A Novel Transfer Learning-Based Algorithm for Detecting Violence Images

Yuyan Meng[1], Deyu Yuan[1,2,*], Shaofan Su[1] and Yang Ming[1]
[1] Department of Police Information Engineering and Cyber Security, People's Public Security University of China, Beijing 100038, China
[e-mail: mengx61033@qq.com]
[2] Key Laboratory of Safety Precautions and Risk Assessment, Ministry of Public Security, Beijing 102623, China
[e-mail: yuandeyu@ppsuc.edu.cn]
[*]Corresponding author: Deyu Yuan

Received March 21, 2022; revised May 3, 2022; accepted May 18, 2022; published June 30, 2022

Abstract

Violence in the Internet era poses a new challenge to the current counter-riot work, and according to research and analysis, most of the violent incidents occurring are related to the dissemination of violence images. The use of the popular deep learning neural network to automatically analyze the massive amount of images on the Internet has become one of the important tools in the current counter-violence work. This paper focuses on the use of transfer learning techniques and the introduction of an attention mechanism to the residual network (ResNet) model for the classification and identification of violence images. Firstly, the feature elements of the violence images are identified and a targeted dataset is constructed; secondly, due to the small number of positive samples of violence images, pre-training and attention mechanisms are introduced to suggest improvements to the traditional residual network; finally, the improved model is trained and tested on the constructed dedicated dataset. The research results show that the improved network model can quickly and accurately identify violence images with an average accuracy rate of 92.20%, thus effectively reducing the cost of manual identification and providing decision support for combating rebel organization activities.

Keywords: Deep learning, Image classification, Pre-training, Violence images

This work was supported by Fundamental Research Funds for the Central Universities, People's Public Security University of China (2021JKF215), Key Projects of the Technology Research Program of the Ministry of Public Security(2021JSZ09), the Fund for the training of top innovative talents to support master's degree program, People's Public Security University of China(2021yjsky018) and Key Project of National Social Science Foundation of China(No.20AZD114).

http://doi.org/10.3837/tiis.2022.06.003 **ISSN : 1976-7277**

1. Introduction

Currently, urban populations and infrastructures are in a phase of rapid development and various security problems have been as a result. For this issue, the riots have become the main threat worldwide. According to studies [1], at least one rebel attack occurs in more than 71 countries every year, causing an average of more than 33,000 deaths per year, as well as significant economic losses to countries and people. Counter-riot has become a major topic of study for scholars in most countries around the world. To improve understanding of rebel events, the US Study of Riot and Responses to Riot (START) has published an open-source global riot database with the acronym GTD to enable scholars to better study rebel events [2].

Studies have shown that there is an inevitable link between the seriousness of violent images and audio-visuals and their association with violent activities. However, with the increasing scale of the Internet, the conventional mode of work can hardly adapt to the current trend of Internet development. There are problems such as serious waste of human resources, low work efficiency, and inaccurate images that need to be solved.

Computer vision has become a popular topic of research for scholars in various countries in the past period. Since 2012, neural network models have emerged that achieve high accuracy on normal image datasets, such as AlexNet [3], GoogleNet [4], VGGNet [5], and ResNet [6]. These network models are trained on 1000 classes of 15million images that have been labeled. Therefore, the current image research is a hot topic for many scholars. Z. G. Qu et al. proposed a further study of the image[7-8].

Although deep learning techniques have been used extensively for image classification work and have achieved satisfactory results, the application of neural network models in the field of violent images content detection is still in its early stages. And experiments based on transfer learning and attention mechanisms are still rare because of the huge number of datasets that are traditionally essential for successful training of deep network models and the small number of image samples. The number of experiments based on transfer learning and attention mechanisms is even smaller. This experiment proposes to load the improved neural network model with pre-trained weight files for the detection of violent images. The residual network was chosen as the pre-trained network model due to its excellent performance in the ImageNet competition. A classification model suitable for violent image detection was constructed by fine-tuning the network structure on the constructed dataset. The experimental results show that the residual network with the introduction of transfer learning techniques and the SE attention mechanism is more effective than the original residual network in the self-constructed riot dataset.

The main work in this paper can be summarized as follows.

1) Building a dataset of riot-related data. A dataset of violent images containing obvious features of riot-related elements such as violent acts, bloody scenes, special slogans, etc., is crawled from the Internet and obtained from regulatory authorities.

2) A pre-trained weight file on ImageNet is introduced to the original residual network, while the network structure is fine-tuned to fit the self-constructed violent images dataset. Three different network structures with different layers, ResNet18, ResNet34, and ResNet50, were used as control groups to investigate whether transfer learning works significantly under different network layers.

3) Introducing a single-way attention mechanism SE module in the structural block of the residual network to obtain more accurate attention information by sinking cross-channel information through a one-dimensional convolutional layer. To explore whether the

introduction of the SE module in the shallow neural networks ResNet50 can significantly improve the effectiveness of the model.

4) Analyzing the experimental results, investigating the reasons for their occurrence, comparing the experimental data, and then drawing conclusions and proposing priorities and major tasks for further improvement.

The paper then proceeds to discuss the above in sections II, III, IV, and V in the form of related work, methodological study, experiments, and summary discussion.

2. Related work

The research on the classification of violent images has been more focused on the field of violent videos, and little work has been done specifically on the classification of violent images. L. Yan et al. used integrated learning to solve the problem of few samples and proposed an automatic labeling method for riot images, which effectively assisted in screening out riot information from web pages [9]. M. Chen used transfer learning for image classification [10]. Xinxu Hu et al. chose to improve the post-fine-tune network by adding additional layers to the network, using a new residual network-like structure, to eventually make use of more audio information [11]. Chao Huang et al. used a bag-of-words model for modeling and a support vector machine for classification, and after optimizing several global parameters[12]. S. Sun et al. started with the network side[13]. Meng Caixia et al. followed the filtering method to filter the noise in the fused images and used the calibrated approximate rectangular response technique to delineate the rioters' areas to achieve accurate detection of rioters [14]. A. Kaur and L. Kaur used SIFT and SURF techniques for the detection of self-built image datasets with hidden gun information and showed that the accuracy was up to 90% and SURF was faster than SIFT [15]. N. J. Hussein and F. Hu et al. used infrared sensors to capture real-time RBG and infrared images to detect weapons hidden in clothes and the results were more satisfactory [16].

In addition, the detection of violent behaviors is also a hot topic in the detection of images involving riots. Liu, Y., et al. used the Yolov3 feature pyramid model to detect people holding guns, waving drumsticks and violent behaviors in videos with an accuracy of 92.91%, and 80.5%, respectively [17]. The detection method deployed on IoT to find abnormal behavior in videos with an accuracy of up to 97% [18]. C. Dhiman et al. detected violent and other harmful behavior in videos by using algorithmic techniques with satisfactory results [19]. Z. Guo F. et al. detected violent behavior in videos by modifying the parameters [20]. S. Sudhakaran et al. used AleNet and also introduced the LSTM module to analyze the violence in videos, and finally achieved an accuracy of 94.57% in the Violence-Flows dataset [21]. P. Yun et al. additionally used 2 Deep CNN models as detection to effectively detect the phenomenon of crowding in videos, which was experimentally verified to be as high as 94.57% in the Crowd Violence dataset with an accuracy of 92.5% [22]. F.U.M. Ullah et al. used a pre-trained MobileNet model with reduced parameters and ended up with an accuracy of 94.6% on the hockey fight dataset [23].

S. Chaudhary et al. classified fighting, assault, and stealing as violent behaviors [24]. Further optimization of dataset quality was done by A.B. Mabrouk et al. [25]. E.Y. Fu et al. achieved detection of a small number of people with an accuracy of 90% by using Motion region and Optical flow methods [26]. M. Ahmed explored the state-of-the-art deep learning architecture of convolutional neural networks (CNNs) and inception V4 to detect and recognize violence using video data[27]. M. Al-Nawashi et al. built a video framework that can be used to automatically detect violent behaviors [28].

Most of the previous studies on the detection of violent image content have focused on violent acts, while few studies have used deep learning networks to target other riot-related features. In addition, improvements to the ResNet network are also the focus of many researchers[29-34]. Therefore, in this paper, we select special slogans, weapons, and violent acts as riot-related elements, build a special dataset of riot-related images, and design a deep learning neural network model based on transfer learning techniques and attention module for training, to achieve better detection of violent images.

3. Methodology

3.1 Overall architecture of brute force image classification of ResNet network based on transfer learning

Based on the image classification method of the traditional residual networks, this paper proposes ResNet50-se-pre, a violent image classification model that combines transfer learning technology with SENet, and the overall structure of the model is shown in **Fig. 1**.

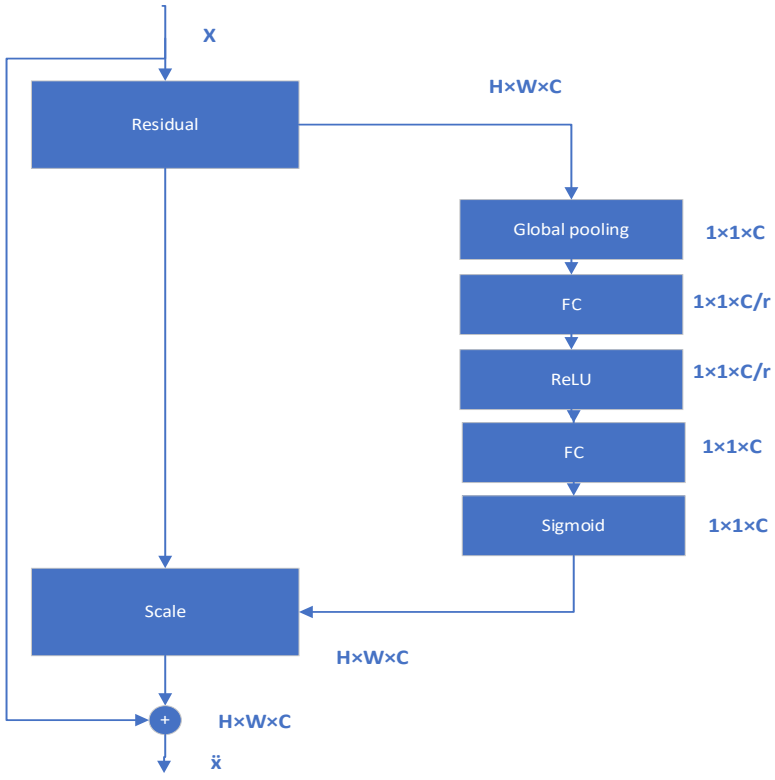


Fig. 1. The overall structure of the ResNet50-se-pre model

Resnet 50-SE-Pre algorithm mainly includes three steps:
(1) Pre-training: the pre-training model trained based on ImageNet data set is introduced;
(2) Residual block adjustment: SE attention mechanism was inserted into the original residual block to enhance feature extraction;
(3) Adjustment of output layer: According to the classification results, there are two results, namely, normal pictures and violent pictures, the output layer result is changed to 2.

3.2 Data preprocessing

As an important technology to enhance the size and quality of training datasets in deep learning, data enhancement technology can prevent the model from overfitting based on maintaining the original model structure and computational complexity. This paper uses the ability to adjust image hue saturation and standardize and normalize images to achieve the generalization capabilities of the model.

3.3 Residual network framework

The training of CNN deep models often requires the support of large-scale data sets. However, preparing a labeled dataset is a complex and urgent task. On the other hand, insufficient data sets not only do not make the model get better results, but also lead to overfitting of CNN models. To solve the problem of insufficient samples and better use the neural network model to solve the problem, the transfer learning technology, and attention module mechanism are introduced.

The residual network (ResNet) was proposed in 2015, mainly by introducing residual blocks, making a reference for the input of each layer, learning to form a residual function, so that the number of layers of the network is greatly deepened while reducing the impact of gradient dispersion. The common residual structure is shown in **Fig. 2**.

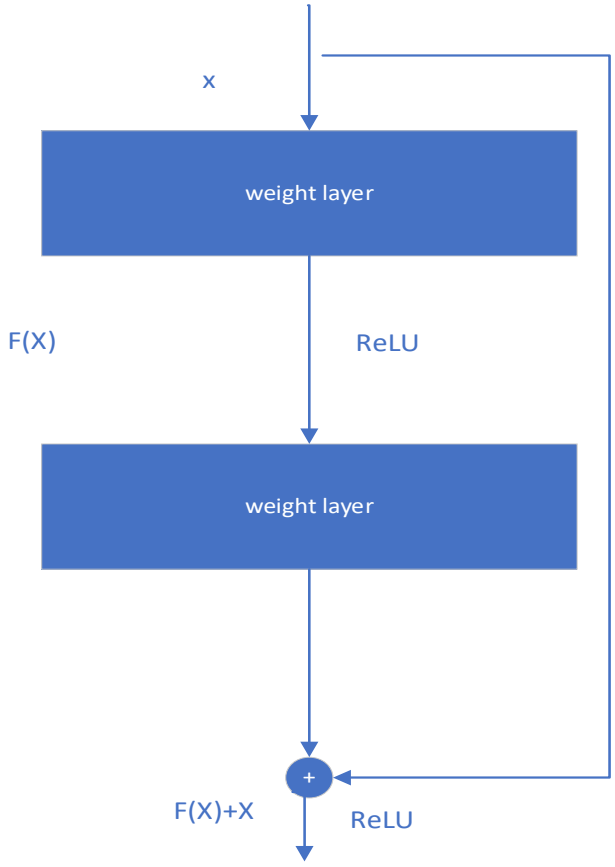


Fig. 2. Schematic diagram of common residual structures

By making a reference(X) to the input of each layer, learning to form a residual function, instead of learning a function without reference(X), such residual functions are easier to optimize and can greatly deepen the number of layers of the network. There are two layers in the residual block in the figure above, the corresponding expression is Eq. (1), which σ represents the nonlinear function ReLU, followed by a shortcut, and the second ReLU to obtain the output y, such as Eq. (2). When it is necessary to change the input and output dimensions (such as changing the number of channels), a linear transformation of x can be WS in the residual structure, as shown in Eq. (3).

$$F = W_2 \sigma(W_1 X) \tag{1}$$

$$Y = F(X, \{W_i\}) + X \tag{2}$$

$$Y = F(X, \{W_i\}) + W_S X \tag{3}$$

3.4 Transfer learning techniques

Transfer learning refers to a machine learning method that uses existing knowledge to solve problems in new areas that differ but still have some relevance to the sample or problem [29]. In transfer learning, the learning model needs to learn related tasks on the source data distribution (source domain). And then transfer the knowledge to specific tasks (target tasks) on the target distribution (target domain), To improve the performance of the model on specific tasks. Here are two important concepts: "domain" and "task"

(1) Field

Domain D can be expressed as the form of a binary group $\boldsymbol{D} = \{\boldsymbol{X}, \boldsymbol{P}(\boldsymbol{x})\}$, where X represents the feature space, $\boldsymbol{x} \in \boldsymbol{X}$, $\boldsymbol{P}(\boldsymbol{x})$ is the marginal probability density function.

(2) Tasks

For a given domain D, the task can also be expressed in the form of a binary group $\boldsymbol{T} = \{\boldsymbol{Y}, \boldsymbol{f}(\cdot)\}$, where Y is the label space and $\boldsymbol{x} \in \boldsymbol{f}(\cdot)$ is the prediction function of the model on domain D. Based on the above two concepts, the definition of transfer learning is given: given source domain D_S and source task T_S, target domain D_T and target task T_T. Transfer learning aims to help the model solve the prediction function $\boldsymbol{f}(\cdot)$ of target task T_T on target domain D_T through the knowledge obtained from source domain D_S and source task T_S when $\boldsymbol{D_T \neq D_S}$ or $\boldsymbol{T_T \neq T_S}$. Transfer learning technology is widely used in the field of image classification because it can reduce the demand for hardware resources, limit the sample size of the data set, effectively shorten the model training time, and enhance the generalization ability of the model. First of all, the ImageNet dataset is used to carry out pre-training, the result is a network weight parameter file, and then the pre-trained network structure is fine-tuned, that is, the weight parameters of the feature extraction layer are retained, and the final output of the model is adjusted from the original 1000 classification to the 2 classifications adapted to the self-built dataset, and the self-built fear-related image dataset is used to continue to train the model. And the adjusted network model structure can be obtained, as shown in **Fig. 3**.

Transfer learning takes the weight of the model trained in the source data domain as the initial weight of the target data set, modifies the output of the full connection layer as required, and retrains the network. Fine-tuning the network process can avoid the over-fitting phenomenon caused by small amount of data.

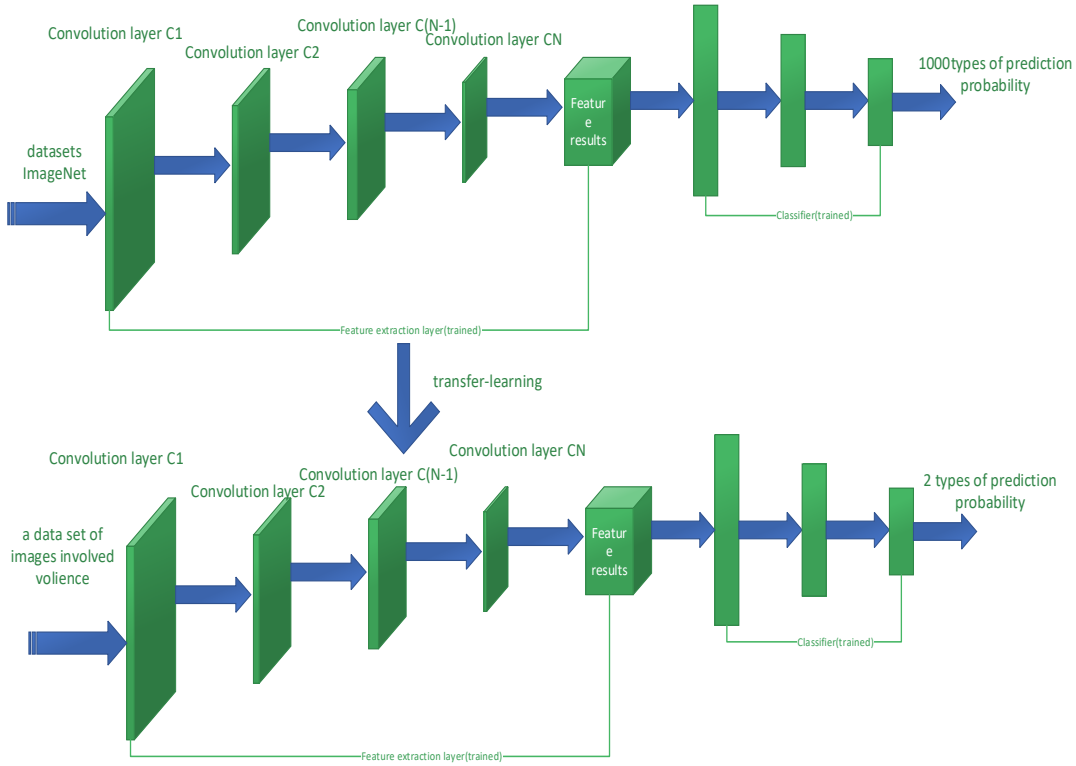


Fig. 3. Training process of transfer learning

3.5 SENet modules

In the routine feature extraction, it is often during the convolution operation that only the result of the convolution of each channel of the original feature map is performed for a simple accumulation operation, and the feature relationship between the channels is ignored. Therefore, how to use the correlation of the channel to better extract features is a major focus of the study. HU et al. innovatively proposed SENet (Squeeze-and-Excitation Network) in 2018 to take advantage of feature channel correlation. The SENet module consists of Squeeze, Excitation, and Reweight. The SE module structure is shown in **Fig. 4**.

SENet obtains feature weights according to the correlation between channels, strengthens important features and reduces the influence of invalid features, so as to train the model in this way to achieve better recognition effect. The basic idea of SE module is to strengthen the extraction of important features and weaken non-important features by controlling scale size, so as to avoid irrelevant feature infection and make feature extraction more targeted.

(1) Feature compression operation (Squeeze): First of all, the characteristic map of the input size H×W×C (H is the length of the input feature map, W is the width of the input feature map, and C is the number of feature channels), so that it changes to a 1×1×C feature vector with a global sensor field, to achieve the purpose of processing semantic information outside the region of each output unit in the bottom layer of the network. As shown in Eq. (4), where U is the feature map after convolution operations, W×H is the spatial dimension of u.

$$Z_C = F_{sq}(U_C) = \frac{1}{H \times W} \sum_{i=1}^{n} \sum_{j=1}^{W} U_C(i, j) \tag{4}$$

(2) Excitation operation: a fully connected layer of the function activated by using ReLU (whose input is the same dimension as the output feature channel) through two layers, and the weight of each feature channel is derived by using the Sigmoid activation function. Finally, the output feature channel weight vector is multiplied by the original input feature map by the Scale operation (Fscale), that is, the original feature calibration on the channel dimension is completed, and the final extracted feature has stronger directivity than the original feature, thereby improving the classification performance. As shown in Eq. (5).

$$S = F_{ex}(Z,W) = \sigma(g(Z,W)) = \sigma(W_2 \delta(W_1 Z)) \tag{5}$$

The SE module assigns a weight value to each channel through the fusion of full connection layer and multiplication features, and in the denoising task, each noise point is assigned a weight, the low weight noise point is automatically removed, the high weight noise point is retained, the network running time is improved, and the parameter calculation is reduced.

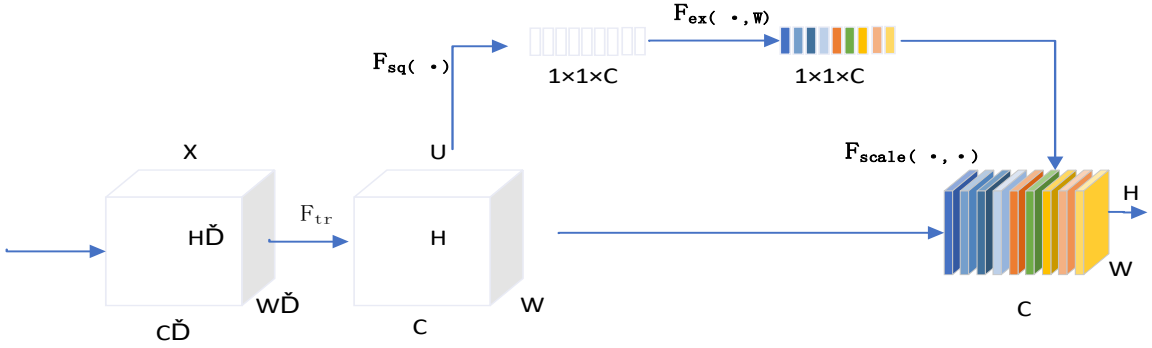


Fig. 4. SE module structure diagram

4. Experiments

This section describes the experimental details of the proposed methodology of the learned features representation model using transfer learning as discussed in section 3.

4.1 Experimental environment

The experimental platform for this article is Windows 10 Pro, version number 20H2, GPU is NVIDIA GeForce RTX 2080 Ti, and intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz 2 processors. This is done using the Pytorch framework in the anaconda 4.10.3 environment.

4.2 Dataset introduction

According to the preliminary survey, violent images are mostly related to beating, hacking and burning and blood. Therefore, the self-built data set selects a total of 4000 violent images (mostly bloody scenes and beating, hacking and burning) and 4000 normal images from the Internet and violent videos, and the positive and negative samples are distributed in a 1:1 ratio, totaling 8000 sample images. Firstly, the image should be clipped and the size should be normalized. At the same time, in order to increase the generalization and robustness of the experimental model, data augmentation methods such as inversion and Angle transformation were used to increase the diversity of data.

4.3 Experimental setup

This section will elaborate on the details of the experimental model of an improved network-based on transfer learning techniques discussed in section 3.

Considering the small sample size, shallow neural networks ResNet18, ResNet34, and ResNet50 are selected to avoid overfitting of layers. The ResNet network model can extract features from the resulting and constructed image datasets, and continuously learn the image characteristics to achieve a clear image category effect. 4,000 normal pictures and 4,000 violent images (including weapons, blood, and other obvious violent features) were selected as the datasets used in the experiment, and the training set and verification set were randomly divided according to the ratio of 8:2. Pre-processing operations such as horizontal flipping, normalization processing, and cropping are performed randomly before the image is entered into the residual network model.

The network model parameters are fine-tuned using batch parameter 32, a constant learning rate of 0.0001. Through 80 epoch trainings in the hope of finding the most suitable network model parameters. Throughout the training process, after the original image is resized to a size of 224×224 pixels, as input to the entire network, the network fine-tuns the error by backpropagating to the previous number of layers. Experiments were performed on two 2080ti GPUs. Experimentally, four sets of models, ResNet18, ResNet34, ResNet50, and ResNet50-se-pre, were trained on 80 epochs on a given dataset. In addition, accuracy rate refers to the ratio between the total number of samples correctly identified in the identification and classification task. In common classification tasks, the higher the accuracy is, the better the effect of the model is. The experiment in this paper takes the accuracy ratio (ACC) as the accuracy of the evaluation model classification, and the calculation formula of the accuracy rate is shown in Eq. (6): where TP is represented, TN is represented, FP is represented, and FN is represented. In this paper, the experiment is performed under a pre-set test set, and the results are judged by the results of the confusion matrix.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

4.4 Analysis of experimental results

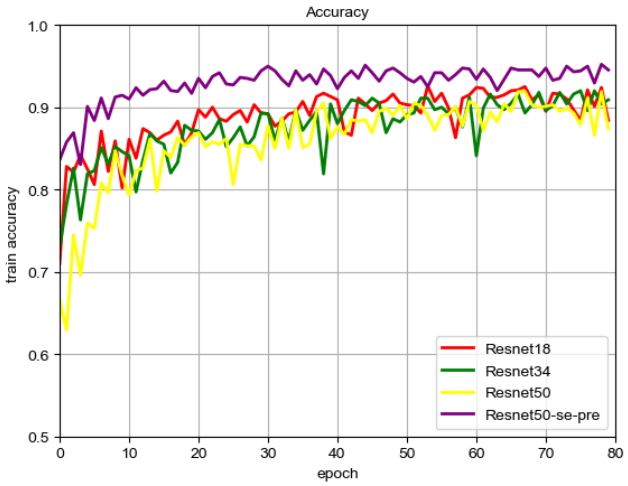
To verify the performance gain of the model improvement method, this paper takes the ResNet50 network as the basic model, and compares the detection accuracy on the self-built dataset by introducing the pre-trained weight file and embedding the SE module, and the experimental results are shown in **Table 1**.

Table 1. Average accuracy rate

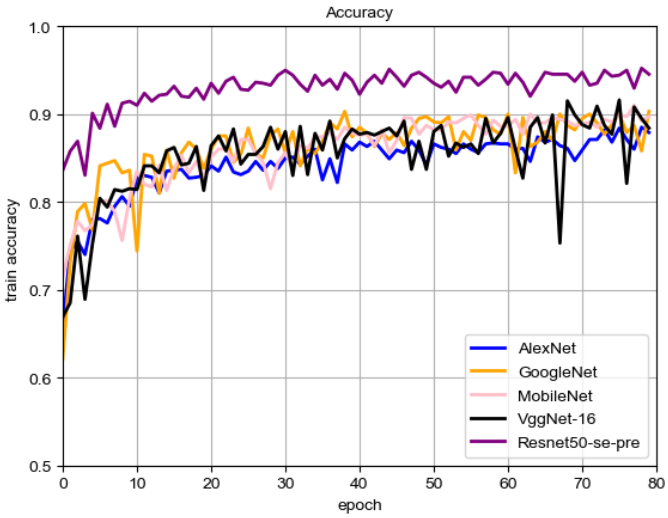
Model	Average classification accuracy /%
ResNet18	89.79
ResNet34	90.72
ResNet50	91.71
ResNet18	89.79

Available from Table 1, the SE module is added to the original model, so that the model can take full advantage of the correlation of the channel, bringing 1.3% performance to the model. In addition, based on the ResNet50 model, **Fig. 5 (a)** shows the change in accuracy when training 80 epochs for a four-class residual network. The results showed that as the

number of layers increased, the accuracy rate increased, and the accuracy of the ResNet50-SE-pre model that introduced pre-trained and SE structures was always higher than that of other control groups, as noted. After adding the SE module, you can speed up the convergence of the model and improve its accuracy of the model.



(a) Residual network training process diagram



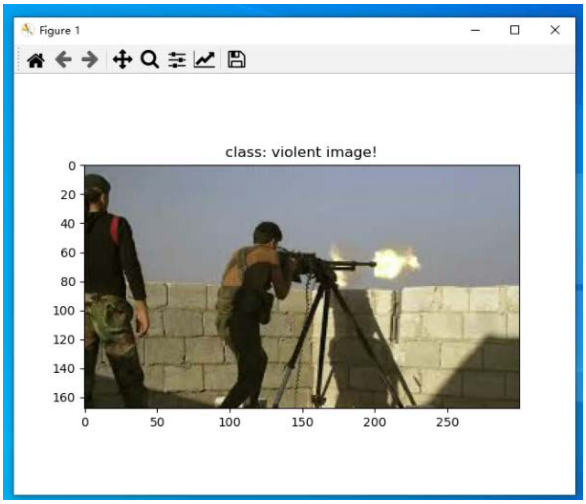
(b) Control group training process diagram

Fig. 5. Model training process diagram

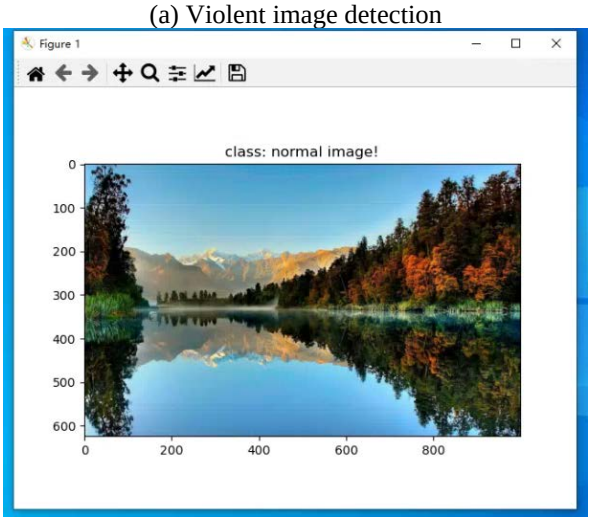
To further verify the performance of the model in this paper, this paper selects other cutting-edge image classification detection methods in this field for comparison, including AlexNet, GoogleNet, MobileNet, VggNet-16, the experimental training process is shown in **Fig. 5 (b)**, and the experimental comparison results are shown in **Table 2**. From the table, it can be concluded that the ResNet50-se-pre proposed in this paper can effectively improve the accuracy of the model in the detection of violent images, which is better than the existing mainstream algorithms. An example of using the ResNet50-se-pre model to detect violent images versus normal images is shown in **Fig. 6**.

Table 2. Comparison of detection accuracy rates of each model Table

Model	Average classification accuracy /%
AlexNet	88.12
GoogleNet	86.37
MobileNet	89.63
VggNet-16	87.61
ResNet50-SE-pre	92.91



(a) Violent image detection



(b) Normal image detection

Fig. 6. ResNet50-se-pre model detection result diagram

Through the longitudinal comparative analysis of the residual network models with different layer numbers, the residual block of ResNet18 is the basic residual block, and the in-block convolutional block structure is Basic[2,2,2,2]; the residual block of ResNet34 is the base residual block, and the in-block convolutional structure is Basic[3,4,6,3]; the residual block of ResNet50 is the residual block, and the in-block convolutional structure is Boostneck[3,4,6,3]. An increase in the number of convolutions can effectively improve the accuracy of the model, but the improvement is limited under the premise of a small data set.

The emergence of the Residual Structure of ResNet50 avoids the phenomenon of gradient explosion to the greatest extent and achieves a relatively good result. The lateral analysis is performed on whether pre-trained weights are introduced under the same number of layers and whether SE attention mechanism blocks are added. The introduction of pre-training blocks, that is, the training of the network model in advance, can effectively improve the accuracy of the network model from the beginning of training. The next step is to consider introducing attention blocks under deep network models such as ResNet101 to see if the results are significantly improved. Further analysis of the results of the model classification institute (taking ResNet50-SE-pre as an example), that is, the focus analysis is incorrectly predicted as a phobia-related picture (False Positive, false positive) and incorrectly predicted as a normal picture (False Negative, false negative), it can be obtained: (1) specific slogans become the most effective features for identifying violent images, and the false negatives and false positives in the predicted results are the lowest among all categories; (2) among the false negatives, Mainly a single characteristic of violence and flames are prone to misjudgment. In summary, the fusion of multiple feature elements to reduce the error rate will be the focus of the work thereafter.

5. Conclusion

As a basic work in the governance of cyberspace, the detection of violent images, especially the detection of violent elements, has been a research hotspot for computer vision researchers in the past period. Although traditional violent image detection uses a small amount of automatic detection based on deep learning models, it is ultimately based on manual judgment. At present, the transfer learning method based on deep learning is widely used in the field of small sample image detection, such as plant disease and pest detection, metal surface defect detection, etc. However, it is still a minority to use transfer learning-based neural network models for phobic image detection studies.

In this study, it is proposed that the deep CNN model be used for the detection of violent-related images. To consider that training from scratch with few samples is prone to overfitting, a transfer learning strategy is adopted and an attention module is introduced. ResNet is used as a basic network framework for its good classification effect. When building the model, the SE attention block is introduced based on the original network; the weight file pre-trained by ResNet on the ImageNet dataset is loaded first, and then fine-tuned according to the built dataset, and the optimal parameters and network structure are discussed. The model used the 8x cross-validation method to train the image input after sizing, and after 80 epochs, the ideal result was finally obtained. Experimental results show that the accuracy of the improved model can reach 92.91% on specific data sets, which is better than other common network models, especially in the detection of specific slogans.

Acknowledgement

This work was supported by Fundamental Research Funds for the Central Universities, People's Public Security University of China (2021JKF215), Key Projects of the Technology Research Program of the Ministry of Public Security(2021JSZ09) and the Fund for the training of top innovative talents to support master's degree program, People's Public Security University of China(2021yjsky018).

References

[1] B. Ionescu, M. Ghenescu, F. Răstoceanu, R. Roman and M. Buric, "Artificial Intelligence Fights Crime and Terrorism at a New Level," *IEEE MultiMedia*, vol. 27, no. 2, pp. 55-61, 1 April-June 2020. Article (CrossRef Link).

[2] T. Xia and Y. Gu, "Building Terrorist Knowledge Graph from Global Terrorism Database and Wikipedia," in *Proc. of 2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 194-196, 2019. Article (CrossRef Link).

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E., "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, 60(6), 84–90, 2017. Article (CrossRef Link)

[4] Szegedy C, Liu W, Jia Y, et al, "Going Deeper with Convolutions," in *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. Article (CrossRef Link)

[5] Simonyan, Karen, and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] He, K., Zhang, X., Ren, S., & Sun, J, "Deep Residual Learning for Image Recognition," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Article (CrossRef Link)

[7] Z. G. Qu, H. R. Sun and M. Zheng, "An efficient quantum image steganography protocol based on improved EMD algorithm," *Quantum Information Processing*, vol. 20, no. 53, pp. 1-29, 2021. Article (CrossRef Link)

[8] Z. G. Qu, Y. M. Huang and M. Zheng, "A novel coherence-based quantum steganalysis protocol," *Quantum Information Processing*, vol.19, no. 362, pp. 1-19, 2020. Article (CrossRef Link)

[9] L.Yan, "Automatic labeling method for violent images based on integrated classification," *Journal of Terahertz Science and Electronic Information Technology*, 18(02), 306-312, 2020. Article (CrossRef Link)

[10] M. Chen, "Automatic identification of violent images based on transfer learning," *Journal of Beijing University of Aeronautics and Astronautics*, 46(09), 1677-1681, 2020. Article (CrossRef Link)

[11] Hu Xinxu, Zhou Xin, He Xiaohai, Xiong Shuhua, Wang Zhengyong, "The Moby Audio Discrimination Method based on Transfer learning," *Computer System Applications*, 28 (11), 147-152, 2019. Article (CrossRef Link)

[12] Huang Chao, Yi Ping, "Research and implementation of content detection systems for video-phobias," *Communication technology*, 51(01), 75-81, 2018. Article (CrossRef Link)

[13] S. Sun, J. Zhou, J. Wen, Y. Wei and X. Wang, "A dqn-based cache strategy for mobile edge networks," *Computers, Materials & Continua*, vol. 71, no.2, pp. 3277–3291, 2022. Article (CrossRef Link)

[14] Meng Caixia., "Terror Detection Simulation based on Fusion Two-Channel Video," *Computer Simulation*, 32(02), 428-431, 2015.

[15] A. Kaur and L. Kaur, "Concealed weapon detection from images using SIFT and SURF," in *Proc. of 2016 Online International Conference on Green Engineering and Technologies (IC-GET)*, pp. 1-8, 2016. Article (CrossRef Link)

[16] N. J. Hussein and F. Hu, "An alternative method to discover concealed weapon detection using critical fusion image of color image and infrared image," in *Proc. of 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI)*, pp. 378-383, 2016. Article (CrossRef Link)

[17] Liu, Y., et al., "Abnormal Behavior Recognition Based on Key Points of Human Skeleton," *IFAC-PapersOnLine*, 53(5), 441-445, 2020. Article (CrossRef Link)

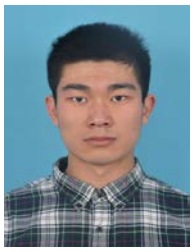
[18] Karthikeswaran D, Sengottaiyan N, Anbukaruppusamy S, "Video surveillance system against anti-terrorism by Using Adaptive Linear Activity Classification (ALAC) Technique," *Journal of Medical Systems*, 43(8), 256, 2019. Article (CrossRef Link)

[19] C. Dhiman and D. K. Vishwakarma, "A review of state-of-the-art techniques for abnormal human activity recognition," *Eng. Appl. Artif. Intell.*, vol. 77, no. August 2018, pp. 21-45, 2019. Article (CrossRef Link)

[20] Z. Guo F. Wu H. Chen J. Yuan and C. Cai, “Pedestrian violence detection based on optical flow energy characteristics,” in *Proc. of 2017 4th Int. Conf. Syst. Informatics ICSAI 2017*, pp. 1261-1265, 2017. Article (CrossRef Link)
[21] S. Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” in *Proc. of 2017 14th IEEE Int. Conf. Adv. Video Signal Based Surveillance AVSS*, 2017. Article (CrossRef Link)
[22] P. Yun J. Jiao and M. L. B, “Trajectory-Pooled Deep Convolutional Networks for Violence Detection in Videos,” in *Proc. of ICVS 2017 Computer Vision Systems*, pp 437–447, 2017. Article (CrossRef Link)
[23] F. U. M. Ullah A. Ullah K. Muhammad I. U. Haq and S. W. Baik, “Violence detection using spatiotemporal features with 3D convolutional neural network,” *Sensors*, vol. 19, no. 11, pp. 1-15, 2019. Article (CrossRef Link)
[24] S. Chaudhary M. A. Khan and C. Bhatnagar, “Multiple anomalous activity detection in videos,” *Procedia Comput. Sci.*, vol. 125, pp. 336-345, Jan. 2018. Article (CrossRef Link)
[25] A. B. Mabrouk and E. Zagrouba, “Abnormal behavior recognition for intelligent video surveillance systems: A review,” *Expert Syst. Appl.*, vol. 91, pp. 480-491, Jan. 2018. Article (CrossRef Link)
[26] E. Y. Fu M. X. Huang H. Va Leong and G. Ngai, “Cross-species learning: A low-cost approach to learning human fight from animal fight," in *Proc. of 26th ACM Int. Conf. Multimedia*, pp. 320-327, Oct. 2018. Article (CrossRef Link)
[27] M. Ahmed, M. Ramzan, H. U. Khan, S. Iqbal, M. A. Khan et al., “Real-time violent action recognition using key frames extraction and deep learning,” *Computers, Materials & Continua*, vol. 69, no. 2, pp. 2217–2230, 2021. Article (CrossRef Link)
[28] M. Al-Nawashi O. M. Al-Hazaimeh and M. Saraee, “A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments," *Neural Comput. Appl.*, vol. 28, no. 1, pp. 565-572, Dec. 2017. Article (CrossRef Link)
[29] P. Siva and N. Nandhagopal, “Classification similarity network model for image fusion using ResNet50 and googlenet,” *Intelligent Automation & Soft Computing*, vol. 31, no. 3, pp. 1331–1344, 2022. Article (CrossRef Link)
[30] R. U. Khan, W. S. Wong, I. Ullah, F. Algarni, M. Inam et al., "Evaluating the efficiency of cbam-ResNet using malaysian sign language," *Computers, Materials & Continua*, vol. 71, no.2, pp. 2755–2772, 2022. Article (CrossRef Link)
[31] L. Sun, Y.L. Wang, Z.G. Qu, N.N. Xiong, "BeatClass: A Sustainable ECG Classification System in IoT-based eHealth,” *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7178-7195, 2022. Article (CrossRef Link)
[32] Yifei Wei, F. Richard Yu, Mei Song, Zhu Han, “User Scheduling and Resource Allocation in HetNets with Hybrid Energy Supply: An Actor-Critic Reinforcement Learning Approach,” *IEEE Transactions on Wireless communications*, vol. 17, no. 1, pp. 680-692, Jan. 2018. Article (CrossRef Link)
[33] J. Hu, Z. Zhang, Y. Zhao, “Identification of bamboo chip defects based on transfer learning,” *Journal of Northwestern Forest College*, vol. 36(5), pp. 190-196, 2021. Article (CrossRef Link)
[34] Y. Zhang, L. Zhu, Yu, “Overview of attention mechanisms in convolutional neural networks,” *Computer Engineering and Applications*, pp. 64-72, 2021.

Yuyan Meng is currently pursuing a master's degree at People's Public Security University of China. Her research interests include image classification, deep learning.

De-yu Yuan, born in 1986, Ph.D, lecturer, Ph.D supervisor. His main research interests include cyber security and complex networks



Shaofan Su is currently pursuing a master's degree at People's Public Security University of China. His research interests include image classification, deep learning.