Business Hall Abnormal Behavior Detection Based on Cascading Deep Neural Network Model

He Yang*, Peng Wu† and Guoqiang Lin‡
Artificial Intelligence on Electric Power System State Grid Corporation Joint Laboratory (GEIRI), Global Energy Interconnection Research Institute, Beijing 102209, China

*Corresponding author e-mail: yh_fc@163.com, †wup@geiri.sgcc.com.cn, ‡linshen1213@163.com

Abstract. In the business halls of our country, video cameras are very common, but many videos are analyzed manually, which is time-consuming and labor-intensive, not smart enough. Many of them are used after-the-fact, unable to meet real-time requirements, and the time is relatively tight. However, our method can analyze the video, get real-time analysis results, and perform early warnings. Neuron network is a common method of artificial intelligence for identifying objects in images and videos. This method requires a large amount of sample data to obtain higher accuracy. This paper proposes a behavior detection framework based on cascading depth neural network model. Based on the advanced drfcn neural network, we trained two models and the two models are cascaded. Based on this framework, we implemented a drfcn-service system, which implements the features of drfcn, and flexibly adds plugins to support single model, two-model cascading and even multiple model cascading. Experiments show that the cascading system can significantly improve the accuracy of recognition.

1. Introduction
There are about 21,000 power business halls nationwide, in which we basically realizes the video surveillance coverage based on the security requirements of the power business hall scene. However, the current processing of video is still based on the on-site staff on duty in the business hall, which is time-consuming, labor-intensive and not smart enough. For example, for disputes during service, we have to collect evidence afterwards, and the incident response is relatively passive.

Common monitoring contents include irregularities such as staff member failing to arrive on time, not wearing work clothes, sleeping during work, playing mobile phones during working hours, smoking during working hours, and non-staff illegally entering work areas and fighting. Because the manual detection is time-consuming and labor-intensive, the target recognition is not timely, and the alarm response cannot be timely. To solve this problem, this paper proposes a cascading model based on the deep convolutional neural network for the detection of abnormal behaviors in the business hall scene (especially smoking during working hours). The following paper is organized as follows: Chapter 2 introduces the framework design of the cascading model, and Chapter 3 implements the drfcn-service system based on the framework design of Chapter 2, and the fourth chapter shows the comparison of experimental results. Chapter 5 draws conclusions and summarizes future research directions.
2. Cascading model framework

2.1. Cascading model overview

The rfcn [1] model discards the fully connected layer and uses the fully convolutional layer to locate and identify the target of interest in the image. Compared with other classical object detection models such as Faster RCNN [2], the computational complexity is significantly improved. In addition, rfcn proposes position-sensitive score maps to solve the position sensitivity problem of target detection, which further improves the detection accuracy.

Based on the rfcn model, drfcn incorporates the deformable convolution and deformable RoI pooling [3] modules. When the two modules process the object, they add offsets on the convolution layer and the pooling layer respectively, so that drfcn has better adaptability to geometric changes than rfcn, thus getting better results on visual tasks. At the same time, both modules are learning the offsets from target tasks, without additional supervision, and can be trained directly through standard back-propagation. We can use these two modules to replace the counterparts in the rfcn, thus improving the detection effect.

Based on drfcn, we further propose a cascading neural network model for the detection of abnormal behaviors, especially smoking during working hours, in business hall scenes.

![Cascading model and ordinary model](image)

**Figure 1.** Cascading model and ordinary model

The traditional image recognition method uses the image of what should be identified for training, without considering the integrity, and the monitoring of the business hall scene should be based on human detection. Therefore, we not only recognize people but also other objects in the behavior recognition task. For example, in the case of recognition of smoking behavior, if the model predicts that it is another object rather than a person, then the subject does not smoke. Further, if a person is detected, considering the position of the smoking may be outside the box of person, in addition to the size of the
box, we enlarge the width of the box appropriately, and then send the content of the enlarged box to the smoking recognition model to identify whether someone smokes. The specific frame is shown in Fig. 1.

As can be seen from Fig., the cascading system we designed requires support for both single-model and multi-model cascading.

2.2. Cascading model details
In order to enable the implemented system to support the single-model image analysis, double model cascading and even multiple model cascading, we designed the predictor and plugin components.

Predictor is the actual deep neural network predictor, predictor needs certain data formats, the final location of image box needs a certain format, and the return message needs a certain format. So these are some repeated processing, we design plugins to perform these processes, completely separated from the predictor. We can configure whether the plugin exists or set the plugin parameters, the plugin can be flexibly placed before or after the predictor as needed. In addition to converting the data format, the plugin has other functions, such as completing the cascading operation.

Taking smoking during working hours as an example, the cascading process is performed by the object recognition model and the smoking model. First, the client sends an identification request to the server, and the server pre-processes the image through the plugin, and then passes it to the predictor of the object recognition model, and then the post-plugin perform the coordinate conversion of the model output box. If the object is recognized as a category other than one person, the next step is not processed. If it is recognized as person, the cascading plugin (callNext) will be executed, and the plugin will transfer the ROI of the human body to the next model, and the height and width of the ROI are adjusted as what is configured, and a message is sent to the next smoking model. The data format of the message is the same as that of the image before being preprocessed by the plugin. The next model first preprocesses the image and then passes it to the smoking model. After the prediction coordinate conversion of the output box of model is executed, and then plugin generates the converted coordinate message and sends it to the previous object recognition model. After the object recognition model receives the message, it generates the final message through the plugin and return the final result to the client. For the specific system implementation, see Chapter 3.

3. Drfcn-service system implementation
In order to complete the requirements of the second chapter framework, we designed a drfcn-service system, in which the predictor part is implemented based on the open source MXNET framework, and both the plugins and the message delivery are completely designed and written by our team. We mainly deliver the message through Redis, for some data that can not be sent directly we use protobuf to generate a message. The system is developed with python and is divided into client and server, respectively; both parts have corresponding py files. The server py files are used to start the service, so that the service is stored in graphics memory, and the client py files send a message to Redis and accept the message returned by the server to obtain the result of the image recognition.

A large number of configuration parameters of the server are configured through the yaml configuration file, and the yaml configuration file is automatically read when the server starts. In addition, the predictor needs a neural network model file, that is, a params file, and the params file is a model file generated through training.

The specific framework of the service is shown in Fig. 2. At first in one configuration file which pipeline (can be understood as a complete process of a model) should be loaded and the name of the predictor, the type of neural network used, the serial number of the used gpu, the address of the log, some common settings of the service, the access of Redis, and some other parameters are configured. Drfcn service then loads the additional pipeline and predictor configuration files according to the specified path in the previous configuration file. Pipeline is a process of each processing, including pre-plugin, predictor and post-plugin. There are many plugins, for example, im_transform is used for image pre-processing, abs2rel converts the coordinates of the model output box, rel2response generates the protobuf message from the processed box, and callNext is used to pass the result of the previous model
processing to the next model. The cascading model mentioned in this article use this very plugin. This plugin will generate a prototype message by converting the content of the picture frame, and then send the message to Redis (similar to the request submitted by client) to complete the cascading of the model.

![Drfcn-service system structure](image)

**Figure 2. Drfcn-service system structure**

4. Cascading model experiment
For the video surveillance needs of business halls, taking smoking as an example, we collected 539 smoking pictures, which is divided into two categories, background and smoking, with 500 pictures as training data and 39 pictures as test data. At the same time, in order to further expand the training data set, we extracted the images of people in the Pascal VOC 2012 data set for training.

For the test data, the test process has two forms as shown in Fig. 1. First, the smoking model and the object recognition model are all trained, and then we need no too complicated operations. Instead, we can simply configure the configuration file to complete the two formats of Fig. 1. The configuration of the two processes are as follows.

1. The first form tests image by cascading object recognition model and smoking model, Add the pre-plugin im_transform in the configuration file of the object recognition model, after that add the predictor, then add the post-plugin abs2rel, and then fill in the cascading plugin callNext, You need to configure the name of the next model in the parameters of callNext. Here we can reuse the configuration file of the second form without modification, directly fill in the name of the smoking model pipeline, moreover we should also configure the enlargement ratio of height and width of the box that the callNext plugin passes to the next model, and finally we fill in the post-plugin rel2response. The object recognition model is executed in order, at first we execute im_transform, then the processed data is passed to the predictor of the object recognition model, and then execute abs2rel, then if it is identified as other categories, there is no further processing, if it is identified as a person, We will execute callNext, the ROI of the human body is scaled and passed to the smoking model. The smoking model first executes im_transform, then passes to the predictor of the smoke model, then executes abs2rel and rel2response, returns the result to the object recognition model, and the object recognition model executes rel2response to return the final result.

2. The second form directly tests images through the smoking model, Add the pre-plugin im_transform, predictor, ie the smoking model, the post-plugin abs2rel and rel2response in order. When executed, the pipeline first executes im_transform and then passes it to the predictor of the smoke model, then execute abs2rel and rel2response.

The test data was tested by both a cascading model and just a smoking model, the cascading model contained an object recognition model and a smoking model, and the smoking model in the cascading model is identical to the independent smoking model.

Run the models separately, get their results csv file, and then compare the result with ground truth, calculate their mAP (mean average precision), the results are shown in Table 1. It can be seen that the cascading mode outperforms the independent smoking model in most cases and on average.
Table 1. Comparison of picture recognition results

|          | Cascading model | Smoking model only |
|----------|----------------|-------------------|
| mAP@0.60 | 0.600058       | 0.590597          |
| mAP@0.65 | 0.578546       | 0.481841          |
| mAP@0.70 | 0.311077       | 0.234337          |
| mAP@0.75 | 0.167706       | 0.147790          |
| mAP@0.80 | 0.047921       | 0.061735          |
| mAP@0.85 | 0.029986       | 0.001000          |
| mAP@0.90 | 0.001136       | 0.000000          |
| Average mAP | 0.248061 | 0.216757          |

5. Conclusion

In this paper, a cascading model framework is proposed for the abnormal behavior detection requirements in the current business hall scene. The framework requires support for both ordinary models and two cascading models or even multiple cascading models. We have implemented drfcn service system according to the framework design. The service system cleverly implements the requirements of the cascading model framework through plugins.

Taking smoking detection as an example, the cascading model first detects the subject as a person and further identifies whether a person is smoking. The experimental results show that the cascading model can improve the recognition accuracy by 3% compared with the single smoking recognition model, and proves the validity of the proposed model.

The implementation of the cascading model proposed in this paper has been considered as an effective method, but it inevitably has some places to continue to improve, specifically the following.

1. The current model is relatively simple, only can identify smoking behavior, and we will increase the detection of abnormal behaviors such as fighting, playing mobile phones, not wearing work clothes, sleeping during work, and not arriving on time;

2. Currently, each frame of the video is processed independently, without considering the context. Future research will incorporate the correlation between frames to further improve the detection accuracy of abnormal behavior.

3. At present, some real-time video processing may not meet the speed requirement, and may require frame skipping. For example, a video is 24 frames per second, we take one frame every two frames, and then synthesize the video, and the video becomes 12 frames per second, further study is needed to accelerate the model.

Acknowledgments

The work is supported by State Grid Corporation of China Headquarters Project (SGGR0000JSJS1701223) which is 'Application Research of New Generation Artificial Intelligence Technology in Grid Customer Service'.

References

[1] Dai J , Li Y , He K , et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks [J]. 2016.
[2] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. 2015.
[3] Zhu X , Hu H , Lin S , et al. Deformable ConvNets v2: More Deformable, Better Results [J]. 2018.
[4] Girshick R B, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [J]. computer vision and pattern recognition, 2014: 580-587.
[5] Girshick R . Fast R-CNN [J]. Computer Science, 2015.
[6] Redmon J, Divvala S K, Girshick R B, et al. You Only Look Once: Unified, Real-Time Object Detection [J]. computer vision and pattern recognition, 2016: 779-788.
[7] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector [J]. european conference
on computer vision, 2016: 21-37.

[8] Chen L, Papandreou G, Kokkinos I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs [J]. international conference on learning representations, 2015.

[9] He K, Gkioxari G, Dollar P, et al. Mask R-CNN [J]. international conference on computer vision, 2017: 2980-2988.

[10] Chen L, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40 (4): 834-848.