International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n1.34
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Optimizing Matrix Capsule Networks for Contraband detection Research

Zhiming Yan[1, 2], Xinwei Li[1, 2, *], Yi Yang[1, 2]

[1]School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, China;
[2]Henan Key Laboratory of Intelligent Detection and Control of Coal Mine Equipment, Jiaozuo, China

ABSTRACT

An optimized matrix capsule network is proposed for the problem of contraband in parcels with different poses, different sizes, random occlusion and sample imbalance. The network improves the recognition accuracy with the help of the matrix capsule network's recognition ability for object poses, and is mainly composed of a multi-branch feature extraction network and a side branch matrix capsule network, which is used to extract large and small targets; the side branch matrix capsule network uses a larger capsule convolution kernel, which is capable of detecting larger targets, and uses the operation of randomly discarding the capsules in the side branch, which makes the parameter amount to be reduced while enhancing the learning ability of the network. The region of interest of the network is obtained by using heat map approach with the help of weight back propagation mapping back to the original map to localize the contraband. Through a large number of experiments on the SIXray dataset, it is proved that the network in this paper improves the detection accuracy by 9.43% and the processing speed of the model by about 1/3 compared with the original capsule network.

KEYWORDS

Contraband classification; Matrix capsule network; Multi-feature extraction; Discarded Capsules

1. INTRODUCTION

Security check is the most important means to ensure the safety of transportation, and in public places such as airports and railway stations, X-ray machines are the most important way to detect luggage and parcels. X-ray images are X-ray images when they absorb different amounts of energy through items of different densities and materials, so that the energy that reaches the receiving medium is different. At present, the items in the baggage package are mainly inspected manually through X-ray images, but this method has continuously promoted the development of automatic detection technology due to factors such as high labor intensity and low detection efficiency [1]. Due to the diversity and complexity of the items in the luggage package and the characteristics of the X-ray images, the X-ray images have the characteristics of different attitudes, different sizes, superimposed occlusion and unbalanced samples. These circumstances pose a great challenge to the automated detection technology of aviation [2].

In addition to the traditional X-ray contraband detection methods, there are also semi-manual detection methods that rely on image processing, such as feature matching methods based on visual bag of words [3], support vector machine (SVM) detection methods [4], image segmentation methods [5], millimeter wave imaging methods [6], foreground and background separation methods [7], and feature extraction methods using neural networks after image enhancement [8].

The traditional manual and semi-manual contraband detection methods are slow to detect and cannot meet the requirements of the rapidity of security inspection, so a rapid security inspection method is urgently needed.

With the development of deep learning, convolutional neural networks have gradually become the most important method for contraband detection. S. Akcay et al. [9] introduced deep learning into X-ray contraband detection for the first time, and the detection network was migrated from the AlexNet network, and its detection speed was qualitatively improved compared with traditional manual and semi-automatic detection methods; Gao Qiang et al. [10] carried out a research on the automatic detection of dangerous goods at airports based on CNN network; in order to make up for the mismatch between positive and negative samples of contraband in X-ray images, YANG J F [11] The generative adversarial network GAN is used to increase the image dataset containing contraband and estimate the pose of the obtained contraband images in the Cartesian coordinate system, and the original contraband dataset and the generated dataset are used at the same time during training, which solves the problem of uneven distribution of contraband. Zhang et al. [12] proposed an asymmetric convolutional multi-view neural network (ACM-Net), which is based on the deep learning network SSD [13], which fuses feature layers of different dimensions and uses dilated convolutional layers to obtain the feature relationship between local and global, so as to reduce the impact of mutual occlusion between objects on classification accuracy. Caijing Miao et al. [14] proposed the Class Balanced Hierarchical Refinement (Resnet-CHR) method based on the residual network Resnet [15], which fuses features of different scales through residual branching, and uses a hierarchical refinement strategy to enhance the stability of samples in the case of imbalance. Based on the Cascade RCNN model, You Xi et al. [16] proposed a spatially adaptive attention module by introducing deformable convolution, which effectively reduced the problem of false detection and missed detection of contraband. Su Xingwang et al. [17] proposed a Yolov5S model combining deformable convolution and attention mechanism improvement for contraband detection in order to solve the problems of variable shape scale and overlapping occlusion of contraband. Li et al. [18] improved yolov7 by adding a high-resolution detection head and adding a MobileNetViTv3-block module at the end of the backbone to achieve a balance between detection speed and accuracy.

In summary, deep convolutional neural networks can extract deeper semantic information from images, and their detection speed is faster and more accurate than manual classification. However, the current contraband detection still faces some problems, the posture and size of contraband are different, occluding each other, and the background and foreground are complex, and the contraband detection in these complex scenarios makes the automatic detection face great challenges.

And most importantly, these scalar-based CNN models cannot actively identify the different postures of objects, and need a large amount of data for generalization, which is not efficient, and the positioning mode of the calibration frame is also relatively cumbersome.

2. NETWORK CONSTRUCTION

2.1. Matrix Capsule Network

The scalar CNN network cannot recognize the posture of different objects, and it has the invariance of recognition, that is, the recognition of different postures of objects needs to be generalized through the dataset of different angles and different postures during training. The vector capsule network [19] uses vectors to represent various properties of objects, the length of the vector represents the probability of the object's existence, and the direction of the vector represents the object's color, pose, texture, and other properties. The matrix capsule network [20] uses the 4×4 attitude matrix P and an activation probability a to form a capsule, where the attitude matrix is used to represent the different postures of the object and the activation probability is used to represent the existence probability of the object. This allows the matrix capsule network to achieve high detection accuracy without the

need for a large amount of data generalization. At the same time, there are n transformation parameters of the matrix capsule network, which has higher efficiency than the n2 of the dynamically routed capsule network.

The matrix capsule network is mainly composed of the following four parts: the traditional convolutional layer, which uses a 5×5 convolutional layer for feature extraction; a primary capsule layer for the generation of capsules; convolutional capsule layer, forward calculation using EM algorithm; Classify the capsule layer, which is used to classify and output the class probability. The network architecture is shown in Figure 1.

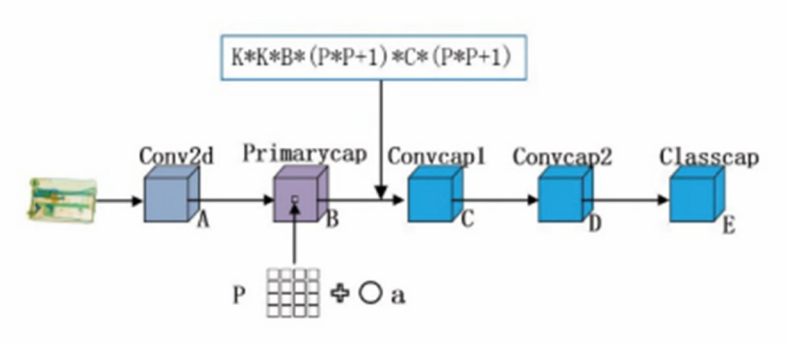

Figure 1. Matrix Capsule Network Architecture

The network uses two convolutional capsule layers for feature extraction and forward propagation.

The classification capsule layer evolves from the convolutional capsule layer, and the 1×1 convolution of the convolutional capsule layer is used to calculate the class probability of the output. Matrix capsule networks are often used for classification tasks and cannot be used directly in detection tasks.

2.2. EM Routing

The EM (Expectation Maximization) algorithm consists of E (Expectation) steps and M (Maximization) steps. For capsule element j, multiple low-level matrix capsules are input, and one matrix capsule is output through EM routing. Among them, the attitude matrix of the low-level capsule needs to be transformed by the linear transformation of the visual coordinates to obtain the voting matrix, while the activation probability is not transformed. The matrix w_{ij} represents the visual coordinate transformation from the lower capsule i to the upper capsule j The activation probability of a high-level capsule a_j indicates the probability that the feature corresponding to capsule element j will be activated. The EM routing process is shown in Figure 2 below.

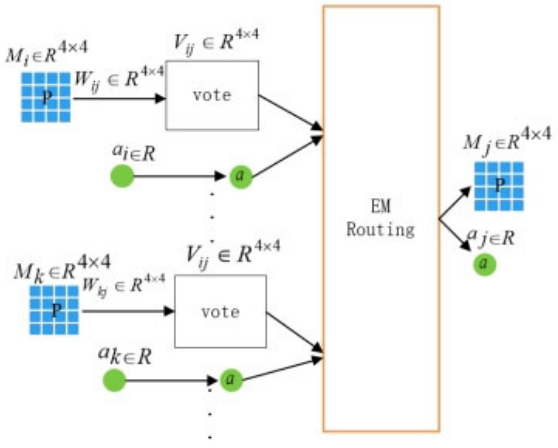

Figure 2. EM Routing Voting Diagram

Suppose w_L and w_{L+1} represent low-level capsules and high-level glue, respectively

The set of sacs, V_{ij}^h represents the h_{th} element after the voting matrix V_{ij}^h transformed into vectors, assuming that the number of routing iterations is T, the specific iteration process of EM routing is divided into two steps.

M steps (a, R, V, j) are defined as; Traversing the capsule $i \in w_L$ to correct the connection probability R_{ij}, see equation (1):

$$R_{ij} = R_{ij} * a_i \tag{1}$$

Where R_{ij} is the connection probability, and a_i is the activation probability of the underlying capsule.

For any hth element in the voting matrix V_{ij}^h, calculate the mean μ_j^h and variance $(\sigma_j^h)^2$ of the Gaussian mixture model, see equation (2):

$$\mu_j^h = \frac{\sum_i R_{ij} V_{ij}^h}{\sum_i R_{ij}},\ \left(\sigma_j^h\right)^2 = \frac{\sum_i R_{ij}\left(V_{ij}^h - \mu_j^h\right)^2}{\sum_i R_{ij}}. \tag{2}$$

Where μ_j^h is the calculated mean, $(\sigma_j^h)^2$ is the calculated variance, and R_{ij} is the connection probability of the bottom capsule and the upper capsule.To calculate the cost for element x, see Equation (3):

$$\cos t_j^h = \left(\beta_u + \log\left(\sigma_j^h\right)\right)\sum_i R_{ij}. \tag{3}$$

Where $\cos t_j^h$ is the calculated cost, β_u is the coefficient obtained by the network in the learning process, and σ_j^h is the mean square error obtained from the Gaussian mixture model.Calculate the activation probability based on the cost, see equation (4):

$$\alpha_j = sigmoid\ \lambda\left(\beta_\alpha - \sum_h \cos t_j^h\right). \tag{4}$$

Where a_j is the probability of activation of the high-level capsule according to the calculation, the β_a and the β_u can be learned, and λ is the annealing coefficient, which increases with the increase of the number of iterations.

Estep (μ, σ, a, V, i) is defined as:Traversing the capsule $j \in w_{L+1}$ for the voting matrix from capsule i to capsule j, calculate the joint probabilities of the Gaussian distribution corresponding to the 16 elements, see equation (5):

$$P_j = \frac{1}{\sqrt{\sum_h^{16} 2\pi\left(\sigma_j^h\right)^2}} \exp\left(-\sum_h^{16} \frac{\left(V_{ij-}^h \mu_j^h\right)^2}{2\left(\sigma_j^h\right)^2}\right). \tag{5}$$

The P_j is the joint distribution probability from the bottom capsule i to the upper capsule j. Traversing capsule $j \in w_{L+1}$, updating the connection probability from capsule i to capsule j, the matrix capsule network iteratively executes steps M and E steps to perform forward propagation, where the number of iterations of the EM route is T, and the default is 3.

2.3. Loss Function

Matrix capsules use spread loss to avoid the sensitivity of the network to hyperparameters in the early stages of training. The loss function is constructed based on the fold function [21] and simplified on the basis of the vector capsule loss function. For a sample, if the activation probability of the target category t of the label in the capsule output is a_t and the activation probability of the other categories i is a_i, then the loss L_i between category i and target category t is calculated as shown in equation (6):

$$L_i = \left(\max\left(0, m-\left(\alpha_t - \alpha_i\right)\right)\right)^2. \tag{6}$$

When the activation probability of the target class t is less than that of the other class i, the loss value increases. where m is the limit boundary of the fold, which gradually increases from 0.2 to 0.9 during the training process, and the increasing fold boundary is conducive to gradually softening the probability distribution of the output and enhancing the generalization ability of the model. All losses for a sample are the sum of all category losses, as shown in equation (7):

$$L = \sum_{i \neq t} L_i. \tag{7}$$

If the value is fixed m=1, the function L is formally equivalent to the loss of the multi-class support vector machine.

In summary, the matrix capsule network relies on the matrix capsule to identify the posture and existence probability of objects, and does not require a large amount of data for generalization. However, the feature extraction layer is relatively simple and cannot deal with the classification problem in complex backgrounds. The matrix capsule layer structure is fixed, and multi-feature extraction cannot be achieved. The capsule will be expanded during forward operation, which makes it have a huge amount of parameters and calculations, which makes the capsule channel not too deep and the capsule structure difficult to improve. Matrix capsules are used for classification, which cannot fully meet the requirements for the detection and localization tasks in this paper. How to solve the above problems and use them for contraband detection with the advantages of matrix capsules is the next problem to be solved in this article.

3. MMCN NETWORK CONSTRUCTION

Therefore, this paper proposes a multi-branch matrix capsule network (MMCN). The structure of the network is shown in Figure 3 below:

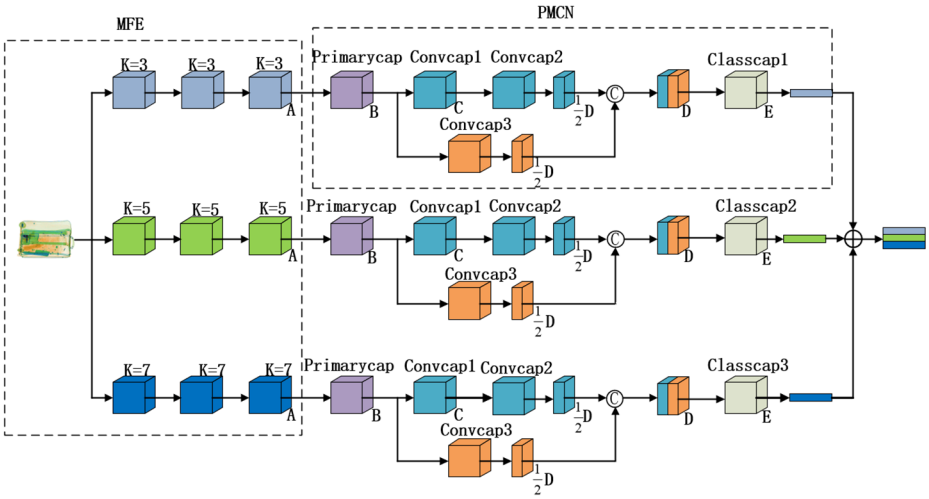

Figure 3. MMCN Network Architecture

The network structure is divided into two parts: the multi-feature extraction network MFE and the side-branch matrix capsule network PMCN.

3.1. MFE Multi-feature Extraction

The feature extraction part designed in this paper is derived from the multi-feature extraction network Inception V1 [22], and its module structure is shown in Figure 4.

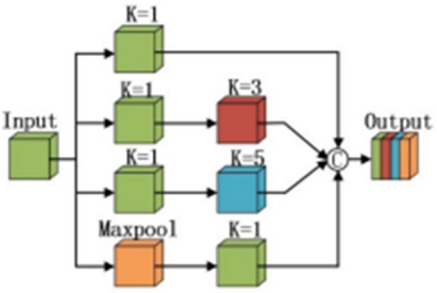

Figure 4. Inception V1 Structure Diagram

The InceptionV1 module uses a 4-branch structure. This method of multi-feature extraction enriches the feature extraction, and at the same time, it almost does not increase the number of parameters and calculations, which is suitable for multi-target detection tasks.

In this paper, a multi-feature extraction network (MFE) is designed, and its network structure is shown in Figure 5. First, after the image is input, three convolutional kernels of size 3, 5 and 7 are used. The convolutional layer receptive field of 3×3 is small, which is suitable for extracting the feature information of small targets, such as guns and pliers. The 5×5 convolutional layer receptive field is large, which is suitable for extracting the feature information of larger targets, such as scissors and knives. The convolutional layer of 7×7 has a larger receptive field and is suitable for extracting feature information from larger objects, such as knives and wrenches, while being able to extract background information. Different convolutional layers are used to extract the underlying feature information, which avoids feature loss, and this extraction method is equivalent to separating the features and backgrounds of contraband of different sizes, which can improve the detection rate of the network.

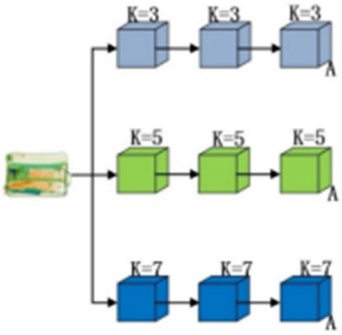

Figure 5. Multi-feature Extraction Structure

After the feature extraction of different convolutional layers, this paper does not fuse them in the channel dimension, but adopts a distributed propagation method, and each branch is calculated forward separately, which is related to the way of extracting the body information in the matrix capsule network. Different from the scalar convolutional neural network pixel-level feature extraction, the matrix capsule is sensitive to the whole pose relationship between objects, which is located and classified by recognizing the object posture, if the feature information extracted by different convolutional layers is fused in the channel, the pose information between different objects will change, and the original unrelated objects will overlap, which is called the "overlap effect" in this paper, which is a big interference for the matrix capsule. Therefore, the multi-feature extraction

network MFE designed in this paper uses multi-branch feature extraction, which is directly sent to different collateral matrix capsule layers for feature extraction and classification after feature extraction from different branches.

3.2. Construct A Collateral Matrix Capsule Network

The collateral matrix capsule network consists of three parts, the primary capsule layer, the convolutional capsule layer that can change the feature map, and the classification capsule layer.

3.2.1. Capsule Layers With Variable Size Of Feature Diagrams

Due to the fixed convolutional capsule layer network structure of the traditional matrix capsule, the convolution of 3×3 is used for feature extraction, the size of the convolution kernel cannot be changed arbitrarily and the padding is used, and the size of the output feature map is fixed, and the feature fusion cannot be carried out. This paper modifies the network architecture of the convolutional capsule layer, and adds an improved PAD structure, that is, the capsule is randomly discarded to achieve the purpose of changing the output size of the feature map, the process is shown in Figure 6, similar to the convolution adding a blank part to change the size of the output feature map, this paper adopts the operation of discarding the capsule, so that they do not participate in the operation, so as to change the size of the output feature map, and the influence of the design drop mode on the output size is shown in equation (8):

$$output = \frac{input - kerne_{size} - 2*pad}{stride} + 1. \tag{8}$$

where output represents the size of the output feature map, input represents the size of the input feature map, $\ker ne_{size}$ represents the size of the convolution kernel of the convolution capsule, pad represents whether to drop the capsule and the number of capsule layers discarded, when the parameter is set to non-0, it represents a random discard of a part of the capsule so that they do not participate in the operation, and stride represents the step size of the convolution capsule. The design of the new pad structure enables the convolutional capsule layer to output feature maps of different sizes, which can be spliced at will like the convolutional layer of CNN network, and the feature maps obtained by using different convolutional capsule layers can also be spliced and fused to obtain richer feature information. At the same time, the discarded capsule can reduce the amount of computation and parameters, increase the generalization of the network, and also reduce the overfitting of the network structure too deep.

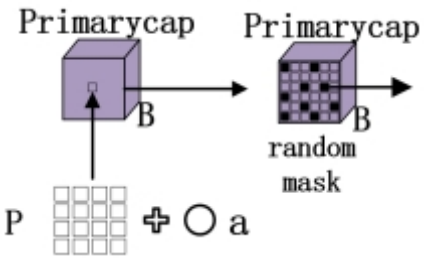

Figure 6. Discard Capsules

3.2.2. Collateral Matrix Capsule Network

In this paper, after redesigning the convolutional capsule layer, a collateral matrix capsule network (PMCN) is designed, and the network structure is shown in Figure 7 below:

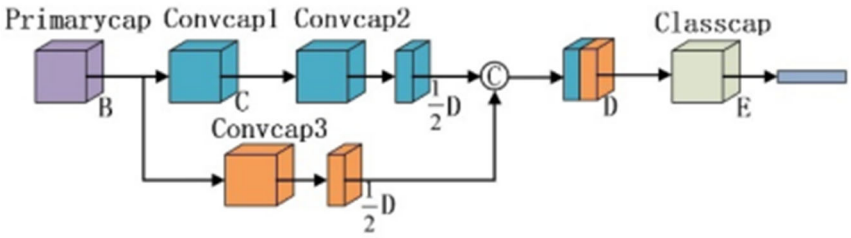

Figure 7. Parabranch Matrix Capsule Network

The network joins a 5×5 bypass branch that uses a capsule drop operation while using a larger convolutional capsule kernel, and the main and bypass branches use half the original number of channels to balance accuracy and operational efficiency. On the one hand, this method can promote feature extraction, and on the other hand, the operation of discarding the capsule does not increase the number of parameters too much.

3.3. Multi-branch Matrix Capsule Network

After the MFE network performs multi-feature extraction, a multi-branch matrix capsule network is used in this paper, and its network architecture is shown in Figure 8 below:

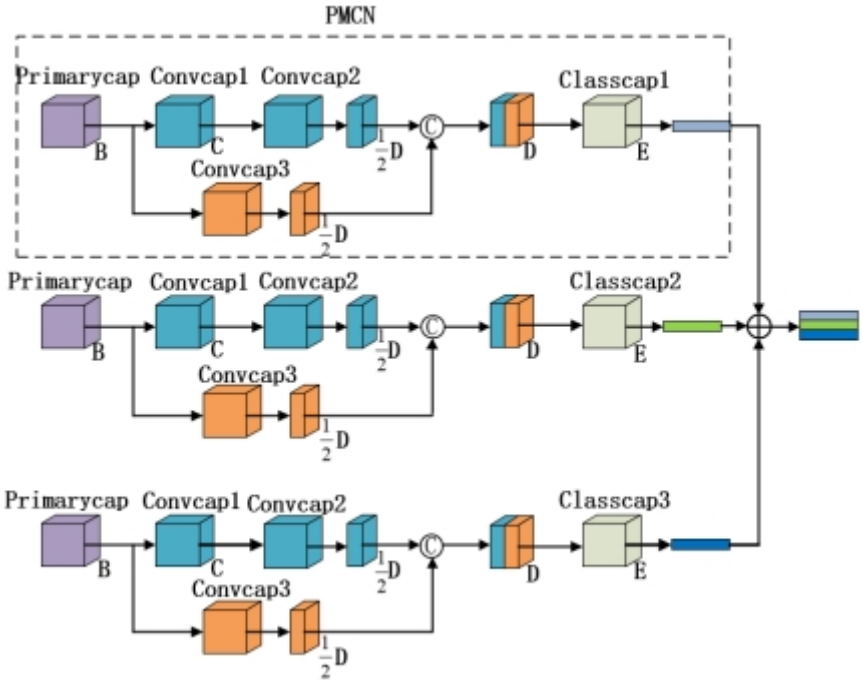

Figure 8. Multibranch Matrix Capsule Network

This multi-branch structure is advantageous for the matrix capsule to recognize objects of different sizes, the recognition process does not interfere with each other, and at the same time makes the network have high computing efficiency, each branch carries out feature extraction and classification at the same time, and finally the classification probability is added and normalized to obtain the final classification probability, and the network model of each branch can be designed very small to reduce the number of parameters and improve the running speed of the whole network.

3.4. Positioning With Grad-cam

Since the matrix capsule network is used for classification tasks and does not have the ability to locate goods, the MMCN model designed in this paper uses the Grad-cam heat map [23] method for contraband positioning, which is simpler than the calibration box method and does not need to redesign the network architecture.

Grad-cam is a weakly supervised visualization model, after the MMCN model training reaches the maximum number of iterations, the trained model is saved, and then the pre-trained model is loaded, and the feature weights of the last convolutional layer of the 5×5 branches of the model are extracted, and then the region of interest of the model is mapped back to the original image through backpropagation to obtain the heat map of the region of interest of the model.

4. EXPERIMENTS

4.1. Experimental Environment And Dataset

The platform for this experiment is ubuntu18.04 operating system, the hardware uses Intel Core i7-8700K CPU, the main frequency is 3.70GHz, 12 threads, the graphics card is NVIDIA GTX-1080Ti, 12G video memory. The Pytorch deep learning framework was used to complete the network design and experiments using the Python programming language. The batch size of the experiment is set to 16, the number of iterations T of the EM algorithm is set to 2, the optimizer is selected as Adam, the initial learning rate of the network is int_lr 3e-2, the learning rate decay is exponentially decayed, the decay rate is 0.96, the decay is performed every 2000 steps, and the epoch of the training period is 500.

The dataset used in the experiment is the SIXray dataset, which was made by Meio Jane et al., which was actually collected from the subway station, with a total of 1059231 images, including 8929 images of the positive sample containing contraband, including knives, guns, wrenches, scissors, pliers and hammers 6 types of contraband, contraband has the characteristics of a wide variety, serious occlusion and complex foreground background, which can be as close to the actual situation as possible. Due to the small number of contraband with positive labels, in order to be closer to the actual situation, the contraband dataset was scaled, rotated at will, changed the color brightness and randomly cropped, so as to enhance the generalization ability of the model. The dataset is divided according to the ratio of training set:test set = 8:2, and a part of the contraband dataset is shown in Figure 9 below:

Figure 9. Partial Sample Of Contraband

4.2. Experimental Results And Analysis

Firstly, the initial parameters of the network are set, and the number of channels and the size of the attitude matrix of MMCN are determined through experiments. Then, comparative experiments and positioning experiments were carried out to highlight the accuracy and speed of MMCN, experiments

on other datasets to highlight the excellent generalization ability of MMCN, and data imbalance experiments were set to simulate the detection of contraband by MMCN in the actual scene. Finally, the effectiveness of the improvement in improving the detection accuracy of contraband was verified by ablation experiments.

4.2.1. Network Parameter Settings

In order to analyze the influence of the number of channels (A, B, C, D) and attitude matrix (P*P) of MMCN on the classification accuracy, a comparative experiment was designed, and the classification results obtained with a maximum number of iterations of 10 are shown in Table 1.

Table 1. Influence Of Channels And Pose Matrix On Classification Accuracy

(A,B,C,D)	(P*P)	Acc/%	Params/M	Images/N
2,2,2,2	2*2	52.17	1.3+8.5*e(-3)	190
	3*3	55.48	1.3+35.4*e(-3)	180
	4*4	58.04	1.3+102.51*e(-3)	179
8,8,8,8	2*2	61.23	1.4+132.75*e(-3)	173
	3*3	60.55	1.4+531*e(-3)	140
	4*4	62.99	1.4+1535.1*e(-3)	118
16,8,8,8	4*4	60.28	1.6+1559*e(-3)	118

(A, B, C, D) represents the number of channels of the model, (P*P) represents the size of the pose matrix of the matrix capsule, Params represents the number of parameters of the model (the first value represents the parameter quantity of MFE, the second parameter represents the parameter quantity of PMCN), Images represents the number of images processed per second, and Acc (Average classification accuracy) represents the average classification accuracy.

Through the experimental results, it can be seen that the classification accuracy of the model is related to the number of channels and the attitude matrix of MMCN, and a larger number of channels can improve the classification accuracy, but will increase the number of parameters and calculations. The attitude matrix can characterize the attitude of the object, and the larger attitude matrix can represent more attitude information of the object, which is beneficial to improve the detection accuracy, and its minimum is 2*2, which is characterized by the coordinate values of the upper left (x1, y1) and lower right (x2, y2) of the object. When the number of channels is small, the use of smaller and larger attitude matrices has little effect on the detection accuracy, while when the number of channels is increased, the smaller attitude matrix is beneficial to reduce the number of parameters and improve the operation speed of the model. In order to balance the detection accuracy and processing speed, the number of channels used in this paper (model=A, B, C, D) is 8,8,8,8, and the attitude matrix size (P*P) used is 4*4.

4.2.2. Comparative Experiments

In order to compare the performance of the proposed method, DenseNet121[24], Resnet50[25], InceptionV3[26], MobilenetV3[27], EfficientNetV2[28], and Matrix-Capsules networks were selected to compare with the MMCN networks proposed in this paper. DenseNet121 enhances forward propagation, uses dense connection, and greatly reduces the number of parameters. Resnet50 uses a residual network connection, and its residual branches can avoid gradient vanishing and gradient explosion, and at the same time, it can extract features and fuse the main branch to achieve better feature extraction capabilities. Inception V3 uses a multi-feature fusion method, which uses multiple Inception modules to extract features, each Inception module uses a 1×1 convolutional layer for dimensionality reduction, and at the same time, the features extracted by multiple branches are fused, which can extract features well without increasing the amount of computation, and achieve multi-feature extraction and feature fusion. MobilenetV3 uses AutoML to find the optimal neural network structure for a given problem; EfficientNetV2 uses an improved progressive learning method to dynamically adjust the regularization method according to the size of the image, which improves

the training speed and accuracy. Matrix-Capsules uses capsule neurons to extract features. The original parameters of each algorithm were selected for experiments to achieve the best performance, and the experimental results are shown in Table 2.

Table 2. Comparative Test Of Classification Results

Algorithm	Classification accuracy of individual contraband %					Average classification accuracy
	Knife	Gun	Scissor	Pliers	Wrench	
DenseNet121	53.22	92.02	83.53	46.10	49.67	66.97
Resnet50	75.00	86.21	64.32	88.33	48.22	72.41
InceptionV3	62.50	90.68	80.00	46.10	51.41	66.14
MobileNetV3	60.46	85.88	75.89	73.88	24.29	64.08
EfficientNetV2	70.86	94.56	83.49	74.26	77.34	77.34
Matrix-Cap	84.90	95.21	55.12	43.38	39.29	64.14
MMCN(ours)	94.00	95.40	56.80	69.20	52.40	73.57

From the experimental results in Table 2, it can be seen that the matrix capsule network can recognize different postures of objects, but it is limited to object classification with simple background and single object, and the detection accuracy of the simpler knife and gun is greatly improved compared with the other five algorithms, but in the face of contraband detection in complex background, its classification accuracy is not only the same as that of MobilenetV3, but also needs to be improved compared with the other four networks, such as scissors, pliers and wrenches. Compared with the original, the average classification accuracy of MMCN proposed in this paper is increased by 9.43%, especially on knives, pliers and wrenches, which are greatly improved, by 9.1%, 25.74% and 13.14%, respectively. The detection accuracy of the knife is increased by 20%~40%, the detection accuracy of the pliers is increased by about 20%, and the average classification accuracy is also significantly improved.

Table 3 shows the number of network parameters for each method, the number of params, the number of floating-point runs, FLOPs, and the average GPU run time for processing an image.

From the results in Table 3, it can be seen that the parameter quantity of the proposed method is slightly higher than that of the matrix capsule network, but the parameter quantity is still small among the 7 networks, and its computational cost is also small among the 7 models, second only to MobilenetV3, which makes the training speed of the model very fast, and the average time to process an image is only 0.008s.

Table 3. Network Complexity Comparison Test

Model	Img_size	Params/M	FLOPs_G	Times/s
DenseNet121	224*224	7.98	5.8	0.77
Resnet50	224*224	23.52	3.86	0.012
InceptionV3	229*229	24.00	15.5	0.61
MobileNetV3	224*224	4.00	0.48	0.0104
EfficientNetV2	224*224	40.74	3.42	0.0108
Matrix_Capsules	128*128	1.66	4.90	0.012
MMCN(ours)	128*128	2.93	3.05	0.008

4.2.3. Contraband Testing Experiments

In order to locate the location of contraband, this paper uses Grad-cam to visualize the parts of the network of interest through heat maps. The original image of the experiment and the resulting heat map are shown in Figure 10.

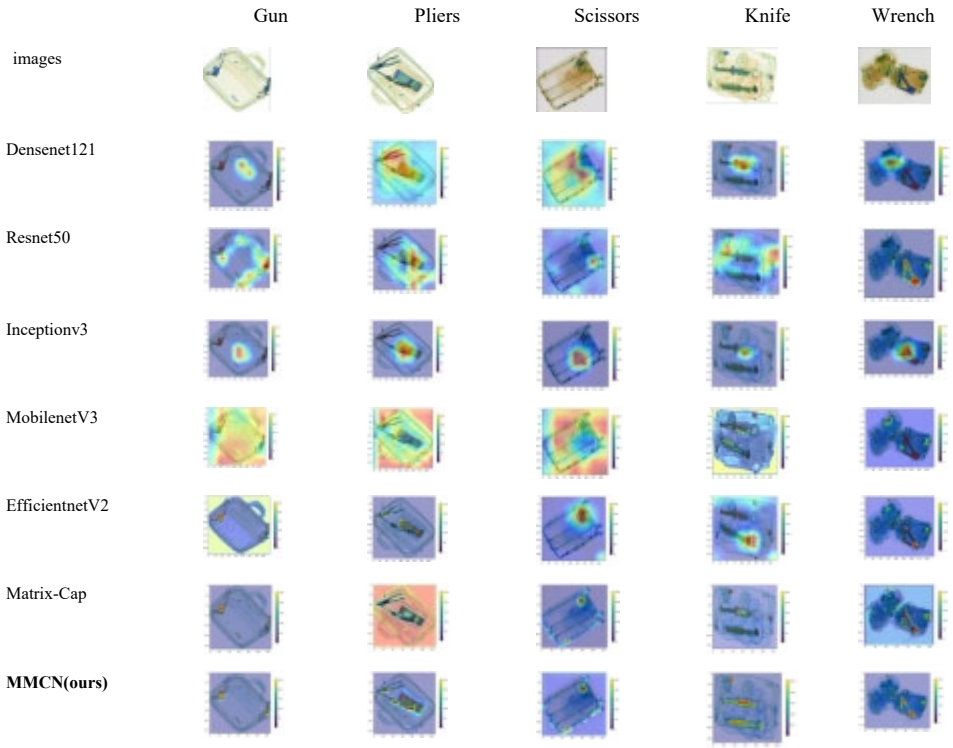

Figure 10. Results Of Contraband Detection

From the heat map obtained in Fig. 10, it can be seen that the method designed in this paper can locate the contraband more accurately than other methods, especially when the size of the contraband is large and the number of contraband is large.

4.2.4. Model Generalization Ability Experiment

In order to test the generalization ability of the proposed method, experiments were carried out on the GDXray [29] dataset, which contains five types of X-ray images, namely castings, welds, luggage, natural objects, and backgrounds, and the X-ray images of luggage in the five categories are used in this paper, and compared with the other six networks, and the results are shown in Table 4.

Table 4. Experimental Results Of GDXray Dataset

Model	Acc/%
DenseNet121	96.10
Resnet50	96.23
InceptionV3	96.07
MobileNetV3	96.73
EfficientNetV2	98.46
Matrix-Cap	97.15
MMCN(ours)	98.62

As can be seen from Table 4, the detection accuracy of the proposed method on the GDXray dataset is 1.47% higher than that of the matrix capsule network, which is basically the same as that of EfficientNetV2 and about 2.5% higher than that of the other four networks, indicating that the proposed method has strong generalization ability.

4.2.5. Data Imbalance Experiments

In order to verify the detection effect of the proposed method in the case of data imbalance, the SIXray10 and SIXray100 datasets were redesigned and experiments were carried out on this basis. Assuming that there are 100 images containing contraband, SIXray10 means that there are 1,000

images that do not contain contraband, and SIXray100 means that there are 10,000 images that do not contain contraband. The experimental results are shown in Table 5.

As can be seen from the results in Table 5, the matrix capsule network has a better recognition effect than the convolutional neural network when the data is unbalanced, thanks to its ability to recognize the object pose. On the basis of the MMCN obtained in this paper, the detection accuracy of the data is unbalanced, which effectively reduces the false detection rate. The false detection rates on SIXray10 and SIXray100 datasets are reduced by 18.95% and 5.14%, respectively, indicating that the method has better adaptability in the face of data imbalance.

Table 5. Unbalances Experimental Results

Model	SIXray 10%	SIXray 100%
Densenet121	35.25	50.00
Resnet50	26.37	36.20
InceptionV3	25.67	43.47
MobileNetV3	23.11	30.23
EfficientNetV2	15.30	44.20
Matrix-Cap	28.67	34.35
MMCN(ours)	9.72	29.21

4.2.6. Ablation Experiments

In order to further prove that the proposed method can effectively improve the detection accuracy of contraband, an ablation experiment was designed, and the results are shown in Table 6.

Table 6. Ablation Test Results

	MFE				
	3*3	5*5	7*7	PMCN	Acc/%
Baseline+	-	√	-	-	64.14
	√	√	-	-	68.25
	√	√	√	-	70.12
	√	√	√	√	73.57

From the results in Table 6, it can be seen that the average classification accuracy is only 64.14% when only 5*5 feature extraction branches are used, and the average classification accuracy increases by 4.11% after adding 3*3 feature extraction branches, and 1.87% after adding 7*7 branches, indicating that the classification accuracy can be effectively improved by using multiple branches. On this basis, the average classification accuracy is increased by 3.45% by the introduction of the bypass branch structure, which indicates the effectiveness of the bypass branch in improving the average classification accuracy.

5. CONCLUSION

In this paper, an optimized matrix capsule network is designed, which is mainly composed of a multi-feature extraction module and a side-branch matrix capsule module, which can cope with the different sizes of contraband, and the side-branch matrix capsule network can cope with different attitudes, overlapping occlusions, and unbalanced data samples. The use of Grad-cam to locate contraband reduces the cost of labeling without the need for object labeling frames. Sufficient experiments are carried out in the above cases, and the results show that the proposed method has significant improvement in accuracy and computational efficiency. In the security scene, X-ray image samples are seriously unbalanced, and how to efficiently and accurately uncover a very small amount of contraband in a large number of items requires further in-depth research.

REFERENCES

[1] WANG Yanqing, REN Jinjin, YU Qin. Impact Analysis of X-ray Machine Conveyor Belt Speed on Airport Security Missing Rate[J]. China Transportation Review, 2017, 39(05):55-59. (in Chinese)

[2] ZHAO Zhenwu, WANG Junjie, ZHANG Yi. Research on reliability optimization of airport passenger security system[J]. China Safety Science Journal, 2023, 33(8): 182-189. (in Chinese)

[3] STERCHI Y， HATTENSCHWILER N， MICHELS， et al. Relevance of visual inspection strategy and knowledge about everyday objects for X-ray baggage screening［C］// International Carnahan Conference on Security Technology，Oct. 23-26，2017，Ma-drid, Spain. IEEE, 2017:1-15.

[4] YAN W，JING H. Object detection in X-ray images based on object candidate extraction and support vector machine［C］// Ninth International Conference on Natural Computation. May 19，2014，Shenyang，China. IEEE，2014：173-177.

[5] Heitz G, Chechik G . Object separation in X-Ray image sets[C]// The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010. IEEE, 2010.

[6] Xiao Y, Wang S , Wang Z ,et al.Coordinated-security based on probabilistic shaping and encryption in MMW-RoF system.[J].Optics letters, 2023, 48 11:, 2989-2992. DOI:10.1364/ol.493644.

[7] Mery D Automated detection in complex objects using a tracking algorithm in multiple X-ray views[C]//Computer Vision & Pattern Recognition Workshops.IEEE, 2011.DOI:10.1109/CVPRW.2011.5981715.

[8] Hassan T, Akcay S , Bennamoun M, et al. Trainable Structure Tensors for Autonomous Baggage Threat Detection Under Extreme Occlusion[C]//2020.DOI:10.1007/978-3-030-69544-6_16.

[9] AKCAY S，KUNDEGORSKI M E，DEVEREUXM，et al. Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery［C］// 2016 IEEE International Conference on Image Processing，Sep. 25-28，Phoenix，AZ，USA. IEEE，2016：1057-1061.

[10] Gao Qiang, Pan Jun, Hong Ruifeng. Research on automatic identification of dangerous goods in airport security check based on CNN[J]. Computer Technology and Development, 2019, 29(10):95-99. (in Chinese)

[11] Yang J, Zhao Z , Zhang H ,et al.Data Augmentation for X-Ray Prohibited Item Images Using Generative Adversarial Networks[J].IEEE Access, 2019, PP(99):1-1.DOI:10.1109/ACCESS.2019.2902121.

[12] Zhang Youkang, SU Zhigang, ZHANG Haigang, et al. X-Ray security images multiscale contraband detection[J]. Signal Processing, 2020, 36(7):11. (in Chinese)

[13] Berg A C, Fu C Y , Szegedy C ,et al. SSD: Single Shot MultiBox Detector. 2015[2024-02-28]. DOI:10.1007/978-3-319-46448-0_2.

[14] Miao C, Su C, Wan F, et al. SIXray: A Large-scale Security Inspection X-ray Benchmark for Prohibited Item Discovery in Overlapping Images.2019[2024-02-28].

[15] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. IEEE, 2016. DOI:10.1109/CVPR.2016.90.

[16] YOU Xi, HOU Jin, REN Dongsheng, et al. Adaptive Security Check Prohibited Items Detection Method with Fused Spatial Attention[J]. Computer Engineering and Applications, 2023, 59(21): 176-186.(in Chinese).

[17] SU Xingwang, Wang Xiaoming, Huang Jinbo, et al. X-ray security contraband detection based on deformable convolution and attention mechanism[J]. Electronic Measurement Technology, 2023, 46(10): 98-108. (in Chinese)

[18] LI Song, Yasenjiang Musa. Improved YOLOv7 X-Ray Image Real-Time Detection of Prohibited Items[J]. Computer Engineering and Applications, 2023, 59(12): 193-200. (in Chinese)

[19] SABOUR S，FROSST N，HINTON G E. Dynamic routing between capsules［C］// 31th Neural Information Processing Systems，Dec. 4-9, 2017, Long beach, CA, USA. NIPS Proceeding, 2017:3856-3866.

[20] Hinton, G.E., Sabour, S., Frosst, N.: Matrix capsules with EM routing. In: International Conference on Learning Representations (2018)

[21] Gao Xin, Yu Jiahao, Zha Sen, Fu Shiyuan, Xue Bing, Ye Ping, Huang Zijian, Zhang Guangyao. An ensemble-based outlier detection method for clustered and local outliers with differential potential spread loss[J]. Knowledge-Based Systems,2022,258.

[22] Szegedy C, Liu W Jia Y Q et al Going deeper with convolutions C // 2015 IEEE Conference on Computer Vision and Pattern Recognition CVPR June7-12 2015 Boston MA USA New York IEEE 2015 15523970

[23] SELVARAJU R R，COGSWELL M，DAS A，et al. Grad-CAM：Visual explanations from deep net works via gradient-based localization［C］// International Conference on Computer Vision，Oct.22-29，2017，Venice，Italy. IEEE，2017：618-626.

[24] HUANG G，LIU Z，LAURENS V，et al. Densely connected convolutional networks［C］// 2017 IEEE Conference on Computer Vision and Pattern Recognition, Jul. 21-26, 2017, Honolulu, HI, USA.IEEE, 2016:2261-2269.

[25] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 770–778.

[26] SZEGEDY C，VANHOUCKE V，IOFFE S，et al. Rethinking the inception architecture for computer vision［C］// 2016 IEEE Conference on Computer Vision and Pattern Recognition，Jun. 27-30，2016，Las Vegas，NV，USA.IEEE，2016：2818-2826.

[27] Howard A, Sandler M , Chu G ,et al.Searching for MobileNetV3[J]. 2019.DOI:10.48550/arXiv.1905.02244.

[28] Tan M, Le Q V .EfficientNetV2: Smaller Models and Faster Training[J]. 2021.DOI:10.48550/arXiv.2104.00298.

[29] Mery, Domingo, et al. "GDXray: The database of X-ray images for nondestructive testing." Journal of Nondestructive Evaluation 34 (2015): 1-12.