Fine-grained Retail Goods Recognition based on Deep Learning

Jinghuang Zhang *

School Of Computer Science, South China Normal University, Guangzhou, China

* Corresponding author email: 20183232108@m.scnu.edu.cn

Abstract. Nowadays, much effort has been devoted to image recognition technology, while the application of image recognition technology to the self-service checkout system of retail goods is still a subject with little research and great application value. For the problems of low efficiency and high cost of manual checkout in supermarkets, this study proposes to apply deep learning to commodity picture recognition. Due to the small difference between classes and the large difference within classes of commodity images, commodity image recognition is modeled as a fine-grained commodity image classification task, and the NTS-Net model is used to classify commodity images. The algorithm automatically extracts the partial features of commodity images and classifies them by combining partial features with the features of the whole commodity images. Training is conducted in a self-built commodity dataset, and tests the recognition effect of the model on commodity images from different angles and environments. This study show that the algorithm can effectively and accurately identify commodities, and the model has a good performance in identifying the side of commodities and the pictures of commodities with low brightness.

Keywords: Deep Learning; Retail Goods Recognition; Fine-grained Image Classification.

1. Introduction

Residents are increasingly shopping-demanding due to the improvement in people's living circumstances. At the same time, several drawbacks of the traditional retail sector have risen to the fore, most notably the low efficiency of manual cashier settlement, high labor cost, and poor settlement experience during peak consumption hours. The "new retail" represented by unmanned stores has entered people's lives recently. The commodity identification technology in self-service settlement systems is one of the technical difficulties of an unmanned store. Choosing a simple and fast method of commodity identification can not only reduce labor costs but also significantly speed up the settlement time of consumers' shopping, thereby expanding sales to bring profits to businesses. As a result, the design of an accurate and fast product identification method has important significance for unmanned stores.

Deep learning has facilitated the development of numerous computer vision [1-2] tasks, how to apply deep learning-based image identification to supermarket commodity settlement to increase the accuracy of commodity recognition and settlement speed has become a hot topic for many academics. Liu and Yuan improved the RetinaNet target detection model, combined with the ResNet feature extraction network, and used shelf commodity images for training, with an accuracy of 96.5% [3]; Li and Yan proposed a product image recognition model based on deep residual shrinkage network, which achieved 97.02% accuracy on a dataset containing 51 types of products [4]. Mei Cheng used the AlexNet model to train on a self-built dataset containing 50 commodity categories and achieved an accuracy rate of 81.4% [5]. The above image recognition network can well recognize categories of commodities.

For the commodity images in many retail scenes, however, there are only subtle differences between categories. Many products have a high degree of similarity, reflected in the same packaging type, similar packaging cover, or only differences in packaging specifications, etc. These differences are difficult to identify accurately in traditional image recognition. In addition, the properties of these commodity photos also include significant differences within the categories, including different angles of view, brightness, backgrounds of images, and so on, due to different shooting environments.

The above problems make it difficult to perform this task using traditional image recognition methods alone, and accurately identifying these visually highly similar com classification modity categories can be considered a fine-grained classification task. The difficulty of fine-grained image classification includes accurately locating the discriminative key regions and extracting useful features from the detected key regions for classification.

In this study, the NTS-NET network [6] is applied to the fine-grained classification of product images, which is trained on a self-built retail product dataset; good classification outcomes are also attained. Additionally, the model is a weakly supervised [7] fine-grained picture classification model, which only utilizes the category label information of the images for training. There are also some works [8-9] that utilizes bounding box and annotations solely in the learning stage, in which case, lots of bounding boxes and annotations need to be manually labeled, which will be costly in terms of staffs and financial resources. Therefore, this model is more suitable for recognizing merchandise images in retail scenarios. In addition, the quality of commodity images will be affected by various factors when commodity settlement is carried out in real scenes. For example, the commodities photographed will be at different angles, distances, lighting, etc. This study also compares the effect of commodity recognition under the influence of these factors. The experiments prove that this model achieves good results in the self-built product dataset and can predict the images under different environmental influences well, which verifies the effectiveness and feasibility of the algorithm.

Focusing on the above issues, this paper introduces the basic principle of the NTS-NET algorithm and analyzes its structure in detail in Section 2. A detailed experimental setting and results are subsequently reported in Section 3. At last, an overview of the paper's overall contents is provided and discusses the future research direction of commodity image recognition technology.

2. Methods

2.1 Revisiting NTS-NET

NTS-Net model is put forward based on the following assumption: for the object to be recognized, some areas are more informative than others. That is, these information-rich areas can help to express the object better. Therefore, according to this assumption, the model's classification performance for the object can be effectively improved by fusing the features extracted from the above information-rich areas and the feature of the whole image.

The NTS-Net is composed of three components, including Navigator, Teacher, and Scrutinizer. The Navigator was influenced by the Feature Pyramid Networks (FPN) in its design [10], it will generate several candidate boxes on feature networks of different scales and then score each candidate region according to the pre-designed anchor point. In order to reduce the occurrence of regional redundancy, the method of Non-Maximum Suppression (NMS) is adopted for each candidate box according to its information quantity, and then the top N regions with the largest amount of information are selected to be input into the Teacher network. The Teacher network evaluates the probability that each area belongs to the target area, calculates respective confidence, and evaluates and guides the Navigator network to update the areas with more information. When the Navigator network converges, the Scrutinizer network carefully examines the candidate areas proposed by the Navigator network, enlarges each candidate area to the same size, and fuses the feature of these areas with the feature of the whole picture as the final fine-grained image classification features.

Generally speaking, the NTS-Net model can be regarded as a case of Actor-Critic reinforcement learning [11], and the Navigator network and Teacher network play the roles of actor and critic, respectively. Specifically, improving the positioning ability of the Navigator network for key areas can effectively improve the training and prediction ability of the Teacher network and make the confidence value of its prediction better reflect the amount of key information contained in local areas extracted by the Navigator network. Correspondingly, if the quality of confidence evaluation of the Teacher network can be improved, it will also be beneficial to the training of the Navigator network, so that the area extracted by Navigator will contain more key information. Through this mechanism,

the prediction performance of the above two networks can be effectively improved, and finally, a multi-label classification model that can accurately locate the area with high information in the input image can be obtained.

The main goal of this network is to locate the area with the richest information. Assuming that all areas are rectangular, the A is represented as the set of all areas in a given image. I is an information function for calculating the information a region. C, a confidence function, is defined as a classifier to access how likely it is that a region R belongs to the ground truth clas. The following conditions should be satisfied, as regions with more information should have a higher confidence level:

$$R_1, R_2 \in A \tag{1}$$

$$C(R_1) > C(R_2), I(R_1) > I(R_2) \tag{2}$$

To approximate the information function I and the confidence function C, the network utilizes the Navigator network and the Teacher network, respectively. To keep things simple, M regions are chosen from set A. The Navigator network evaluates the information $I(R_I)$ for each region A_M, and then its confidence $C(R_I)$ will be evaluated by the Teacher network. In order to meet the conditions, this paper adopts the sorting loss function [12] to make the $\{I(R_1), I(R_2), \dots, I(R_M)\}$ and $C(R_1), C(R_2), \dots, C(R_M)\}$ have the same order.

To sum up, the network expresses the M with the richest information predicted via the Navigator network as $R = \{(R_1), (R_2), \dots, (R_M)\}$ and the information of the region as $I = \{(I), (I_2), \dots, (I_M)\}$. Teacher network predicts its confidence as $C = \{(C_1), (C_2), \dots, (C_M)\}$. The central concept behind the algorithm is to make the confidence C have the same order as the information I to optimize the information area. The functions of Navigator loss and Teacher loss are as the formula (3) and (4), respectively.

$$L_I(I, C) = \sum_{(i,s):c_i<c_s} f(I_s, -I_i) \tag{3}$$

The function $f(x) = \max\{1 - x, 0\}$; if $C_s > C_i$, then encourages $I_s > I_i$. The Navigator loss function encourages I and C arranged in the same order.

$$L_C = -\sum_{i=1}^{M} \log C(R_i) - \log C(X) \tag{4}$$

$-\sum_{i=1}^{M} \log C(R_i)$represents the sum of cross-entropy loss for every region, while $\log C(X)$ is the cross-entropy loss of the whole image. When the Navigator network finds the $\{(R_1), (R_2), \dots, (R_K)\}$ with the richest information, the review network generates a fine-granular identification output $P = S\{X, (R_1), (R_2), \dots, (R_K)\}$, and finally uses the cross-entropy loss for classification loss:

$$L_S = -\log S(X, R_1, R_2, \dots, R_K) \tag{5}$$

The purpose is to splice each region with the whole image and then calculate its cross-entropy. The final loss is the weighted sum of the three, as shown in the formula (6). $\lambda = \mu = 1$ in the experimental setup of the original text.

$$L_{total} = L_I + \lambda L_S + \mu L_c \tag{6}$$

2.2 Model Construction for Commodity Recognition

Generally speaking, the overall training method of the model is as follows: for the input commodity image, first, scale the picture to a specific size and then input it into the network; Firstly, the network collects the feture map of this commodity picture, then inputs it into the Navigator network to calculate the information of all areas. Then, NMS selects M regions according to the top information ranking (the information of each region is represented as $I = \{(I), (I_2), \dots, (I_M)\}$), cuts them out from the complete commodity image, adjusts them to the specified size, then inputs them into the Teacher network to obtain the confidence $C = \{(C_1), (C_2), \dots, (C_M)\}$ of these regions. Through optimization, the order of $I = \{(I), (I_2), \dots, (I_M)\}$) and $C = \{(C_1), (C_2), \dots, (C_M)\}$ is the same. Finally, these regions are cascaded with the features of the source picture and input into the Scrutinizer network for classification. Finally, the category of commodity images is predicted.

2.3 Parameter Design

The main hyper-parameters of the NTS-Net model are image input size, m, k, learning rate, weight attenuation, and so on. The input size of the images in the input network is set to 448x448 pixels, and the images in the dataset will be scaled to this size first, and then input into the network for training and prediction. M is the number of high information areas sent by the Navigator to the Teacher, and K is the number selected by the Teacher network from these areas and finally used as input to the discriminant network. The larger the value of K, the more local areas will participate in the prediction. In addition, the experiment uses ResNet-50 [13] pre-trained on ILSVRC2012[14] dataset as the feature extractor, and the parameters in the feature extractor are shared by Navigator, Teacher, and Scrutinizer. The model uses Batch Normalization (BN) [15] to realize regularization. In addition to standardizing the input layer, Batch Normalization (BN) also normalizes the input to each intermediate layer of the network before activating the function. This helps ensure that the output follows a normal distribution with a mean of 0 and a variance of 1, preventing the problem of variable distribution deviation. The optimizer used in this case is SGD with momentum, whose learning rate was fixed at 0.001. Both in supervised learning and deep learning, the learning rate plays a crucial role in determining how fast the target function approaches a local minimum. An appropriate learning rate allows the target function to reach a local minimum at a suitable time. The weight decay value is set to 0.0001, which helps to prevent the model parameters from becoming too large, thereby controlling the complexity of the model and addressing the issue of overfitting.

3. Experiment

3.1 Dataset

A total of 20050 images of the commodity dataset were used in the experiment, including common drinks, food, daily necessities, etc. The training set contains 18000 pictures, 154 categories in total, 115 pictures in each category on average, and 2050 pictures in the test set. All images are collected manually in physical retail stores with natural lighting, which is highly matched with the actual application scene and has excellent adaptability to practice. Before inputting the image data into the network for training, the image size in the dataset is converted to 448×448, and then the image enhancement techniques, such as random horizontal flipping and image normalization, are adopted, and then input into the network for training. Because the model used in the experiment is a weakly supervised model, only the class labels in the dataset are used, so there are only class labels in the dataset, but no bounding box and part annotations.

3.2 Evaluation Indicators

The primary focus of this experiment examines the ability of the classification algorithm and measures the network model performance through the use of accuracy as the key metric. Accuracy is a measure of how well a classifier correctly classifies samples in a given test dataset. It is calculated as the ratio of successfully classified samples to the total samples. Accuracy is positively correlated with the ability of the classifier to successfully identify positive samples. The specific calculation formula of accuracy rate is shown in the formula:

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})} \tag{7}$$

3.3 Experimental Results

3.3.1 Comparison of Classification Accuracy

Firstly, this model is compared with the deep residual networks Resnet50, whose results can be seen in Table 1. ResNet-50 is a powerful baseline and can achieve an accuracy of 89.6%, while the accuracy of NTS-Net can reach 97.8%, an increase of about 8.15%. According to the study results,

the model featured in our work exhibits a better recognition accuracy compared to the deep residual network.

Table 1. Comparison of accuracy with other methods

Algorithm	Accuracy rate
ResNet-50	89.6%
NTS-Net(K=4)	97.8%

3.3.2 Ablation Study of K

In addition, the performance of the model under different K values is compared, whose values refer to how many features of partial areas in the Scrutinizer network participate in the classification. The following Table 2 displays the performance of the model for various values of K. When the complete image is used as the input of Scrutinizer, that is, the hyperparameter k is 0, and the accuracy rate is 91.56%, which is better than the performance of the ResNet-50 models. When k increases from 0 to 2, the accuracy rate reaches 97.3%, which increases by about 5.7%. It shows that the local discriminant features are really helpful for the correct recognition of the whole image, and the network can effectively extract the discriminant area of fine-grained objects. The accuracy rate only rises by 0.5% when K goes up from 2 to 4., but the feature dimension almost doubles, which shows that simply increasing the feature dimension will only get a small improvement.

Table 2. Accuracy with various K

Algorithm	Accuracy rate
NTS-Net(K=0)	91.5%
NTS-Net(K=2)	97.3%
NTS-Net(K=4)	97.8%

3.3.3 Effects Analysis in Different Environments.

Table 3. Recognition accuracy for commodities with different environments

	Frontage	Acc	Side	Acc	Dark	Acc	Blurry	Acc
Totole chicken sauce		100%		83%		93%		100%
Kangshifu sour plum soup		100%		73%		100%		100%
Vitasoy 250ml		100%		100%		100%		79%
Oriental Leaf green tea		100%		68%		100%		100%

The experiment also compares the performance of commodity images in different angles and environments on the model. Four commodities with different outer packaging shapes were selected for prediction, and tested under four conditions: front, side, dark and blurry. Each category was tested with about 30 pictures. The test results are followed below (Table 3).

The front of the package is often the main selling point of the goods, which contains the name of the goods and the brand logo, which are important information for consumers to recognize the goods quickly. The front pictures are usually carefully designed by designers, which can attract consumers' attention and promote sales. On the side of the package, it is mainly the nutritional composition table of the product, or the information of the production date and address, which is challenging to be learned by the model compared with the trademark and renderings in the front. Therefore, the front of the package contains more distinguishing features than the side. In the test, the picture used to test the prediction effect of the side of the product only retains a small part of the front of the product. That is, the front area of the product accounts for a small proportion of the whole picture. From the test reports, it is clear the model can accurately identify those front images of goods, with 100% recognition accuracy for the front of four kinds of goods. In contrast, the recognition effect for the side of most goods is poor, with only 68% recognition accuracy for the goods with the worst recognition accuracy.

For the recognition of the front image of goods with low brightness, the recognition effect is the same as that of the front images with high brightness, which can be accurately recognized. This shows that the brightness has little effect on the recognition accuracy of the model. This may be because the model has learned how to deal with images with various brightness levels in the training process, and the recognition accuracy will not be affected by the different brightness of the images. In addition, the front of the package usually contains more visual features, such as the brand logo, logo, etc. These features are also easier to identify for the model. Therefore, even low-brightness images can be recognized accurately.

In addition, the recognition of low-pixel images is also tested. The pixel size of each group of images is set to 3% of the original product front image size, and the longest side of the images is no more than 80 pixels. According to the test results, in this case, the recognition rate of other commodities reached 100%, except for Vitasoy, whose recognition rate was only 79%. This shows that our model can accurately identify the commodities in low-pixel images and can achieve good recognition results in this case.

4. Discuss

Commodity image recognition is widely used in the retail scene. For example, in e-commerce websites or mobile phone applications, users can quickly search for product information by taking pictures of products or uploading pictures of products. In physical retail stores, product image recognition technology can be used for quick checkout so that consumers can put the products into the shopping cart and then check out quickly through the product image recognition system, which greatly saves waiting time in line. In addition, commodity image recognition technology can also be used for inventory management, helping merchants to quickly scan commodity information and determine the inventory situation, thus improving the efficiency of inventory management. In a word, commodity image recognition technology has a wide application prospect in the retail scene, to significantly make commodity search and checkout more efficient.

In addition, in practical application, this image recognition system can be lightweight deployed, and it can use a cloud server with powerful computing power to support the running of the model. Lightweight cloud deployment can also allow users to access through mobile devices or web pages, so there is no need to build a local server to support the operation of the image recognition system. In addition, by adjusting the size and complexity of the model through lightweight deployment, the system is easier to operate and maintain. Another advantage is lower cost. Lightweight deployment can reduce the required computing resources, which can reduce the cost of deploying the system. In

addition, lightweight deployment can also bring higher scalability. Because less computing resources are used, the system can be easily extended to more users or larger datasets. In short, the lightweight deployment of commodity image recognition system in the cloud can provide faster deployment time, lower cost and higher scalability, which is very suitable for retail store.

When this model is put into practice, it still needs to be further improved: the types of commodity image dataset and the number of images still need to be expanded. The types of common commodities in supermarkets are far more than those included by the commodity image dataset constructed in our study. For the better fulfillment the practical application, it is necessary to supplement the commodity image data not involved in the dataset, so that the coverage of the system is wider. In addition, the image used in the current research only contains a single commodity, and many different commodities are usually purchased when shopping, so how to recognize of multiple commodities needs further research.

5. Conclusion

Because of the small differences between classes and large differences within classes of commodity pictures, commodity picture recognition becomes a challenging task. Therefore, this study model commodity image recognition as a fine-grained commodity image classification problem, and use the NTS-Net model to complete the classification of commodity images. This algorithm has the capability to automatically extract the local features of commodity images and classify them by fusing the features of the complete image. The training was conducted on a self-built commodity dataset. The recognition effect of the model on commodity pictures in different angles and environments is also tested. The algorithm has been shown to be effective and accurate in identifying classified goods, even the side of goods and low-brightness pictures can be well identified. Besides being applied to the self-checkout system in unmanned supermarkets, this technology can also realize faster and more accurate inventory management and commodity sales analysis for retailers.

References

[1] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep con-volutional neural networks. In: NIPS. pp. 1097–1105 (2012).

[2] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CVPR (Nov 2015).

[3] Liu,Yuan. Product Recognition on Shelves Based on Deep Neural Network [J]. Packaging Engineering,2020,41(01):149-155.DOI:10.19554/j.cnki.1001-3563.2020.01.023.

[4] Li, Yan. Commodity Image Recognition Based on Deep Residual Shrinkage Network [J]. Journal of Test and Measurement Technology,2021,35(04):294-299+322.

[5] Mei, Lyu. Research on Commodity Image Recognition Based on Deep Learning [J]. Mechanical & Electrical Engineering Technology,2018,47(09):28-31+151.

[6] Yang, Ze, et al. "Learning to navigate for fine-grained classification." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[7] Zhou, Zhi-Hua. "A brief introduction to weakly supervised learning." National science review 5.1 (2018): 44-53.

[8] Branson, S., Horn, G.V., Belongie, S., Perona, P.: Bird species categorization using pose normalized deep convolutional nets. In: BMVC (2014).

[9] Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based rcnn for fine-grained detection. In: ECCV (2014).

[10] Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (July 2017).

[11] Grondman, Ivo, et al. "A survey of actor-critic reinforcement learning: Standard and natural policy gradients." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42.6 (2012): 1291-1307.

[12] Liu, T.Y.: Learning to rank for information retrieval. Found. Trends Inf. Retr. 3(3), 225–331 (Mar 2009).

[13] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016).

[14] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015 Dec;115(3):211-52.

[15] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International conference on machine learning. PMLR, 2015.