Security and Safety in the “Metaverse”

Other Fields

An accurate identification method for network devices based on spatial attention mechanism

Xiuting Wang1,2,3,* , Ruixiang Li2,3 , Shaoyong Du2,3,4 , and Xiangyang Luo2,3,*

1 Henan Polytechnic Institute, Nanyang 473000, China
2 Henan Province Key Laboratory of Cyberspace Situation Awareness, Zhengzhou 450001, China
3 State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China
4 Institute of Information Engineering, State Key Laboratory of Information Security, Beijing 100093, China

Received: 16 December 2022 / Revised: 23 February 2023 / Accepted: 22 March 2023 / Published online: 03 May 2023

Abstract With the metaverse being the development direction of the next generation Internet, the popularity of intelligent devices, and the maturity of various emerging technologies, more and more intelligent devices try to connect to the Internet, which poses a major threat to the management and security protection of network equipment. At present, the mainstream method of network equipment identification in the metaverse is to obtain the network traffic data generated in the process of device communication, extract the device features through analysis and processing, and identify the device based on a variety of learning algorithms. Such methods often require manual participation, and it is difficult to capture the small differences between similar devices, leading to identification errors. Therefore, we propose a deep learning device recognition method based on a spatial attention mechanism. Firstly, we extract the required feature fields from the acquired network traffic data. Then, we normalize the data and convert it into grayscale images. After that, we add a spatial attention mechanism to CNN and MLP respectively to increase the difference between similar network devices and further improve the recognition accuracy. Finally, we identify devices based on the deep learning model. A large number of experiments were carried out on 31 types of network devices such as web cameras, wireless routers, and smartwatches. The results show that the accuracy of the proposed recognition method based on the spatial attention mechanism is increased by 0.8% and 2.0%, respectively, compared with the recognition method based only on the deep learning model under the CNN and MLP models. The method proposed in this paper is significantly superior to the existing method of device-type recognition based only on a deep learning model.

Keywords Metaverse, Device identification, Deep learning, Spatial attention

Citation Wang XT, Li RX, Du SY and Luo XY. An accurate identification method for network devices based on spatial attention mechanism. Security and Safety 2023; 2: 2023002. https://doi.org/10.1051/sands/2023002

1 Introduction

The metaverse is a world made up of computers, and the concept of the metaverse has evolved since it first appeared, with a variety of descriptions. Typically, the metaverse is thought of as a virtual shared space that blends the physical, human, and digital worlds [1], and is where the next generation of the Internet is headed following the Web and mobile Internet revolution [2]. At present, with the popularity of intelligent devices, the metaverse has become a new development direction of the Internet.

* Corresponding authors (email: wangxiuxiu1997@163.com (Xiuting Wang); luoxyieu@sina.com (Xiangyang Luo))

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© The Author(s) 2023. Published by EDP Sciences and China Science Publishing & Media Ltd.
of intelligent devices and the maturity of various emerging technologies [3–5], more and more intelligent devices try to connect to the Internet so that computer users can experience another kind of life in the virtual world. There is no doubt that massive network device greatly facilitates people's daily production and life, but it also brings a variety of security problems. Through these network devices, people's life and privacy are easy to be obtained and monitored. In 2016, a large-scale DDoS attack took place in the United States, which caused a large area of Internet breakdown in the United States. Behind the attack was a botnet composed of about 1.5 million network devices [6]. The attack had a huge impact on the daily production and life of the American people. Therefore, to maintain network security, it is urgent to identify network devices on the Internet. Efficient and accurate identification of network devices is the basis for enhancing the security of cyberspace [7], realizing asset evaluation [8], and network situational awareness [9]. Therefore, in recent years, the research on network device identification technology has attracted wide attention at home and abroad.

At present, the network device identification method based on traffic is one of the hot research spots. In the process of identifying network devices, the traffic-based network device identification method mainly extracts the characteristic attributes of application layer protocol packets, network layer protocol packets, and transport layer protocol packets. However, both device identification methods based on network traffic packets and device identification information can identify network devices, but in the actual identification process, the data may be subject to malicious changes and forgery, and it is easy to cause identification errors when the characteristics are processed manually, which will reduce the identification accuracy. Therefore, it is of great academic value and practical significance to study fast and accurate network device identification methods.

By extracting application layer protocol features, Gao et al. [10] construct the fingerprint of a fixed device and realizes the identification of the network device. Luo et al. [11] propose to generate fine-grained fingerprints based on the subtle differences between the hardware information of the device, use natural language processing technology to process the hardware information of the device and verify the effectiveness of the proposed method through the system. Genevieve et al. [12] propose to apply deep learning to network traffic to automatically identify network devices connected to the network. Wang et al. [13] propose a network device identification method based on MAC boundary inference, which identifies device types by inferring MAC address boundaries corresponding to device types. In reference [14], clock offset information is used as device identification, and device fingerprint is generated by extracting the timestamp difference between the detection packet and the response packet, so as to realize the identification of network devices. The feature attributes extracted by the existing device identification based on network traffic often have redundant features and interference features, which increase the time cost of identification and reduce the identification accuracy.

In view of the above problems, by analyzing the relationship between network traffic and deep learning, this paper studies the network device identification technology based on deep learning, combines the characteristics of image for data enhancement, and the grayscale image of different device types has great differences, and uses conversion rules to convert network traffic data into a grayscale image for deep learning processing. By adding the spatial attention mechanism to make the network device identification model has a certain interpretability and increase the small differences between devices, a deep learning network device identification method based on the spatial attention mechanism is designed to improve the accuracy of network device identification.

The main work of this paper is as follows:

- This paper proposes a gray image conversion technology, which splits and reorganizes the traffic data of network devices based on sessions, and generates gray images corresponding to different device types as the input of the deep learning model. Compared with the existing methods, the proposed method can reduce the error of device identification.
- This paper proposes an optimal deep learning network device identification technology based on spatial attention mechanism, and uses the image processing method of deep learning to process the network gray image. By adding a spatial attention mechanism, the identification accuracy of network devices is improved. Compared with the existing methods, this method can effectively improve the identification accuracy of the device.
- The experimental results on public data sets show that compared with the typical identification methods based only on convolutional neural networks, the accuracy of the proposed convolutional
neural network device identification method based on spatial attention mechanism is improved by 0.8%. The proposed multi-layer perceptron network device identification method based on the spatial attention mechanism has an improved accuracy of 2.0% compared with the identification method based only on the multi-layer perceptron model.

The rest of this paper is organized as follows: Section 2 introduces the work related to network device identification, Section 3 describes the method in detail, Section 4 introduces the experiments and analyzes the results, and Section 5 summarizes the full text and looks into future work.

2 Related work

In recent years, network device identification has gradually become a research hotspot in the field of cyberspace security. At present, the mainstream identification method is to identify devices based on network traffic. Firstly, this method captures data packets [15] and firmware information [16] of network devices through detection tools. Then the captured traffic data is processed, the optimal features are selected, and interference features and redundant features are removed. Finally, various learning algorithms are used to identify the network device.

Among them, Greis et al. [17] construct the fingerprint of a specific network device by extracting the application layer protocol features and using the deep learning model to realize the identification of network devices. Zhang et al. [6] propose to use natural language processing to extract web content, use machine learning to build a classification model, and use network scanning technology to achieve real-time, non-invasive network crawling.

Deep learning is widely used in the field of network device identification due to its autonomous learning ability of features. Umair et al. [18] propose to apply deep learning to analyze network traffic to automatically identify network devices connected to the network. Zhu et al. [19] propose an efficient classification method for network devices based on multi-level deep learning. In this method, deep neural networks are used to extract traffic features and maximum entropy classifiers are used to classify Internet traffic. Although the machine learning traffic classification system based on a shallow neural network proposed in this paper has achieved very good classification effect and achieved a very high identification accuracy, this method does not carry out further fine-grained identification of network devices. Meidan et al. [20] design a traffic monitoring system based on the C5.0 decision tree and time series analysis. The system uses a CNN-LSTM (Convolutional Neural Networks, Long short-term Memory) combination model to autonomously learn the traffic, to avoid human intervention features. However, this system has poor applicability. Kotak et al. [21] propose a method based on machine learning to analyze the traffic of network devices, and then identify the network device. Firstly, researchers monitor and obtain TCP packets of devices. Then, the feature extraction tool is used to convert each TCP packet of the device into a feature vector and construct the feature space. Finally, an optimal classifier is constructed for each device type, and the device is identified by the machine learning algorithm. The proposed method achieves high identification accuracy. However, the single use of one protocol packet cannot satisfy the existing identification of various types of network devices.

Shiv et al. [22] train ten deep learning models for ten types of network devices to detect the traffic generated by network devices and non-network devices. In each model, one type of deep learning model is regarded as a positive class and the rest as a negative class. The deep learning model is trained separately for each network device. Finally, the effectiveness of this method is proved through identification accuracy. Although this method can improve the identification accuracy by training a deep learning model for each network device, it still has some problems. Firstly, each deep learning model needs to train all samples in the training process. As the amount of sample data increases, the time cost of model training will also increase. Secondly, because in this mode, the number of positive and negative samples is very unbalanced, with the increase of sample data, this asymmetry will be gradually serious. Thirdly, in practical application, if new network device types are added, all ten deep learning models need to be retrained.

According to the experimental results of the existing methods, the existing identification methods all need to manually process the features extracted from the data and have better identification performance for the network devices with a large degree of differentiation. However, manual feature selection and processing often difficult to capture the small differences between similar devices, which will lead
Figure 1. Framework of network device identification method based on spatial attention

3 Methods

This section describes network device identification techniques based on spatial attention mechanisms. By analyzing the relationship between network traffic and deep learning, this method studies the network device identification technology based on deep learning. Combining the characteristics of gray images of different device types with great differences, this method uses conversion rules to convert network traffic data into grayscale images convenient for deep learning processing. By adding the spatial attention mechanism, the network device identification model has a certain interpretability and increases the small differences between devices, improving the accuracy of network device identification.

3.1 Method framework

In response to the problems raised in Section 2, this section proposes a method for identifying network device types based on spatial attention mechanisms. Firstly, the traffic data of the network device is split and reorganized based on the session. Then, grayscale images corresponding to different device types are generated as the input of the deep learning model. Secondly, the spatial attention mechanism training optimal classification model is added to the convolutional neural network and the multi-layer perceptron respectively. Finally, the network devices are classified based on the optimal classification model. Through the deep learning model, the autonomous learning ability of the model is improved, to avoid the identification error caused by manual processing. The addition of an attention mechanism makes the deep learning model of the network device identification with certain interpretability, further improving the identification accuracy. The following mainly introduces the basic principle framework and main steps of the network device identification method based on spatial attention mechanisms. The basic framework is shown in Figure 1. The framework mainly includes four parts: data preprocessing, grayscale image generation, optimal classification model training, and network device classification identification.

The specific workflow of the network device identification method based on the spatial attention mechanism is as follows:
1. Data splitting: Since there are many protocols in a pcap (packet capture) file, this paper only needs TCP protocol packets, so the existing original pcap file is parsed to extract the TCP traffic data.

2. Data reorganization: The TCP traffic packets are reorganized according to the session, and finally constitute a complete TCP session.

3. Data preprocessing: Normalizes the reassembled TCP traffic data.

4. Data filling: Since convolutional neural network and multi-layer perceptron are to be used for classification, and the input requirements of deep learning model are consistent, the length of traffic data must be fixed, that is, the pixel size of the grayscale image must be fixed. Most of the effective information in communication packets is concentrated in the header. Therefore, this paper selects 784 bytes to ensure identification accuracy. If the length of bytes is larger than 784, it will be truncated, and if the length is smaller than 784 bytes, it will be processed by 0 filling.

5. Generating gray image: The reconstructed traffic data is converted into a 28*28 pixel gray image, which is the input of a convolutional neural network and multi-layer perceptron, so as to train the optimal deep learning model. In this paper, the generated gray image is divided into two parts. One part is used as the training data to construct the optimal deep learning model, and the other part is used as the validation data set to verify the learning effect of the deep learning model.

6. Construct the deep learning model: the gray image generated in the previous step is used as the input of the convolutional neural network and the multi-layer perceptron model, and the spatial attention mechanism is added to the two deep learning models. Through the autonomous learning of the deep learning model, the weights and parameters are adjusted iteratively, and the deep learning model is evaluated according to the loss function until it is trained into the optimal identification model.

7. Network device identification: The trained optimal deep learning model is used to classify the traffic of all network devices in the data set, and the type of network devices is determined by classification.

8. Performance evaluation: After classification, this paper evaluates the performance of the deep learning model, and evaluates the effectiveness of network device classification based on the spatial attention mechanism of convolutional neural network and multi-layer perceptron model by calculating the accuracy rate, recall rate, and F1 value.

3.2 Generating grayscale images

The existing network device identification methods mainly analyze the traffic generated by the network device in the communication process and identify the device by processing the features extracted from the traffic data. In the process of extracting the traffic feature field, there will be two problems. One is that the artificial processing of the feature is easy to appear information loss, resulting in the identification error. Second, for the data set with a small sample size, the model learning is not sufficient and the identification error will occur. On the one hand, the grayscale image can ensure that there will be no information loss. on the other hand, the data set can be expanded by means of data augmentation such as translation, rotation, and cropping. Therefore, this chapter converts the traffic data of network devices into the gray image to expand the data set, reduce the error of device identification and improve the identification accuracy.

Since the premise of generating the grayscale image is data division and data reorganization, this section introduces data preprocessing, data division, and data reorganization, then introduces the generation of a grayscale image, and finally displays the grayscale image of some network device generated.

The original network traffic file obtained by this method contains a variety of protocol information, only part of the traffic data can be used for identification, so it is necessary to analyze the header information of the network traffic file pcap and extract the required protocol through the identification field. In this experiment, this paper extracts the TCP protocol for network device identification.

The traffic packet is divided according to the session, then an original pcap file can be divided into several small pcap files. In this method, the traffic data after re-partitioning is normalized. Because the convolutional neural network and multi-layer perceptron are used for classification, the length of the input traffic data must be fixed, that is, the size of the gray image must be the same. However, most of the traffic characteristic data required by this method are distributed near the front. Therefore, 784 bytes are fixed in this paper. If the traffic byte is larger than 784, it will be truncated. If it is smaller than 784 bytes, it will perform the 0-filling operation at the end of the data. The specific grayscale image generation algorithm is shown in Algorithm 1.
Algorithm 1 Grayscale image generation algorithm

**Input:** Network device traffic list $L_t$, Length of flow $flow\_len$, Flow length after cutting $cut\_len$;

**Output:** List of grayscale images of network devices $P_t$;

1: $cut\_len = 784$;
2: $png = 28*28$;
3: **for each** $flow\_len \in L_t$ **do**
4:   **if** $flow\_len > 784$ **then**
5:     $cut\_len = flow\_len$ ;
6:   **else**
7:     add 0x00 until $flow\_len = cut\_len$ ;
8: **end if**
9: **end for**
10: **for each** $flow\_len \in L_t$ **do**
11:     transform $flow\_len$ to $png$;
12:     $png\_width = 28$;
13:     $png\_height = 28$;
14:     insert $png$ into $P_t$;
15: **end for**

Figure 2. Grayscale image of part network device. (a) D-LinkCam, (b) D-LinkDoorSensor, (c) D-LinkSwitch, (d) D-LinkWaterSensor, (e) Lightify, (f) WeMoSwitch.

According to the traffic data generated by different network devices, the restructured traffic data is converted into a 28*28 pixel gray picture, marked with the device label, as the input of convolutional neural network and multi-layer perceptron. The grayscale images of some of the network devices generated at last are shown in Figure 2, where 2a represents the grayscale image of the Youxun camera, 2b represents the grayscale image of the door sensor, 2c represents the grayscale image of the switch, 2d represents the grayscale image of the water sensor, 2e represents the grayscale image of the lamp, and 2f represents the grayscale image of the WeMo switch.

### 3.3 Construct an optimal deep learning model

By inputting grayscale images of network devices generated in Section 3.2, an optimal deep learning network device identification model based on a spatial attention mechanism is constructed. In this paper,
a relatively simple deep-learning model is adopted. The convolutional neural network CNN is taken as an example to train the optimal network device identification deep learning model.

In the training of the deep learning model, this paper converts the pre-processed traffic feature data into a gray image, which is taken as the input of the model. Part of the gray image is used as the training set data to train the optimal identification model, and the results of each layer are calculated through forward propagation until the output, which is used as the prediction result. If the predicted results meet the expectation and the training times are enough, it shows that the model is the optimal deep learning model for network device identification. If the predicted result does not meet the expectation, the parameters will be adjusted through backpropagation, and the process will be repeated until the optimal deep learning model is found and the network device type is finally output.

Forward propagation refers to calculating and storing the results of each layer of the deep learning model from input to output [23]. Its calculation formula can be expressed as

$$\alpha^k = \sigma (s^k) = \sigma (\alpha^k \ast W^k + b^k)$$  \hspace{1cm} (1)

where $k$ stands for the number of layers, $W$ for weight, $b$ for bias, and $\sigma$ for activation function.

In this chapter, two layers of convolution are used. The first layer of convolution uses 16 convolution kernels, the step size is 1, the activation function is the Relu function, and the convolution process can be expressed as

$$s(i,j) = (X \ast W)(i,j) + b = \sum_{k=1}^{n_{in}} (X_k \ast W_k)(i,j) + b$$  \hspace{1cm} (2)

where $s(i,j)$ represents the corresponding position of the convolution kernel $W$ in the final output matrix, $n_{in}$ represents the number of input matrices, $X_k$ represents the $k$th matrix, and $W_k$ represents the $k$th sub-convolution kernel matrix.

When the spatial attention mechanism is added to the process, the contribution of different features to the identification device is different in the network device identification based on the traffic packet. Similarly, not all areas in the image are equally important in terms of their contribution to identifying devices. Therefore, by adding the spatial attention mechanism, the image features can be independently learned to find out the parts of the gray image that contribute more to the identification device for enhancement processing, give them weight, weaken the features that contribute less to the identification, to improve the identification accuracy. A simple convolutional neural network model based on the spatial attention mechanism is shown in Figure 3.

The Relu activation function is used in the above convolution process, and its formula can be expressed as

$$r(\alpha) = max(0, x)$$  \hspace{1cm} (3)

This method adopts maximum pooling, the pooling window is set to $[2, 2]$, the step size is set to 1, and its formula can be expressed as 3.4. The Pooling layer can be located in multiple convolution layers and used to compress the image. It can compress images and reduce the dimension of features, leaving only the most important features for network device identification. The most important thing is that the pooling operation can prevent overfitting, which is more conducive to the optimization of deep learning model. Dropout is also used to prevent overfitting, setting this parameter to 0.25.

$$MaxPooling = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$$  \hspace{1cm} (4)
Figure 4. CNN model parameters based on spatial attention mechanism
The second layer of convolution adopts 36 convolution cores with a step size of 1, which also uses the Relu activation function and maximum pooling and has the same parameters as the first layer.

In this method, parameters are adjusted and corrected through the cross entropy loss function. The larger the cross-entropy loss is, the larger the gap between the two outputs is. Otherwise, it indicates that the two outputs are closer. Its formula can be expressed as:

$$ L(Y|f(x)) = -\sum_{i=1}^{N} Y_i \log f(x_i) \quad (5) $$

where $Y$ represents the true value and $f(x)$ represents the predicted value. The specific CNN model parameters and process based on the spatial attention mechanism are shown in Figure 4.

The parameter setting and process of CNN based on the spatial attention mechanism are shown above. This model is also applied to the multi-layer perceptron MLP model. Assuming that the vector $X$ represents the input layer of the multi-layer perceptron, then the hidden layer can be represented as $f(W_1X + b_1)$, where $w_1$ is the weight, which can also be called the connection coefficient. $b_1$ represents the bias and the function $f$ is the activation function. The output layer can then be represented as $\text{softmax}(W_2X_1 + b_2)$, where the $X_1$ represents the hidden layer’s output $f(W_1X + b_1)$. Combined with the above explanation, the multi-layer perceptron model can be represented as

$$ f(x) = \text{softmax}\left( b^{(2)} + W^{(2)} \left( b^{(1)} + W^{(1)}x \right) \right) \quad (6) $$

This model converts network traffic data into grayscale images, and then takes the grayscale images as the input of the deep learning model. The image processing method of deep learning is used to process the grayscale images on the network. By adding a spatial attention mechanism, the identification accuracy of network devices is improved.

4 Experiment and result analysis

In order to evaluate the performance of the network device identification method proposed in this paper, experimental verification of real data is carried out in this section. The deep learning network device identification method proposed in this paper based on spatial attention mechanism is compared with the existing network device identification method based only on CNN. The data set in reference [24] was used in the experiment, and the identification accuracy was compared with that of reference [22], which only used CNN for network device identification.

4.1 Experimental setup

In this experiment, the original network traffic data pcap file captured in the paper [24] is adopted. The network traffic data set contains the traffic generated by 31 kinds of smart home network devices during communication, including monitoring cameras, switches, smartwatches and other smart home network devices. Each type of network device was set repeatedly for at least 20 times, and the original network traffic data pcap file was obtained. Based on these original traffic data files, the new traffic data was generated by splitting and reassembling. The specific device types of this experimental data set are shown in Table 1.

4.2 Optimal deep learning model training

This section makes a comparative analysis of the optimal deep learning model constructed in Section 3.3. Take the convolutional neural network as an example. The network device identification model-based on the spatial attention mechanism adopts two layers of convolution. The first convolutional layer uses 16 convolutional cores, the step size is 1, and the activation function is the Relu function. The second layer of convolutional layer uses 36 convolutional cores with a step size of 1 and a Relu activation function. At the same time, maximum pooling, cross-entropy loss function, and Adam optimization algorithm are adopted.
### Table 1. Network device types

<table>
<thead>
<tr>
<th>Type of device</th>
<th>Type of device</th>
<th>Type of device</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Aria</td>
<td>11 EdimaxPlug1101W</td>
<td>22 SmarterCoffee</td>
</tr>
<tr>
<td>1 D-LinkCam</td>
<td>12 EdimaxPlug2101W</td>
<td>23 TP-LinkPlugHS100</td>
</tr>
<tr>
<td>2 D-LinkDayCam</td>
<td>13 EdnetCam1</td>
<td>24 TP-LinkPlugHS110</td>
</tr>
<tr>
<td>3 D-LinkDoorSensor</td>
<td>14 EdnetCam2</td>
<td>25 WeMoInsightSwitch</td>
</tr>
<tr>
<td>4 D-LinkHomeCam</td>
<td>15 EdnetGateway1</td>
<td>26 WeMoInsightSwitch2</td>
</tr>
<tr>
<td>5 D-LinkSensor</td>
<td>16 HomeMaticPlug</td>
<td>27 WeMoLink</td>
</tr>
<tr>
<td>6 D-LinkSiren</td>
<td>17 HueBridge</td>
<td>28 WeMoSwitch</td>
</tr>
<tr>
<td>7 D-LinkSwitch</td>
<td>18 HueSwitch</td>
<td>29 WeMoSwitch2</td>
</tr>
<tr>
<td>8 DLinkWaterrSensor</td>
<td>19 iKettle2</td>
<td>30 Withings</td>
</tr>
<tr>
<td>9 EdimaxCam1</td>
<td>20 Lightify</td>
<td></td>
</tr>
<tr>
<td>10 EdimaxCam2</td>
<td>21 MAXGateway</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 5.** Identification accuracy and loss values based on CNN model

Figures 5 and 6 respectively show the changing process of identification accuracy and loss value under the convolutional neural network model only, and the changing process of identification accuracy and loss value under the convolutional neural network based on the spatial attention mechanism. Figure (a) shows the changing trend of identification accuracy of a network device in the training process, and Figure (b) shows the changing trend of loss rate in the training process. It can be seen from the figure that the convolutional neural network based on the spatial attention mechanism has less fluctuation and lower loss value.

In this experiment, the same comparison experiment was performed again under the multi-layer perceptron. Figures 7 and 8 respectively show the change process of identification accuracy and loss value under multi-layer perceptron only, and the change process of identification accuracy and loss value under multi-layer perceptron based on spatial attention mechanism.

Figures 7 and 8 respectively show the change process of identification accuracy and loss value under the multi-layer perceptron model only, and the change process of identification accuracy and loss value under the multi-layer perceptron model based on spatial attention mechanism. It can also be seen from the figure that the multi-layer perceptron model based on the spatial attention mechanism has less fluctuation and lower loss value.

#### 4.3 Identification accuracy experiment

This paper proposes a network device identification method based on a spatial attention mechanism, which uses a convolutional neural network and multi-layer perceptron model to autonomously learn features in network devices, and at the same time adds spatial attention mechanism to build an identification model.
Figure 6. Identification accuracy and loss value of CNN model based on spatial attention mechanism

Figure 7. Identification accuracy and loss values based on multi-layer perceptron MLP model

Figure 8. Identification accuracy and loss value of MLP model based on spatial attention mechanism
Kotak et al. [22] propose an automatic identification method for network devices based solely on convolutional neural networks. This experiment generates its own grayscale image data set by analyzing network traffic data files, compares the identification accuracy of this method with that of the reference [22] by calculating the confusion matrix, and further tests the two methods based on the multi-layer perceptron model to evaluate the effectiveness of this method. In this experiment, the numbers 0–30 are used in the confusion matrix to represent a total of 31 types of network devices from Aria to Withings.

It can be seen from Figures 9 and 10 that the accuracy of devices 3 and 14, namely D-LinkDoorSensor and EdnetCam2, is low. The accuracy of device identification in the convolutional neural network model only is 75% and 33%. In the convolutional neural network model based on the spatial attention mechanism, the accuracy rate is significantly improved, which is 83% and 67%. This indicates that the addition of a spatial attention mechanism can improve the identification accuracy for the data with a small sample size. However, the reason for the low identification accuracy of the device EdnetCam2 is that the data sample size of the network device is small, and the deep learning model cannot learn all the features, so the identification accuracy is low.

It can be seen from Figures 11 and 12 that devices 3, 9, 11, 12, 22, 25, 26, and 29 are less accurate. In this paper, a bar chart is used to show in detail the comparison of identification accuracy based on the multi-layer perceptron model before and after adding the spatial attention mechanism.

As can be seen from Figure 13, the identification accuracy of network devices is 75%, 71%, 88%, 88%, 50%, 97%, 88%, and 85% in the MLP model based only on multi-layer perceptron. In the multi-layer perceptron MLP model based on spatial attention mechanism, the accuracy rate is significantly improved, which is 83%, 100%, 100%, 100%, 100%, 100%, 100%, and 100%. It can be seen that the addition of a spatial attention mechanism can significantly improve the identification accuracy.

In order to evaluate the performance of the method in this chapter, three performance indexes including Accuracy, Recall, and F1-Score were still used in this experiment to evaluate the effectiveness of this
Figure 10. CNN confusion matrix based on spatial attention mechanism

Figure 11. Confusion matrix based on MLP
method in terms of identification accuracy. Table 2 shows the performance comparison of CNN and MLP before and after the addition of spatial attention mechanisms.
The above experimental results show that the network device identification method based on the spatial attention mechanism proposed in this paper has higher accuracy than the identification method based on deep learning alone. Through the evaluation of the public data set, the identification performance of the proposed method is better than that of the existing deep learning-based network device-type identification methods.

### 5 Conclusion

Existing traffic packet-based meta-universe network device identification methods often require manual participation, and it is difficult to capture the small differences between similar devices leading to identification errors. In this paper, a deep learning network device identification method based on a spatial attention mechanism is proposed. Firstly, the required feature fields are extracted from the acquired network traffic data. Then the data is normalized as the input of the deep learning algorithm, and converted into a gray image. Then, the spatial attention mechanism is added into the convolutional neural network and the multi-layer perceptron respectively to increase the differences between similar network devices, to improve the feature autonomous learning ability of the model and further improve the identification accuracy. Finally, network devices are identified based on the deep learning model. A large number of experiments were carried out on 31 types of network devices such as web cameras, wireless routers, and smartwatches. The results show that the accuracy of the proposed CNN identification method based on the spatial attention mechanism is increased by 0.8% compared with the typical identification method based on CNN only. The proposed MLP network device identification method based on the spatial attention mechanism has an improved accuracy of 2.0% compared with the identification method based only on the MLP model. In future work, this paper will focus on the large-scale identification method of network devices in the metaverse and enhance the applicability, to achieve accurate identification of network devices and improve the security of network devices in the meta-universe.

**Conflict of Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Data Availability**

The original data are available from corresponding authors upon reasonable request.

**Authors’ Contributions**

Methodology, Wang XT; validation, Li RX and Du SY; formal analysis, Wang XT, Li RX and Du SY; investigation, Wang XT, Li RX and Du SY; data curation, Wang XT, Li RX and Du SY; writing—original draft preparation, Wang XT; writing—review and editing, Luo XY; visualization, Wang XT; supervision, Luo XY and Du SY; funding acquisition, Luo XY. All authors have read and agreed to the published version of the manuscript.

**Acknowledgements**

We would like to thank Pei Zhang, Guo Wei and others for helping us check the details and providing us with valuable suggestions in this paper.

**Funding**

This work was supported by the National Key Research and Development Program of China (No. 2022YFB3102900), the National Natural Science Foundation of China (No. U1804263, 62172435 and 62002386) and the Zhongyuan Science and Technology Innovation Leading Talent Project, China (No. 214200510019).

---

**Table 2.** Compares the performance of the deep learning model before and after adding the spatial attention mechanism.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on CNN [22]</td>
<td>98.7%</td>
<td>97.1%</td>
<td>97.9%</td>
</tr>
<tr>
<td>Based on CNN+ Spatial attention</td>
<td>99.5%</td>
<td>98.2%</td>
<td>98.3%</td>
</tr>
<tr>
<td>Based on MLP</td>
<td>97.6%</td>
<td>93.2%</td>
<td>94.3%</td>
</tr>
<tr>
<td>Based on MLP+ Spatial attention</td>
<td>99.0%</td>
<td>98.9%</td>
<td>98.8%</td>
</tr>
</tbody>
</table>
References


Xiuting Wang received the B.S. degree from Zhengzhou University, Zhengzhou, China, in 2019. She received her M.S. degrees from the State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China, in 2022. She is currently teaching at Henan Polytechnic Institute. Her major research direction is network device identification and data analysis.

Ruixiang Li received the B.S. and M.S. degrees from the State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China, in 2016 and 2019 respectively. He is currently pursuing his PhD in the state Key Laboratory of Mathematical Engineering and Advanced Computing. His research interests include IP geolocation, device recognition and data analysis.
Shaoyong Du received the B.E. degree in software engineering from Zhengzhou University, in 2012, and the Ph.D. degree in computer science and technology from Nanjing University, in 2019. His current research focuses on security and privacy in mobile computing.

Xiangyang Luo received the B.S., M.S., and Ph.D. degrees from the State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China, in 2001, 2004, and 2010, respectively. He is currently a professor at Zhengzhou Science and Technology Institute and the State Key Laboratory of Mathematical Engineering and Advanced Computing. His research interests lie in multimedia security and cyberspace surveying and mapping.