Abnormal Behavior Recognition Based on Transfer Learning

Li ZhanLi1,a Ye JiaWei2,b
1Xi’an University of Science and Technology
Xi’an University of Science and Technology, Xi’an, Shaanxi, 710054
13384914816, 86
2Xi’an University of Science and Technology
Xi’an University of Science and Technology, Xi’an, Shaanxi, 710054
18682919049, 86
a lizl@xust.edu.cn b 843470772@qq

ABSTRACT: In order to identify abnormal behavior in image samples, a transfer learning recognition model based on Inception-V3 neural network is proposed in this paper. The whole process consists of two parts: 1) Inception-V3 neural network is used to extract the features of image samples. 2) The abnormal behavior recognition model is established to classify the obtained features. In this paper, three-layer feedforward neural network algorithm is used to compare with other common machine learning classification models (k-nearest neighbor algorithm, random forest, LightGBM, SVM, two-layer feedforward neural network). The results show that the performance of anomalous behavior recognition based on three-layer feedforward neural network is better than other classification models.

1. INTRODUCTION

Abnormal behavior detection is the core of intelligent monitoring system. In recent years, it has attracted wide attention from academia and industry, and has become an important research topic of computer vision. However, the complexity of scenes and the diversity of abnormal behavior make abnormal behavior detection still a challenging task. According to the feature used, abnormal behavior detection methods can be divided into two categories: abnormal detection based on visual features and abnormal detection based on trajectories. In the aspect of abnormal detection based on object trajectories, Junejo et al. [1] proposed abnormal behavior detection based on DBN with the feature of trajectory position, velocity and space-time curvature. Yang et al. [2] proposed a local abnormal detection method based on trajectory and multi-instance learning. Li et al. [3] used cubic splines curves to represent target trajectories and abnormal behavior detection based on trajectory sparse is proposed. Mo et al. [4] proposes anomaly behavior detection method based on joint sparse representation of trajectories. Kang et al. [5] uses HMM to represent the trajectory motion pattern, and proposes an anomaly detection method based on HMM. However, in the case of target occlusion, the tracking performance decreases significantly, leading to anomaly detection.

At present, deep learning feature has been successfully used in abnormal behavior detection. Xu et al. [7] uses stacked denoising coder to extract appearance features and motion features and uses multi-class SVM to detect abnormal behavior. Erfani et al. [8] uses Deep Belief Networks to extract behavior features, and uses SVM to detect abnormal behavior. Zhou et al. [9] uses space-time convolution neural
network to extract behavior features, and carries out human behavior detection. Fang et al.\cite{10} used saliency information of image and multi-scale optical flow histogram as low-level features, and then used PCANet to extract more effective features from these low-level features for anomaly event detection. Hu et al.\cite{11} constructed a deep-enhanced slow feature analysis network, which was used for behavior feature extraction and anomaly detection. Abnormal behavior detection is to learn complex non-linear behavior features from the original image by means of depth structure and multi-layer non-linear transformation. Compared with the abnormal detection based on artificial features, the features acquired by deep learning often have certain semantic features and stronger discriminative ability, which can more effectively represent the behavior characteristics. Therefore, abnormal behavior based on deep learning is more effective. At the same time, abnormal behavior detection based on deep learning reduces the computational complexity. However, the performance of abnormal behavior detection based on deep learning falls far short of people's expectations. The main reasons are as follows: 1) deep learning method needs a large number of samples for training, while the general behavior database samples are relatively small; 2) in order to balance the computational cost, deep learning behavior representation usually adopts down-sampling strategy, which results in information loss\cite{12}. In this paper, anomalous behavior recognition in NBA court is studied, which enables computer to automatically extract features through Inception-V3\cite{13-14} transfer learning, and reduces the steps of redesigning neural network and training neural network.

In order to identify abnormal behavior in NBA field, this paper uses three-layer feedforward neural network based on transfer learning characteristics to establish abnormal behavior recognition model. The experimental results show that the detection result of the algorithm is the best, and it has a high reference value for the detection of abnormal behavior in sports venues.

2. FEATURE EXTRACTION

2.1 Convolutional Neural Network

Convolutional neural network is a multi-layer supervised learning neural network. Each layer consists of several two-dimensional planes, and each plane contains many independent neurons. Typical convolution neural networks consist of convolution layer, pool layer and fully connected layer three types of two-dimensional plane layers. The convolution layer enhances the characteristics of the original signal by convolution operation; the pooling layer is also called the down-sampling layer, which uses the principle of image local correlation to downsample the input image, while reducing the amount of data processing, retains the feature information; the output layer achieves the effect of classifier by integrating the feature information with category discrimination.

(1) In the convolution layer, the input image or the feature map of the upper layer is convoluted and biased with the convolution filter of the layer, and the feature map of the convolution layer is output through a non-linear activation function. The specific calculation is as follows:

\[
z = \sum_{i \in I} k_i \ast x_{i-1} + b_i \tag{1}
\]

\[
x_i = f(z) \tag{2}
\]

Formula: \(x_{i+1}\) is the upper output of the feature map; \(z\) is the convolution operation of the feature map output. \(x_i\) is the final feature map output of the convolution layer obtained by the non-linear activation function; \(k\) for the convolution kernel; \(b\) for bias; \("\ast\) for the convolution calculation (step size 1), \(f(\cdot)\) for the non-linear activation function. In this algorithm, ReLU function is chosen as the non-linear activation function. Generally used non-linear activation functions such as sigmoid and tanh, because their gradients in positive and negative saturation regions are close to 0, will have vanishing gradient, while the partial gradient of ReLU function greater than 0 is constant, so the problem of vanishing gradient is avoided. At the same time, ReLU function is used as the non-linear activation function. Because only the input value is positive and the output value is zero, the network has a moderate sparsity.
For the convolution layer, it improves the network’s data representation ability and speeds up the convergence of the network training process.

(2) In the pooling layer, the output characteristic map of the convolution layer is sampled down to reduce the dimension of the data. Maximum pooling uses the maximum value in the pooling region as the down-sampling output, avoiding the weakening of the larger elements in the region by zero elements.

$$x_i = \max(x_{i-1})$$  \hspace{1cm} (3)

Formula $x_{i-1}$ a pooled window corresponding to the output feature map of the previous layer; the maximum pooled output corresponding to the pooled window is shown in the formula $x_{i-1}$.

(3) The output layer of traditional CNN uses the full connection layer as the output classification result of the network output layer. In this algorithm, to use fine-tuned full-connection layer as the output layer, and the pool layer is used to output a feature map of N number (corresponding to N categories of samples respectively), and then the softmax is used to normalize it. Softmax regression model is a generalization of logistic regression model on multi-classification problems. For a given input $x$, the hypothetical function of Softmax is defined as follows:

$$\begin{bmatrix}
z_1 \\
z_2 \\
\vdots \\
z_k \\
\end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{w_j^Tx + b_j}} \begin{bmatrix}
e^{w_1^Tx + b_1} \\
e^{w_2^Tx + b_2} \\
\vdots \\
e^{w_k^Tx + b_k} \\
\end{bmatrix}$$  \hspace{1cm} (4)

(4) Formula: $z_k$ is the output of the k neuron; $w^T$ and b are weighted and biased respectively. It can be seen from formula (4) that the Softmax regression model constructs the output of neurons into probability distribution and plays a role of normalization. In this algorithm, the loss function is defined as cross-entropy loss function.

$$C = -\sum_k y_k \log z_k$$  \hspace{1cm} (5)

In the formula: $y_k$ is the real value corresponding to the k class, and the value is 0 or 1.

2.2 Transfer Learning
Transfer learning is to apply a model trained on a problem to a new problem through simple adjustment. Convolution neural networks contain a large number of parameters to be trained. At the beginning of training, these parameters are usually initialized randomly, which makes the initial error of the network in a relatively large position, and easily leads to poor convergence and over-fitting of the network. To solve this problem, a supervised pre-training method of transfer learning based on feature selection is proposed. The purpose is to obtain common feature representations in source domain and target domain, and then to realize knowledge transfer based on these feature representations. Transfer learning is a machine learning method that uses existing knowledge to solve different but related domains. It relaxes two basic assumptions in traditional machine learning: 1) training samples and new testing samples for learning should satisfy the conditions of independent and identical distribution; 2) sufficient training samples must be available to obtain a good model. Transfer learning includes Source Domain and Target Domain, which are defined as follows:

$$D(s) = \{x, P(x)\}$$  \hspace{1cm} (6)

$$D(t) = \{x, P(x)\}$$  \hspace{1cm} (7)

In the formula, D(s) and D(t) are the source domain and the target domain respectively; x and P(x) are the marginal probability distributions corresponding to the feature space and the feature space in one domain respectively. Source task D(s) in source domain is defined as image recognition, and target task D(t) in target domain is defined as Abnormal Behavior recognition. The network parameters in the
target task are initialized by the pre-training model acquired by the source task, and the feature information transfers from the source domain to the target domain.

2.3 Inception-V3
Christian et al. published GoogleNet (Inception-V3), which further optimized the structure of GoogleNet under the original framework, balanced the width and depth of network model, and improved network performance by maximizing network information flow by balancing the number and depth of filters at each stage; replaced large convolution with small convolution, such as replacing $5 \times 5$ convolution with two $3 \times 3$ convolution cores to reduce the number of parameters, and the Inception module is constructed by using the convolution of $1 \times n$ and $n \times 1$ instead of $n \times n$ convolution. The purpose of each Inception module group is to simplify the spatial structure and transform the spatial information into high-order abstract feature information, as shown in Figure 1.

![Inception module](image1.png)

The model structure consists of five convolution layers, three pooling layers and 11 mixing layers as shown in Figure 2:

![Model structure](image2.png)

2.4 Feature Extraction
In this paper, the 1-19 layers of GoogleNet (Inception V3) model are adopted, which are mainly composed of input layer, convolution layer, Inception module and pooling layer. Inception module is widely used in the model. Its function is similar to convolution layer. It simplifies the spatial structure, transforms information into high-order abstract feature information map, and enriches the model expression ability. Overload inception-v3 parameters, and then input the original image, after training, the model extracts features, in the pool3 layer, the final output of a size of $1 \times 1 \times 2048$ feature vector, as the next stage of the feature. The schematic structure of the algorithm is shown in Figure 3.

![Model structure](image3.png)
3. **CLASSIFICATION AND EVALUATION**

3.1 *Feedforward Neural Network*

Multilayer feedforward neural network, also known as BP (Back Propagation) neural network, consists of input layer, hidden layer and output layer. The topology of a three-layer feedforward neural network is shown in Figure 4.

![Feedforward Neural Network Structure](image)

In this paper, two-layer feedforward neural network $2048 \times 2$ and three-layer neural network $2048 \times 1024 \times 2$ are adopted.

3.2 *Random Gradient Authors*

The batch gradient descent method is adopted to train the feedforward network. The pseudocode is as follows:

- Require: learning rate $\varepsilon_k$
- Require: initial parameter $\theta$
- While Stop criteria not met do:
  - Randomly sample $m$ samples from the training set $\{x^i, \cdots, x^m\}$, $x^i$ corresponding label is $y^i$
  - Computational gradient estimation: $g \leftarrow + \frac{1}{m} \nabla \sum_i L\left(f(x^i; \theta); y^i\right)$
  - apply update: $\theta \leftarrow \theta - \varepsilon g$
- Endwhile

Among them, the number of small batch samples $m$ is 100, the learning rate is 0.01, and the iteration training is 5000 times.

3.3 *Evaluating Indicator*

The three-layer feedforward network has the highest classification accuracy among the six models. In order to further evaluate the performance of the model, other evaluation indicators were used to evaluate the model. For the binary classification problem, the data set can be divided into four cases: true case (TP), false positive case (FP), true negative case (TN) and false negative case (FN), in which TP means
that the positive class is predicted to be positive, FN means that the positive class is predicted to be negative, FP means that the negative class is predicted to be positive, and TN means that the negative class is predicted to be negative. The performance indicators of the two classification problems include accuracy rate (P) and recall rate (R). It is not comprehensive to use these two indicators alone. In fact, the value of $F_1$ can reflect the overall performance of accuracy and recall. $F_1$ is defined as the harmonic average of accuracy and recall.

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$$  \hspace{1cm} (8)

$$F_1 = \frac{2TP}{2TP + FP + FN}$$  \hspace{1cm} (9)

In order to better compare the prediction performance of different models, ROC curve was used to evaluate the prediction performance. By comparing the area under ROC curve (AUC value), the accuracy, recall rate, $F_1$ value and AUC value of different models were calculated respectively.

3.4 Evaluating Indicator

In order to verify the effectiveness of the proposed method, to use 500 NBA of fighting, trip, hugging, shoot at the basket picture set, in which fighting is defined as abnormal behavior in this paper, and other normal behavior, using inception-v3 feature extraction based on the use of these methods for abnormal behavior recognition comparative experiments.

Evaluated by the accuracy, precision, recall, area under ROC curve AUC, $F_1$ indicators. Detailed results are shown as follows:

| model | acc  | pre  | rec  | $F_1$ | AUC  |
|-------|------|------|------|-------|------|
| lightgb | 0.813 | 0.785 | 0.578 | 0.666 | 0.917 |
| svm    | 0.813 | 0.821 | 0.575 | 0.676 | 0.816 |
| rf     | 0.855 | 0.642 | 0.720 | 0.679 | 0.782 |
| knn    | 0.796 | 0.821 | 0.547 | 0.657 | 0.816 |
| DFFN   | 0.916 | 0.855 | 0.688 | 0.762 | 0.909 |
| TFFN   | 0.941 | 0.873 | 0.759 | 0.812 | 0.929 |

4. CONCLUDING REMARKS

Based on Python language, in the integrated development environment of Anaconda, to use tensorflow reload the inception-v3 model to extract features, and build feedforward neural network model base the features. The batch random gradient descent method is used to train the features, and trained by batch random gradient descent method to automatically recognize the abnormal behavior of the NBA image. The results show that the three-layer feedforward neural network (TFFNN) has better performance than other classification models (k-nearest neighbor algorithm (knn), random forest (RF), LightGBM, SVM, two-layer feedforward neural networks (DFFNN)) in anomaly detection based on transfer learning.

REFERENCES

[1] JUNEJO I N. Using dynamic Bayesian network for scene modeling and anomaly detection [J]. Signal Image and Video Preessing, 2010,401):1 – 10.

[2] YANG W Q, GAO Y , CAO L B. TRASMIL: A local anomaly detection framework based on trajectory segmentation multi-instance learning [J]. Computer Vision and Image Understanding, 2013,117(10):1273 – 1286.

[3] LIC, HAN Z, YE Q M, et al. Visual abnormal behavior detection based on trajectory sparse reconstruction analysis [J] Neurocomputing, 2013,119 <6>:94 – 100.
[4] MO X, MONGAV BALR, et al. Adaptive sparse representations for video anomaly detection [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(4):631-645

[5] KANG K, LIU W B, XONG W W. Motion pattern study and analysis from video monitoring trajectory [J]. IEICE Transactions on Information and Systems, 2014, 97(6):1574 — 1578.

[6] SALIGRAMA V, CHEN Z. Video anomaly detection based on local statistical aggregates [C]// IEEE Computer Vision and Pattern Recognition. Providence: IEEE, 2012: 2112 — 2119.

[7] XU D, RICCI E. Learning deep representations of appearance and motion for anomalous event detection [C]// British Machine Vision Conference. Swansea: Spring, 2015: 1 — 12.

[8] ERFANI S M, RAJASEGARAR S, KARUNASEKERA S, et al. A linear one-class SVM with deep learning [J]. Pattern Recognition, 2016, 58(C): 121 — 131.

[9] ZHOU S F, SHEN W, ZENU D, et al. Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes [J]. Signal Processing - Image Communication, 2016, 47: 358 — 368.

[10] FANG Z, FEI F, FANG Y. Abnormal event detection in crowded scenes based on deep learning [J]. Multimedia Tools and Applications, 2016, 75(22): 14617 — 14639.

[11] HU X, HU S Q, HUANU Y P, et al. Video anomaly detection using deep incremental slow feature analysis network [J]. IET Computer Vision, 2016, 10(4): 258 — 267.

[12] ZHU F, SHAG L, XIE J, et al. From handcrafted to learned representations for human action recognition: A survey [J]. Image and Vision Computing, 2016, 55(1): 42 — 52.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1 — 9, 2015.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. arXiv preprint arXiv:1512.00567, 2015.