prevent the gradient from disappearing, this method adds batch normalization layers (BNs) and regularization constraints. The data source of the invoice is a photograph obtained by a smartphone. However, this method only classifies 3 categories. It is difficult to meet the requirements of a practical financial accounting application.

III. THE NETWORK MODEL

The basic process of classification is shown in Figure 1.

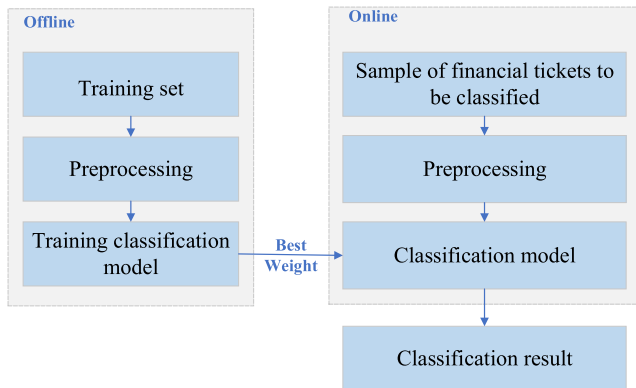


FIGURE 1. Basic flow chart of classification.

Different types of financial tickets are collected to make a training set, which is then preprocessed. The model is trained on the GPU server and the parameters are adjusted accordingly. The trained Best Weight is deployed to the server to classify the ticket samples online. Its input is a sample of the tickets to be classified, and its output is the corresponding label.

A. THE DATASET

Through long-term collation, we divide financial tickets into 12 major-classes with a total of 482 subclasses. Our actual financial software has accumulated more than 10 million real tickets, which continue to increase by approximately 10,000 every day. The names of the major classes of tickets and the number of corresponding subclasses are shown in Table 1.

TABLE 1. Classification of financial tickets.

| Class Index | Class name | Number of subclasses |
|-------------|------------------------|----------------------|
| 1 | Bank ticket | 149 |
| 2 | VAT | 104 |
| 3 | Toll | 85 |
| 4 | Quota ticket | 51 |
| 5 | Admission ticket | 25 |
| 6 | Common printed invoice | 21 |
| 7 | Taxi ticket | 21 |
| 8 | Plane ticket | 1 |
| 9 | Train ticket | 1 |
| 10 | Insurance policy | 7 |
| 11 | Receipt | 2 |
| 12 | Others | 15 |

Bank tickets include those obtained from 16 banks such as the China Construction Bank (CCB), the Industrial and Commercial Bank of China (ICBC), the People’s Bank of China (PBOC), the Bank of China (BOC), and the Bank of Communications (BCM). There are 149 common categories of bank tickets, which include remittance receipts, payment receipts, withdrawal receipts, refund receipts, interest receipts, tax payment vouchers and accounting vouchers. The 104 VAT

categories include ordinary invoices, electronic invoices and special invoices.

There are 85 categories of tolls, which mainly include vehicle occupation fees, highway tolls, parking fees, and passenger tickets in most places. There are 51 categories of quota tickets collected from most of the regions. Admission tickets consist of 25 categories of common scenic tickets. There are 21 categories of general machine printed invoices issued by tax bureaus in various regions. There are 21 categories of taxi tickets obtained from various regions. Insurance policies include 7 categories such as HUA Insurance, CPIC Insurance, and Life Insurance. There are two categories of receipts: machine-printed receipts and handwritten receipts. The Other class includes 15 categories such as payroll, tax payment certificates, and patent fees.



FIGURE 2. Financial ticket examples.

As shown in Figure 2, the major categories of financial tickets have few similarities and large interclass variances. The interclass tickets for each major category have high similarity and small variance. At the same time, there are many uncertain factors that increase the difficulty of classification, such as a large number of categories of financial tickets, similar structures, incompleteness, occlusions, folds, different sizes, low or uneven brightness, similar colors, geometric distortions, complex shooting backgrounds, and blurring or deformation. Therefore, this work needs to preprocess the dataset to improve the accuracy of classification.

B. TICKET PREPROCESSING

The purpose of ticket preprocessing is to clean and enhance the ticket samples in the dataset. The task consists of the following steps:

1) SEGMENTATION

Some of the single samples in the dataset may contain multiple financial ticket images, or the background of the ticket images may be very complicated. To ensure that the input sample contains only a single ticket image and reduces background interference, it is necessary to segment the largest area containing a single ticket image. Considering the different styles and sizes of financial tickets, we used the SSD (Single Shot MultiBox Detector) method for automated financial tickets segmentation. Because SSD can predict the detection results from the feature maps of CNN at different levels, so it

can better detect and segment out tickets of different sizes. Figure 3 shows an example of the segmentation of financial tickets.

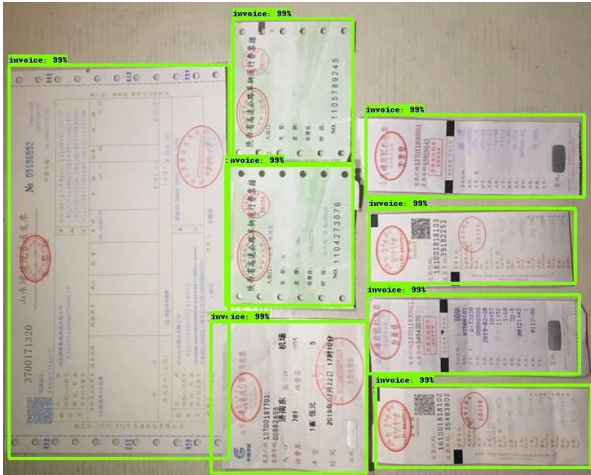


FIGURE 3. An example of segmentation of financial tickets.

2) REMOVE NON CLASSIFIED AND DUPLICATE TICKET IMAGES

A large number of repeated data will affect the training results and may lead to overfitting. We use the similarity between images to remove duplicate ticket images to ensure the uniqueness of each ticket image in the training dataset.

3) ROTATION CORRECTION

The experimental results show that the classification accuracy of the non-rotated training set data is poor. Therefore, without changing the original features of the ticket image, we rotate the samples in all training sets before training to ticket angles 90° , 180° , 270° and 0° .

4) INCREASE THE NUMBER OF RARE TICKET SAMPLES IN THE CLASSIFICATION CATEGORY

As the number of types of financial tickets continues to increase, the number of tickets in some categories will be relatively small, which will seriously affect the accuracy of classification. The automated classification method needs to be constructed and expanded to ensure that there are enough samples to improve the accuracy of the model, ensuring that the characteristics from even rare categories of tickets can be learned. The number of ticket images for each category in the training set cannot be less than k (In our experiment, $k = 300$). When this condition is not satisfied, enhancement operations to increase the number of samples for these classes of ticket images, such as random change, in contrast, Gaussian blur, random rotation (-45° to 45°), affine transformation, dirty point simulation, and deformation, are carried out. For training the network, each ticket image in the dataset counts as a different input. Increasing the diversity of the dataset will help to achieve more accurate classification results for the different tickets and enhance the generalization ability of the model.

C. NETWORK ARCHITECTURE

This section mainly describes the network structure and analyzes the improvement in the weakly supervised fine-grained classification network based on the Discriminative Filter Learning (DFL-CNN) [18] model. In the financial ticket classification scenario, based on the general DFL-CNN model, we have the following two targeted improvements: 1. higher classification accuracy; 2. faster processing speed.

Because a fully connected structure will not only widen the depth and width of the entire network but also cause too many repeated features, increase background noise interference, waste computer computing resources, and increase system memory consumption, we use a fully convolutional network (FCN) to replace the fully connected (FC) structure and extract local salient features instead of the features of the entire image. The DFL model uses a softmax loss function in the G-stream, P-stream, and Side branch modules. The softmax loss function is good at optimizing the distance between major classes, but the accuracy is not high when classifying within subclasses. By analyzing the principles of several different loss functions (Softmax, L-Softmax, A-Softmax loss, AM-Softmax loss) [22], [27], [28], considering the characteristics of the financial ticket data set, this work uses the large-margin softmax (L-softmax) loss function [22] to replace the softmax loss function of the original network for more accurate classification. The L-softmax loss function can make the features learned within the class more compact, make it easier to separate the subclasses, and effectively prevent overfitting. Algorithm 1 describes the calculation process of the L-softmax loss function.

Algorithm 1 L-Softmax Loss Function

Input:

$\{x_i\}$ is the i th input feature, and its corresponding label is y_i ; W_{y_i} is the weight of the i th input feature corresponding to the label;

$m: m \in N^*$

λ, β, γ : coefficients of loss

Output: $loss_{total}$

Function:

$k \leftarrow K(p)$

{if $\cos(p) > \cos(\frac{\pi}{m})$ $k=0$, elseif $\cos(p) \leq \cos(\frac{\pi}{m})$ $k=1$ }

$$(1) \cos(\theta_j) \leftarrow \frac{W_j^T x_i}{\|W_j\| \|x_i\|}$$

$$\cos(m\theta_{y_i}) \leftarrow C_m^0 \cos^m(\theta_{y_i}) - C_m^2 \cos^{m-2}(\theta_{y_i}) (1 - \cos^2(\theta_{y_i})) +$$

$$(2) C_m^4 \cos^{m-4}(\theta_{y_i})(1 - \cos^2(\theta_{y_i}))^2 + \dots (-1)^n C_m^{2n} \cos^{m-2n}(\theta_{y_i})(1 - \cos^2(\theta_{y_i}))^n + \dots$$

$$(3) k \leftarrow K(\cos(\theta_j))$$

$$(4) \psi(\theta_{y_i}) \leftarrow (-1)^{k*} \cos(m\theta_{y_i}) - 2*k$$

$$(5) L_i \leftarrow -\log \left(\frac{e^{\|W_{y_i}\| \|x_i\| \psi(\theta_{y_i})}}{e^{\|W_{y_i}\| \|x_i\| \psi(\theta_{y_i})} + \sum_{j \neq y_i} e^{\|W_j\| \|x_i\| \cos(\theta_j)}} \right)$$

$$(6) loss_{total} \leftarrow \lambda * loss_g + \beta * loss_p + \gamma * loss_s$$

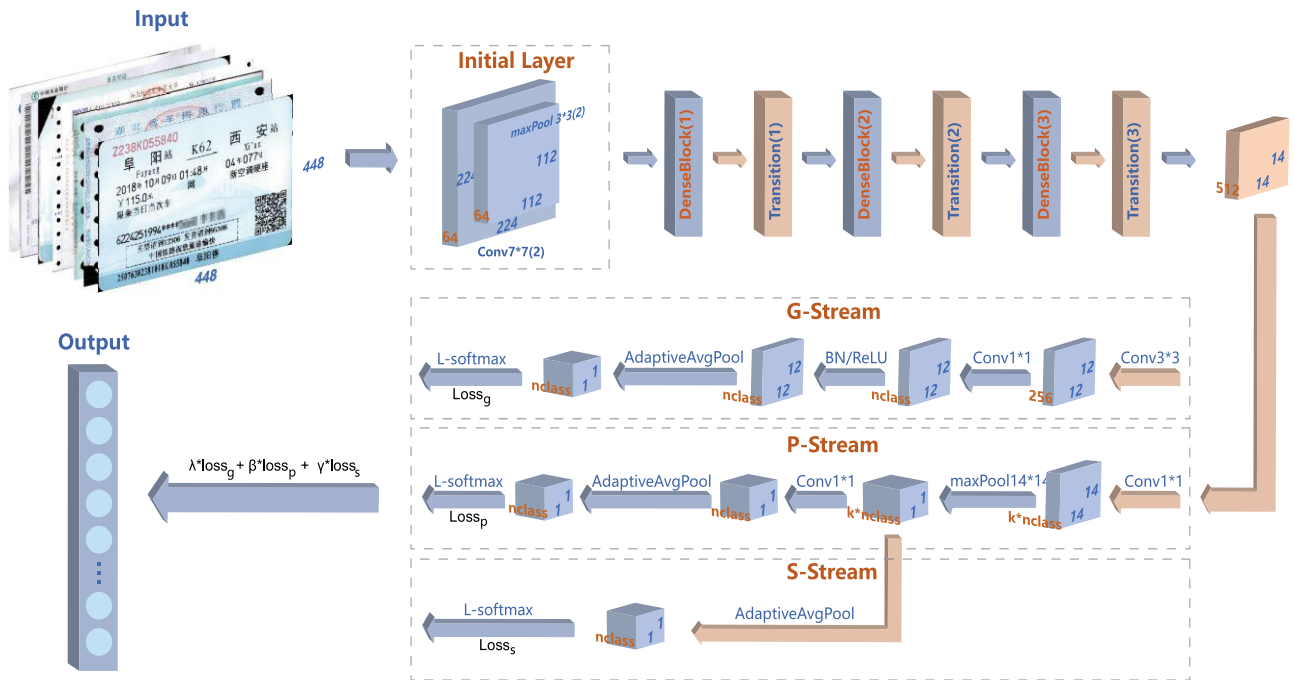


FIGURE 4. The structure of Financial Ticket Classification (FTC) network.

The FTC network consists of four modules: basic network module, G-stream module, P-stream module and Side branch module. A detailed description of the network structure is shown in Figure 4.

The basic network module is used to extract features. AlexNet [23], VGGNet [24], ResNet [25] and other networks can be used for feature extraction. However, the number of layers in the AlexNet network is shallow, and so higher-dimensional features cannot be extracted. The large parameters of VGGNet will result in long training time and require a device with a large storage capacity. ResNet is deeper than the other two networks, and a large number of parameters are reduced without using a fully connected structure. Compared with the above feature extraction networks, the DenseNet [26] network has fewer parameters and is easier to train. It has a greater depth of network layers, can extract high-dimensional feature information, and can easily prevent over-fitting. Therefore, we use DenseNet with its ability to extract highly descriptive features as the basic network for feature extraction. To reduce training time and, at the same time, guarantee the accuracy of the extracted features, we use some layers of DenseNet121 as the basic network, including an initialization layer, 3 DenseBlock layers, and 3 Transition layers. The initialization layer consists of a convolution (Conv) of size $7 * 7$ with stride 2, a BN, a rectified linear units (ReLU), and a max pooling layer (maxPool) with a kernel of size $3 * 3$ with stride 2. The pooling layer is used to reduce the dimensions of the data. DenseBlock 1, DenseBlock 2 and DenseBlock 3 are composed of 6, 12 and 24 dense units, respectively, each of which in turn contains:

$$BN \rightarrow ReLU \rightarrow Conv_{(1*1)} \rightarrow BN \rightarrow ReLU \rightarrow Conv_{(3*3)}.$$

Each transition layer consists of a BN, a ReLU, and an average pooling layer (avgPool) with a kernel of size $2 * 2$ with stride 2. The $1 * 1$ convolution reduces the number of output channels and improves the compactness of the model. The G-Stream module focuses on global information. Its input is a $512 * 14 * 14$ feature map output by Transition Layer 3. Then, a Conv of size $3 * 3$ with stride 1 is used to obtain a feature map of size $256 * 12 * 12$. Then, the features are transformed into $n_{class} * 1 * 1$ column vectors through an FCN layer. The FCN layer contains a $1 * 1$ Conv, a BN, a ReLU and a $1 * 1$ AdaptiveAvgPool2D. Finally, the loss $loss_g$ is obtained by passing the $n_{class} * 1 * 1$ feature vector through the loss function L-Softmax.

The P-Stream module focuses on key local information. Its input is the feature map output by Transition Layer 3. Then, the feature map of size $512 * 14 * 14$ is convolved by a Conv of size $1 * 1$ with stride 1 to obtain a feature map of size $(k * n_{class}) * 14 * 14$, which is followed by global max pooling with a kernel size of $14 * 14$. Then, the output is passed to an FCN layer, which contains a $1 * 1$ Conv and a $1 * 1$ AdaptiveAvgPool2D. Finally, the loss function L-Softmax is used to obtain the loss $loss_p$.

The S-Stream module is a supervision module that highlights the distinguishing parts in the feature map. The input is the global maximum pooling feature map from the P-stream module. It uses average pooling on the $k * n_{class} * 1 * 1$ vector with k as the growth rate to obtain an $n_{class} * 1 * 1$ vector. This operation is called Cross-Channel pooling. Finally, the loss function L-Softmax is used to obtain the loss $loss_s$.

IV. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the FTC network in the classification of financial tickets, we designed two sets of

TABLE 2. Detailed configuration information for the model.

| MODEL NAME | Layer | Output Layer (C*H*W) | Number of parameters | |
|---------------------|------------------------------------------------------------------------------------------------------|---------------------------------------|---------------------------------------|-----------------------------------|
| Initialization | Input | 3*448*448 | 149 | |
| | 7*7(2)Conv | 64*224*224 | 104 | |
| | BN/ReLU | 64*224*224 | 85 | |
| | 3*3(2)maxPool | 64*112*112 | 51 | |
| Base CNN network | DenseBlock(1) $\begin{bmatrix} BN / ReLU \\ conv_1*1 \\ BN / ReLU \\ conv_3*3 \end{bmatrix} * 6$ | $[64 + 32i]*112*112, \{i \in [0,5]\}$ | $128 + 64i, \{i \in [0,5]\} / 0$ | |
| | | 128*112*112 | $8192 + 4096i, \{i \in [0,5]\}$ | |
| | | 128*112*112 | 256/0 | |
| | | 32*112*112 | 36864 | |
| | Transition Layer(1) | BN/ReLU | 256*112*112 | 512/0 |
| | | 1*1conv | 128*112*112 | 32768 |
| Dense Net 121 | DenseBlock (2) $\begin{bmatrix} BN / ReLU \\ conv_1*1 \\ BN / ReLU \\ conv_3*3 \end{bmatrix} * 12$ | $[128 + 32i]*56*56, \{i \in [0,11]\}$ | $256 + 64i, \{i \in [0,11]\} / 0$ | |
| | | 128*56*56 | $16384 + 4096i, \{i \in [0,11]\}$ | |
| | | 128*56*56 | 256/0 | |
| | | 32*56*56 | 36864 | |
| | Transition Layer(2) | BN/ReLU | 512*56*56 | 1024/0 |
| | | 1*1conv | 256*56*56 | 131072 |
| | DenseBlock (3) $\begin{bmatrix} BN / ReLU \\ conv_1*1 \\ BN / ReLU \\ conv_3*3 \end{bmatrix} * 24$ | | $[256 + 32i]*28*28, \{i \in [0,23]\}$ | $512 + 64i, \{i \in [0,23]\} / 0$ |
| | | | 128*28*28 | $32768 + 4096i, \{i \in [0,23]\}$ |
| | | | 128*28*28 | 256/0 |
| | | | 32*28*28 | 36864 |
| Transition Layer(3) | BN/ReLU | 1024*28*28 | 2048/0 | |
| | 1*1conv | 512*28*28 | 524288 | |
| G-stream | 2*2(2)avgPool | 512*14*14 | 0 | |
| | 3*3Conv | 256*12*12 | 1179904 | |
| | 1*1Conv | nclass*12*12 | 257 nclass | |
| | BN/ReLU | nclass*12*12 | 2nclass/0 | |
| | AdaptiveAvgPool2D | nclass *1*1 | 0 | |
| | | L-Softmax($loss_g$) | | |
| P-Stream | 1*1Conv6 | k*nclass *14*14 | 513*k* nclass | |
| | 14*14(14)maxPool | k*nclass *1*1 | 0 | |
| | 1*1Conv | nclass *1*1 | $\sum_{i=1}^{k-2} (31+20i)$ | |
| | AdaptiveAvgPool2D | nclass *1*1 | 0 | |
| | | L-Softmax($loss_p$) | | |
| Side branch | avgPool | nclass *1*1 | 0 | |
| | | | L-Softmax($loss_s$) | |

DenseBlock consists of multiple dense units, *12 means there are 12 dense units, nclass is the number of classes, and k is the growth rate of the network.

experiments: (1) classification experiments with different classification network structures, and (2) an experimental comparison of different parameters. The experimental setup is shown in Section IV A, and the dataset is shown in Table 3. The input image for the experiments is an RGB image.

A. EXPERIMENTAL SETUP

The algorithm in this work uses the deep learning Pytorch framework to build the network in the Python programming language. The model is deployed on a Linux server with CentOS Linux release (Core) 7 operating system. The GPU is a single NVIDIA Tesla P40 with 24 GB video memory, the

CPU frequency is 2.20 GHz, and the computer has 32 GB memory. Details of the layer configuration of the model in the experiment are shown in Table 2.

B. THE DATASET

The equipment for obtaining the dataset sample in these experiments includes imaging equipment such as scanners, mobile phones, and digital cameras and can also collect financial ticket images in the SaaS financial software platform. The experiment uses 10 major classes of financial tickets, including a total of 100 subclasses of commonly used tickets, each of which consists of a sample of 300 tickets, and the total dataset is 30000. The training data and validation data are allocated according to 5:1, that is, there are 250 training data and 50 validation data in each subclass. The total of the training data is 25000, and the validation data is 5000. The dataset in the experiment is shown in Table 3, which covers 86 kinds of bills of 48 banks and VAT, toll, taxi, quota, train ticket. 13 of the 100 classes contain multiple ticket styles, therefore, the fine grained classifier can fully show its advantages. To improve the classification accuracy, the aspect ratio of each input image is maintained during the training and testing stages so that the features of the image are not distorted. We resize the short sides of the input image to 448 pixels, and the long sized are resized to maintain the original aspect ratio.

TABLE 3. Training dataset for the experiment.

| Serial number | Class name | Number of class |
|---------------|------------------|-----------------|
| 1 | Bank ticket | 86 |
| 2 | VAT | 2 |
| 3 | Toll | 4 |
| 4 | Quota ticket | 1 |
| 5 | Taxi ticket | 1 |
| 6 | Plane ticket | 1 |
| 7 | Train ticket | 1 |
| 8 | Insurance policy | 2 |
| 9 | Receipt | 2 |

C. THE METRICS

We use the accuracy and recall as the basic index to measure the algorithm and model, and focus on the time consumption of a single sample from the perspective of the financial accounting application requirements. The metrics are shown in Table 4.

TABLE 4. The metrics.

| Term | Unit | Definition |
|----------|----------|----------------------------------------------------------------------------------------------------------------|
| Accuracy | % | The ratio of predicted labels which exactly match the corresponding ground truth. |
| Recall | % | The ratio between the number of positive samples correctly predicted and the total number of positive samples. |
| Time | ms/image | Forward operation time of single sample. |

D. THE RESULTS

We use stochastic gradient descent (SGD) with a momentum algorithm to optimize the model. The initial value of the learning rate (lr) is set to 0.001, the size of the momentum is set to 0.9, the size of the weight decay is set to 0.000005, the training batch size is set to 8, and the maximum number of iterations is set to 7,546.

1) EXPERIMENT 1 CLASSIFICATION WITH DIFFERENT NETWORK STRUCTURES

To verify the accuracy and real-time classification of the deep learning algorithm on the financial ticket training set, we use four different mainstream CNN networks to classify financial tickets: ResNet50, VGG16, DenseNet121, and DFL-DenseNet121. All networks implement the softmax loss function. The loss coefficients λ, β, γ of DFL-DenseNet121 are 1, 1, and 0.1, respectively.

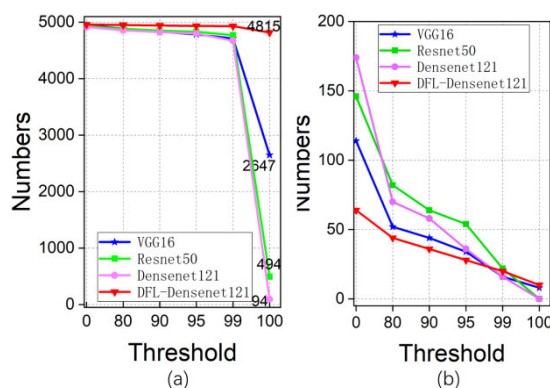


FIGURE 5. (a) The images successfully classified by the different network models. (b) The images misclassified based on (a).

Figure 5 (a) shows a plot of the number of samples that were successfully classified as the threshold was increased using different structured network models. Figure 5 (b) shows a plot of the number of misclassified samples for the different structured network models with increasing threshold based on Figure 5 (a). Due to the need for a low fault-tolerance rate for the financial accounting application, it is necessary to reduce the amount of labor while ensuring high accuracy. Figure 5 shows that the number of correctly classified samples for the four networks is similar when the confidence threshold is less than 99%. However, with a 100% confidence level, the DFL-DenseNet121 network yielded a small number of unclassified samples, compared to the abovementioned networks, which accounted for only 3.7% of the total test samples.

The analysis of the experimental results in Table 5 suggests that the feature extraction capabilities of the four CNN networks are very good, and they can successfully classify financial tickets. With a confidence threshold requirement of 99%, the prediction accuracy and recall of the ResNet50, VGG16 and DenseNet121 has reached more than 93%. The prediction accuracy and recall of DFL-Densenet121 are 98.16% and 98.68%, respectively. When

the confidence threshold is 100%, the classification accuracy of DFL densenet121 network is 96.10%, and the recall is 96.35%, which is much higher than the other three networks. However, the classification accuracy of ResNet50, VGG16 and DenseNet121 networks is only 9.88%, 52.78% and 1.88% respectively, and the recall is only 9.81%, 52.38% and 1.85% respectively. Based on the above situations, this study chooses to use the fine-grained network structure DFL-DenseNet121 as the network model for financial ticket classification.

TABLE 5. Comparison of different models and parameters.

| Model | Model Size (MB) | Confidence 99% | | Confidence 100% | | Time (ms/image) |
|----------------|-----------------|----------------|--------|-----------------|--------|-----------------|
| | | Accuracy | Recall | Accuracy | Recall | |
| ResNet50 | 181.2 | 94.92 | 95.15 | 9.88 | 9.81 | 198.97 |
| VGG16 | 1027.57 | 93.88 | 93.60 | 52.78 | 52.38 | 224.85 |
| DenseNet121 | 54.37 | 93.08 | 93.38 | 1.88 | 1.85 | 240.48 |
| DFL (DenseNet) | 628.18 | 98.16 | 98.68 | 96.10 | 96.35 | 221.87 |

2) EXPERIMENT 2 COMPARISON OF DIFFERENT PARAMETERS

One of the goals of this work is to improve the weakly supervised fine-grained classification network DFL-CNN. To verify the improvement effect, the following comparative experiments were performed.

TABLE 6. Experimental comparison of different parameters.

| Model | Parameters | Time (ms/image) | Accuracy (%) |
|-----------------|---------------------|-----------------|--------------|
| DFL-DenseNet121 | FC Softmax | 221.87 | 99.10 |
| DFL-DenseNet121 | FC L-Softmax (m=2) | 215.17 | 99.18 |
| FTC Network | FCN L-Softmax (m=2) | 213.85 | 99.36 |

TABLE 7. Comparison of related methods.

| | Classes | Accuracy | Data sources types |
|-----------|---------|----------|-----------------------------------------------------------------------------------|
| Ref.[19] | 3 | 98.4 | --- |
| Ref.[20] | 3 | --- | scanner |
| Ref.[21] | 3 | 99.05 | smart phones |
| This work | 100 | 99.36 | scanners, smart phones, digital cameras, actual financial SaaS software platform. |

It can be seen from Table 6 that the fully connected structure uses the whole characteristics of the image, which contains more background noise and reduces the accuracy. In this work, the detailed information of the salient area (locally significant features) is combined with fully convolutional layers to remove the interference of background noise interference and achieve an accuracy of 99.36%. Table 7 shows that our accuracy is improved by 0.96% and 0.31%, respectively,

compared to reference [19] and reference [21]. The number of classification categories in this article is very large and is increased while ensuring similar accuracy. In short, this work is far superior to the methods in references [19]–[21] in terms of the categories of financial tickets supported and the input sources of these supported financial tickets.

V. CONCLUSION

The classification of a vast number of types of financial tickets results in a heavy classification workload and low work efficiency for accounting staff. Therefore, we proposed an FTC network to automatically classify financial tickets, captured by scanners, smart phones, digital cameras and an actual financial accounting software platform. We used an end-to-end network structure and a deep convolution network, DenseNet, capabilities as the basic network to extract highly descriptive features, as well as a fully convolutional network (FCN) to replace the fully connected (FC) structure of DFL-CNN, which can effectively reduce the depth and width of the entire network, the number of training parameters and system resource consumption. Additionally, we used the L-Softmax loss function, which is good at optimizing the distance between classes to improve the classification accuracy. To further improve the accuracy of classification, the input ticket images were preprocessed during the training and testing stages, which included ticket background segmentation, cleaning, rotation, dataset enhancement, and some further operations. Additionally, the short sides of the input images were resized to 448 pixels, and the aspect ratio was maintained to avoid distortions. To improve the generalization ability of the model, 448 * 448 image blocks were randomly cropped from the resized image. The experimental results indicate that the proposed FTC network can achieve both high accuracy (up to 99.36%) and high processing speed. Compared with existing financial ticket classification methods, this method can support a greater number of financial tickets types and achieve higher classification accuracy and processing efficiency.

REFERENCES

- [1] R. Miikkulainen, J. Liang, E. Meyerson, and A. Rawal, "Evolving deep neural networks," in *Artificial Intelligence in the Age of Neural Networks and Brain Computing*, R. Kozma, Ed. New York, NY, USA: Academic, 2019, ch. 15, pp. 293–312.
- [2] E. Charniak, *Introduction to Deep Learning*. Cambridge, MA, USA: MIT Press, 2019.
- [3] X. Feng, Y. Jiang, X. Yang, M. Du, and X. Li, "Computer vision algorithms and hardware implementations: A survey," *Integration*, vol. 69, pp. 309–320, Nov. 2019.
- [4] A. I. Solis and P. Nava, "Domain specific architectures, hardware acceleration for machine/deep learning," *Proc. SPIE*, vol. 11013, May 2019, Art. no. 1101307.
- [5] N. O'Mahony, S. Campbell, A. Carvalho, "Deep learning vs. traditional computer vision," in *Advances in Computer Vision*. Cham, Switzerland: Springer, 2020.
- [6] M. Jha, M. Kabra, S. Jobanputra, and R. Sawant, "Automation of cheque transaction using deep learning and optical character recognition," in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Nov. 2019, pp. 309–312.
- [7] S. Srivastava, "Optical character recognition on bank cheques using 2D convolution neural network," in *Applications of Artificial Intelligence Techniques in Engineering*. Singapore: Springer, 2019.

- [8] H. T. Ha, "Recognition of invoices from scanned documents," in *Recent Advances in Slavonic Natural Language Processing*, 2017, pp. 71–78.
- [9] S. Shreya, Y. Upadhyay, M. Manchanda, R. Vohra, and G. D. Singh, "Optical character recognition using convolutional neural network," in *Proc. 6th Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2019, pp. 55–59.
- [10] A. Subashini and N. D. Kodikara, "A novel SIFT-based codebook generation for handwritten tamil character recognition," in *Proc. 6th Int. Conf. Ind. Inf. Syst.*, Aug. 2011, pp. 261–264.
- [11] C. Yi, X. Yang, and Y. Tian, "Feature representations for scene text character recognition: A comparative study," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 907–911.
- [12] N. A. Jebri, H. R. Al-Zoubi, and Q. A. Al-Haija, "Recognition of handwritten arabic characters using histograms of oriented gradient (HOG)," *Pattern Recognit. Image Anal.*, vol. 28, no. 2, pp. 321–345, 2018.
- [13] C. Boufenar, A. Kerboua, and M. Batouche, "Investigation on deep learning for off-line handwritten arabic character recognition," *Cognit. Syst. Res.*, vol. 50, pp. 180–195, Aug. 2018.
- [14] S. Zheng, X. Zeng, G. Lin, C. Zhao, Y. Feng, J. Tao, D. Zhu, and L. Xiong, "Sunspot drawings handwritten character recognition method based on deep learning," *New Astron.*, vol. 45, pp. 54–59, May 2016.
- [15] C. Zengzhao, H. Xiuling, and Y. Y. D. Cailin, "Bill classification technology based on multiple characteristics fusion and its applications," *Comput. Eng. Appl.*, no. 9, pp. 202–204, 2006.
- [16] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2014, pp. 834–849.
- [17] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.
- [18] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4148–4157.
- [19] A. S. Tarawneh, A. B. Hassanat, D. Chetverikov, I. Lendak, and C. Verma, "Invoice classification using deep features and machine learning techniques," in *Proc. IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol. (JEEIT)*, Apr. 2019, pp. 855–859.
- [20] J. Yang, Y. Gao, Y. Ding, Y. Sun, Y. Meng, and W. Zhang, "Deep learning aided system design method for intelligent reimbursement robot," *IEEE Access*, vol. 7, pp. 96232–96239, 2019.
- [21] Y. Sun, J. Zhang, Y. Meng, J. Yang, and G. Gui, "Smart phone-based intelligent invoice classification method using deep learning," *IEEE Access*, vol. 7, pp. 118046–118054, 2019.
- [22] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. 33rd Int. Conf. Int. Conf. Mach. Learn.*, vol. 48. New York, NY, USA, 2016, pp. 507–516.
- [23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1106–1114.
- [24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, pp. 1409–1556, 2015.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
- [27] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 212–220.
- [28] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.



HANNING ZHANG received the B.S. and M.S. degrees in computer science and technology from Northwest University, China, in 2007 and 2010, respectively. He is currently pursuing the Ph.D. degree in computer science and technology with Xi'an Jiaotong University, China. His research interests include deep learning, computer vision, knowledge graph, and distributed systems.



BO DONG (Member, IEEE) received the Ph.D. degree in computer science and technology from Xi'an Jiaotong University, in 2014. He did postdoctoral research at the MOE Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, from 2014 to 2017. He currently serves as the Research Director with the School of Continuing Education, Xi'an Jiaotong University. His research interests focus on intelligent e-Learning and data mining.



BOQIN FENG is currently a Professor with the Department of Computer Science, Xi'an Jiaotong University, China. He is the National Award for Distinguished Teacher (The Ministry of Education, 2003), the Dean of the Computer Teaching and Experiment Center, Xi'an Jiaotong University. He is a Specialist in teaching methods of the fundamental computer technology. He has also authored or coauthored more than 80 research articles.



FANG YANG was born in Zhongwei, China, in 1991. She received the master's degree in engineering from Ningxia University, in July 2018. She is currently an Algorithm Engineer focusing on deep learning. Her current research interests include computer vision, deep learning, and OCR.



BO XU received the M.M.S. degree from Xi'an Communications University. He is currently an Algorithm Engineer focusing on deep learning. His current research interests include computer vision and OCR.

• • •