Review on CNN-based violent detection method

Guanjie Liang

Queen's University Belfast

gliang02@qub.ac.uk

Abstract. Violent behaviors seriously endanger people's life and property safety, and also undermine social stability and development. In order to monitor the occurrence of violent behavior, the monitoring system plays a vitally important role. As the surveillance is used widely in daily life and video data is growing rapidly, it is increasingly unrealistic to manually detect violent behavior in surveillance, because it will consume excessive manpower. Therefore, it is necessary to establish an automated system for detecting violent behavior. And the convolutional neural network (CNN) plays a leading role in automatic detection. This paper has done a lot of research to study and understand the development of CNN-based violent behavior recognition. First, the original CNN is introduced and the related work is mentioned; Then, the two different improvement paths on CNN, 2DCNN+RNN and 3DCNN, are utilized to enhance the accuracy of violence detection or reduce the calculation. And finally, the discussion about advantages and disadvantages of these two improvement paths is shown in this paper and conclusion is presented. At present, the development of potential of CNN remains to be exploited and the development of CNN is expected.

Keywords: convolutional neural network(CNN), violent detection, violent behavior.

1. Introduction

Violence incidents occur in various places and at all times, which is not only causing harm to personal safety and property, but also seriously affecting the development and progress of the country. Meanwhile, with the development of technology, the monitoring system is widely applied in our daily life to detect the violence. Thus, the video data increases dramatically. It is an unrealistic approach to manually identify the presence of violence in such a large video data and it will wastes too much manpower. If there is a system that automatically detects violent behavior, then all the problems will be solved.

The recognition methods based on handcrafted features and end-to-end deep learning model are two mainly methods for recognizing human action in video [1]. Nam et al. [2] proposed using flame and blood to recognize the violent scenes in videos. Y CHEN and L ZHANG [3] proposed a violent detection based on optical flow context histogram. Y GAO et al. [4] proposed a novel method, namely Oriented VIolent Flows (OViF), to extract features. And this method identifies violence by comparing the changes in histograms computed in the direction of optical flow. A new feature descriptor which is applied to record the changing information of the optical flow was proposed by Mahmoodi J et al. [5]. The above methods are all the first recognition method, have limited generalization ability and some features rely on optical flow information. In this paper, I study the application process of convolutional neural networks in violent behavior detection.

In the following sections, the introduction of the basis of Convolutional neural networks is in section 2. Next, section 3 provides two different improvement methods to detect violence. Section 4 discuss the advantages and disadvantages of these two improvement paths. Ultimately, conclusion concerning the paper is in section 5.

2. Original Convolutional Neural Networks(CNN) for violence detection

As CNN deep learning algorithm develops rapidly, CNN is also gradually being used for violent behavior recognition.

2.1. The original CNN

As indicated in figure 1, the original CNN consists of five different layers and these five layers are divided into two parts. The feature extractor, composed of the first three layers, is to extract the features. The remaining two layers constitute the classifier to classify.

The first layer preprocesses the input information and then sends it to the next layer.

The core layer of the CNN is convolutional layer. It can extract the features and most computations are generated by it.

The operation of selecting the features passed by the convolutional layer is done in the pooling layer. In other words, the input feature maps are compressed. This operation not only decreases the amount of parameters and improves the computational accuracy, but also enhance the fault tolerance of the model.

The fully-connected layer is generally placed at the end of an entire CNN and acts as a classifier in CNN.

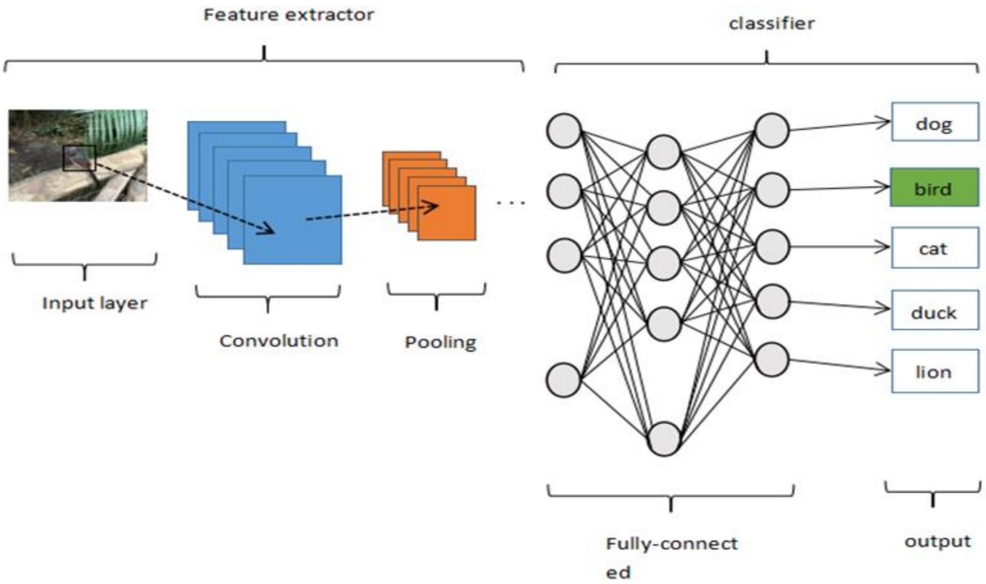

Figure 1. Schematic diagram of CNN.

2.2. The working principle of CNN

The working principle of convolutional neural networks is that the input layer receives the input data, and the features in images are extracted by the convolutional layer; the next layer is employed to reduce the parameter magnitude and prevent overfitting; finally, the fully-connected layer outputs the result.

2.3. Related work

In 1998, LeCun Y et al. [6] proposed LeNet-5, and combined the BP algorithm into the training to form the archetype of CNN. Krizhevsky A et al. [7] employed the CNN AlexNet to win the ImageNet image classification competition in 2012, with an accuracy rate higher than the second 11%. Since then, CNN have occupied a primary position in the field of computer vision.

3. Improvement

Violent features are manifested in both space and time. Thus, based on the premise that both temporal and spatial features need to be extracted, there are two different paths to achieve.

3.1. 2DCNN+RNN

The combination of 2DCNN and RNN is one of the paths to achieve extracting both temporal and spatial features, this combination applies 2DCNN to collect the features in space and then uses the Long Short-Term Memory (LSTM) to extracts the temporal features.

Srivastava N et al. [8] utilized the LSTM Encoder-Decoder framework for video representations learning. Donahue J et al. [9] proposed Long-term Recurrent Convolutional Networks (LRCNs), a new network structure combining CNN and LSTM, which can capture the relationship among the temporal states by training the video recognition models. Dong Z et al. [10] first applied 2DCNN to extract the spatial features on appearance flow, optical flow and acceleration flow, and then employed LSTM to process long-range temporal information for violence recognition. In order to decrease the computational complexity, Sudhakaran S et al. [11] applied the ConvLSTM to extract the spatial features from frame difference, because the ConvLSTM model is able to encode spatiotemporal information in its memory cell and frame difference is capable of suppressing background information. They also explained that training on frame difference can receive a better result. Hanson A et al. [12] realized that both next and previous inputs from a current state can enhance the recognition ability and propose a Bidirectional Convolutional LSTM (BiConvLSTM) architecture, built on ConvLSTM architecture, to detect violence in videos. Islam Z et al. [13] developed a useful method for detecting violence, based on Separable Convolutional LSTM (SepConvLSTM) which utilizes a depthwise separable convolution to replace convolution operation in ConvLSTM, to detect violence with less computation. In order to suppress the background information, they subtract the average frame from each frame. Patel M [14] proposed a pseudo real time violence detection system using the CNN+LSTM to tackle the violence detection challenge and tested different models of CNN to find that ResNet50 is better. Meanwhile, he found that retraining the CNN is capable of enhancing the capability of the network speedily and help the network to find relevant patterns of violence.

3.2. 3DCNN

3DCNN is another path to extract the spatialtemporal features. The extraction of spatialtemporal features are operated by 3D convolution in 3DCNN. As shown in figure 2, 3D convolution is the sliding window operation of the convolution kernel in the input three-dimensional space. In this structure, each feature map is connecting with the adjacent contiguous frames in last layer, which achieves capturing the motion information.

Ji S et al. [15] proposed a new 3D CNN model which is used for action recognition. The spatialtemporal features are extracted by 3D convolutions in this model. Compared with the original CNN, 3D CNN has more information channels from input frame, and the results of all channels is superimposed to gain the ultimate feature map. Ding C et al. [16] developed a new and sophisticated 3D ConvNets model, and first realized the extraction of spatiotemporal violence features based on 3DCNN. Tran D et al. [17] proposed a simple but effective approach named Convolutional 3D (C3D), which is a framework for 3D deep convolutional neural networks, to learn the spatiotemporal features. Zhi Zhou, Ming Zhu and Yahya Khan proposed a 3DCNN-based violent behavior detection method, which directly operates on the input through 3DCNN and can extract the spatiotemporal features of violent behavior well. Ullah F U M et al. [19] put forward a novel deep learning model for violence recognition, which is divided into three stages. At first, they applied MobileNet CNN to recognize the video frames including people to reduce the computation in the frames without people. Then, 3DCNN performs spatialtemporal feature extraction on the 16 frames containing people. Song W et al. [20] proposed a new scheme for detecting violence, based on an improved 3DCNN for improving the accuracy of 3D ConvNet detection. The scheme cuts the video sequence into clips by key frames to reduce redundancy and uniform sampling damage to motion integrity. Li J et al. [21] developed a novel model, based on 3D CNN. In order to capture the spatialtemporal features better and fewer parameters to calculate, this model improves the internal designs, which is improved by utilizing bottleneck units and the DenseNet architecture.

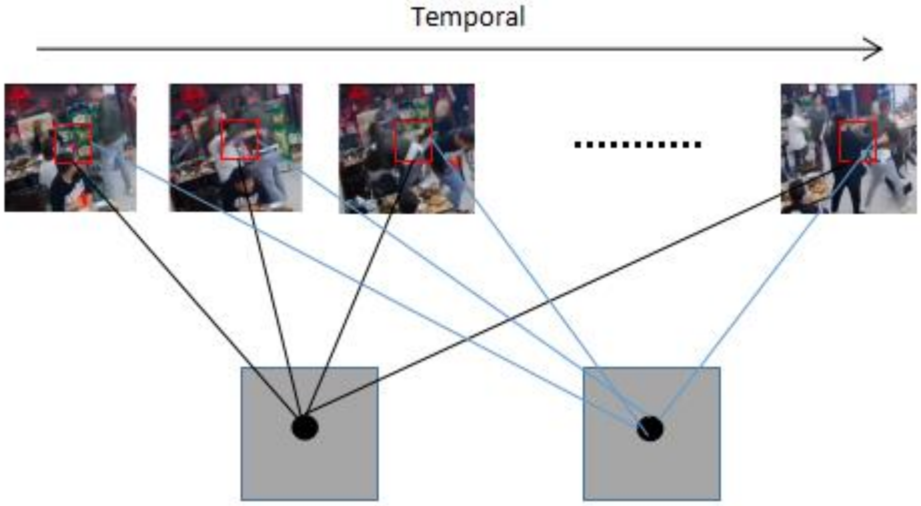

Figure 2. 3D convolution.

4. Discussion

Both paths are undoubtedly superior to the original CNN, but they still have shortcomings that we need to improve. What follows is a discussion of their pros and cons.

2DCNN+RNN architecture applies 2DCNN to extract the spatial features and employs RNN to extract the temporal features, which makes lots of information have temporal and spatial characteristics and achieves better results in violence recognition. However, the spatial and temporal features are extracted separably in 2DCNN+RNN architecture, which ignores the internal connection of spatiotemporal features so that it reduces the feature expressiveness. For RNN, training them requires high performance hardware. When the time span is large, the amount of calculation will be extremely large.

3DCNN takes the contiguous multiple frames as input, which increases the information of temporal dimension and make more expressive features extracted. 3DCNN utilizes 3D convolution to preserve temporal series information so that the results have both temporal and spatial information. However, 3DCNN has a limited convolutional receptive field in the temporal dimension, thus it is solely suitable for short-term time series. And the temporal features of distant frames are attenuated, because the propagation path is longer. Last, the complexity of 3DCNN is quite high, so the cost of research is also high.

5. Conclusion

With the upsurge of convolutional neural networks, more and more models are constructed on this basis. The role of CNN is essential in the development of computer vision, especially in recognition. This paper summarizes the development of violent behavior detection, focusing on the analysis of two paths based on CNN. The first one is 2DCNN+RNN, and the other one is 3DCNN. These two paths are both achieving great work in violent behavior recognition, but there are some shortcomings. In the future work, extracting the features of long-term activities is a worthy research direction under the premise of ensuring that the temporal features are not weakened. Meanwhile, a new outlook is made on the future model for extracting spatiotemporal capabilities.

References

[1] Cheng M , Cai K , Li M . RWF-2000: An Open Large Scale Video Database for Violence Detection[C]// International Conference on Pattern Recognition. IEEE Computer Society, 2021.

[2] Nam J , Alghoniemy M , Tewfik A H . Audio-visual content-based violent scene characterization[C]// Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269). IEEE, 2002.

[3] Chen Y , Zhang L . Fighting Detection Based on Optical Flow Context Histogram[C]// 2011 Second International Conference on Innovations in Bio-inspired Computing and Applications. IEEE, 2012.

[4] Gao Y , Liu H , Sun X , et al. Violence detection using Oriented VIolent Flows[J]. Image & Vision Computing, 2016, 48-49(APR.-MAY):37-41.

[5] Mahmoodi J , Salajeghe A . A classification method based on optical flow for violence detection[J]. Expert Systems with Applications, 2019, 127(AUG.):121-127.

[6] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[7] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[8] Srivastava N , Mansimov E , Salakhutdinov R . Unsupervised Learning of Video Representations using LSTMs[J]. JMLR.org, 2015.

[9] Donahue J , Hendricks L A , Guadarrama S , et al. Long-term recurrent convolutional networks for visual recognition and description[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.

[10] Dong Z , Qin J , Wang Y . Multi-stream Deep Networks for Person to Person Violence Detection in Videos[J]. Springer Singapore, 2016.

[11] Sudhakaran S , Lanz O . Learning to Detect Violent Videos using Convolutional Long Short-Term Memory[C]// IEEE. IEEE, 2017:1-6.

[12] Hanson A , Pnvr K , Krishnagopal S , et al. Bidirectional Convolutional LSTM for the Detection of Violence in Videos[J]. Springer, Cham, 2018.

[13] Islam Z , Rukonuzzaman M , Ahmed R , et al. Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM[C]// 2021.

[14] Patel M . Real-Time Violence Detection Using CNN-LSTM[J]. 2021.

[15] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 221-231.

[16] Ding C , Fan S , Ming Z , et al. Violence Detection in Video by Using 3D Convolutional Neural Networks[C]// International Symposium on Visual Computing. Springer, Cham, 2014.

[17] Tran D , Bourdev L , Fergus R , et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]// IEEE International Conference on Computer Vision. IEEE, 2015.

[18] Zhi Z, Ming Z, Yahya Khan. Violence Behavior Detection Based on 3D-CNN [J]. Computer Systems & Applications, 2017, 26(12):5.

[19] Ullah F U M, Ullah A, Muhammad K, et al. Violence detection using spatiotemporal features with 3D convolutional neural network[J]. Sensors, 2019, 19(11): 2472.

[20] Song W, Zhang D, Zhao X, et al. A novel violent video detection scheme based on modified 3D convolutional neural networks[J]. IEEE Access, 2019, 7: 39172-39179.

[21] Li J , Jiang X , Sun T , et al. Efficient Violence Detection Using 3D Convolutional Neural Networks[C]// 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019.