Multi-classification Sensitive Image Detection Method Based on Lightweight Convolutional Neural Network

Yueheng Mao[1,2], Bin Song[1,2*], and Zhiyong Zhang[1,2], Wenhou Yang[3], and Yu Lan[3]
[1] Information Engineering College, Henan University of Science and Technology
Henan Luoyang 471023, China
[2] Henan International Joint Laboratory of Cyberspace Security Applications
Henan University of Science and Technology, Henan Luoyang 471023, China
[e-mail: songbin@haust.edu.cn]
[3] Sunnetech Ltd., Quzhou, Zhejiang 324003
[*]Corresponding author: Bin Song

Received November 26, 2022; revised April 3, 2023; accepted April 13, 2023; published May 31, 2023

Abstract

In recent years, the rapid development of social networks has led to a rapid increase in the amount of information available on the Internet, which contains a large amount of sensitive information related to pornography, politics, and terrorism. In the aspect of sensitive image detection, the existing machine learning algorithms are confronted with problems such as large model size, long training time, and slow detection speed when auditing and supervising. In order to detect sensitive images more accurately and quickly, this paper proposes a multi-classification sensitive image detection method based on lightweight Convolutional Neural Network. On the basis of the EfficientNet model, this method combines the Ghost Module idea of the GhostNet model and adds the SE channel attention mechanism in the Ghost Module for feature extraction training. The experimental results on the sensitive image data set constructed in this paper show that the accuracy of the proposed method in sensitive information detection is 94.46% higher than that of the similar methods. Then, the model is pruned through an ablation experiment, and the activation function is replaced by Hard-Swish, which reduces the parameters of the original model by 54.67%. Under the condition of ensuring accuracy, the detection time of a single image is reduced from 8.88ms to 6.37ms. The results of the experiment demonstrate that the method put forward has successfully enhanced the precision of identifying multi-class sensitive images, significantly decreased the number of parameters in the model, and achieved higher accuracy than comparable algorithms while using a more lightweight model design.

Keywords: Sensitive image detection, Lightweight convolutional neural network, EfficientNet, Model Pruning.

This work was supported by National Natural Science Foundation of China under Grant 61972133, Project of Leading Talents in Science and Technology Innovation in Henan Province under Grant 204200510021, Program for Henan Province Key Science and Technology under Grant 222102210177 and 232102211060, Henan Province University Key Scientific Research Project under Grant 23A520008, Henan Province Natural Science Fund under Grant 232300420148.

http://doi.org/10.3837/tiis.2023.05.007 **ISSN : 1976-7277**

1. Introduction

During recent years, with the rapid development of Internet technology, multimedia content has also shown an exponential growth trend, and the problem brought by the growth of multimedia content is the pressure from the supervision. Among the multimedia contents, there is no lack of information related to pornography, politics, and violence. If this information is not supervised, it will have a very negative impact on the physical and mental health of young people and social stability. So how to prevent the dissemination of sensitive information through images stably and efficiently, protect the physical and mental health of minors and maintain social stability is an urgent problem we need to solve.

In early studies, many researchers tend to use traditional machine learning algorithms to solve the classification problem of sensitive images [1-3]. Most of these methods are based on low-level features, which extract the feature information such as texture, color, contour and relative size of the image region, and then find the category of each region by genetic algorithm. Finally, the object categories to which the image belongs are added up to calculate the probability that the image belongs to a certain classification.

With the excellent performance of deep learning in computer vision in recent years, many scholars tend to apply deep learning methods to the classification of sensitive images [4-6]. These methods are mainly based on the powerful feature extraction ability of CNN for images, which can effectively discriminate deep features with differentiation, so as to classify sensitive images.

However, in recent years, research on detection tasks for sensitive images has ignored the practical application scenarios and focused on image feature extraction and detection, which cannot accomplish the task of performing sensitive images under the huge data streams of the Internet. To solve this problem, this paper proposes a multi-classification sensitive image detection method based on lightweight convolutional neural network (CNN).

This study proposes a model framework to optimize the feature extraction layer of CNN, and uses the Ghost Module [7] with the addition of the SE [8] channel attention mechanism to achieve feature superposition and complete the expansion of feature layers. The superposed features are then input to the baseline EfficientNet [9] network composed of lightweight flip bottleneck convolution kernels for training. The minimal increase in the number of parameters leads to considerable improvement in accuracy. Next, the Hard-Swish [10] activation function and the model pruning operation are introduced to optimize the original baseline network structure to improve the detection speed of the model. By fusing the above strategies, the multi-classification sensitive image detection method based on lightweight CNNs is proposed, and the detection accuracy and speed of sensitive images are further improved in this paper. Through the comparative test and the attention visualization method, it is shown that the structure proposed in this paper can complete the sensitive information detection task more accurately when the number of model parameters is greatly reduced and the detection speed is significantly increased.

2. Related Work

2.1 Research Progress of Sensitive Image Detection

In the early stage of sensitive image detection, skin color features were mainly detected. In 2011, H. Yin et al. [11] proposed a hybrid large skin region detection method based on the three components of color filtering, texture filtering, and geometric filtering. JORGE et al. [12]

proposed the use of YCbCr color space to detect pornographic images in the same year. In 2017, Safira Nuraisha [13] et al. proposed the use of color distribution histograms with specific skin color features as a discriminatory basis for classification. These methods mainly focus on the comprehensive analysis of skin tone models, texture models of skin, and skin mask image processing for the extraction of sensitive image features.

Later, some researchers shifted their focus to the intimate parts of the human body, so that if some of the intimate organ features were present in the image, it would indicate that the image had a pornographic element, and then the pornographic image would be recognized through this detection. Lintao Lv et al. [14] proposed an improved BoVW model to filter pornographic images by high-level semantic features. A.P.B. Lopes [15] used the HUE-SIFT detector with enhanced luminance to detect local features and finally classify the image with an SVM algorithm. This type of method is more reliable and intelligent compared to skin color detection methods, but it requires the researcher to spend a lot of time adjusting the parameters of the feature extractor. Moreover, in the detection task of multiclassification sensitive images, the generalization of these methods is usually poor due to the diversity of sensitive images and the confusability of physical attribute features.

In recent years, the technique of deep learning has continued to make innovative developments. This approach has attracted many researchers to apply deep learning to image detection with exciting results. In the case of violent image detection, DONGHYEON et al. [16] used a multi-task CNN to classify the presence of protesters in images and identify their activities, and used these as conditions to estimate the level of perceived violence in images. In the area of pornographic image detection, Connie T et al. [17] constructed eight CNNs with different structures to classify images separately, and then used the least square method to calculate the individual weights for all classification results, finally fusing them to give the final recognition results. Lin et al [18] fused four frozen l different shallow DenseNet-121 [19] models for a pornographic image detection task, and this approach improved the accuracy by about 1% over a single DenseNet121 network. Olarik Surinta et al. [20] used AlexNet, GoogleNet and ResNet, three typical CNN frameworks, to identify and compare pornographic images, and finally found that ResNet could obtain higher accuracy in the recognition of pornographic images.

However, studies in recent years have lacked the detection of identification involving political figures and political symbols. Moreover, the above research works are only based on the detection task of sensitive images for accuracy improvement, ignoring the actual production reality. A larger number of model parameters means longer model training and loading times. Therefore, the use of lightweight CNN will be more compatible with the model requirements in real applications.

2.2 Related Work of Lightweight CNN Research

Based on the requirement of model detection speed, some lightweight neural networks have been proposed by researchers and applied in the field of computer vision, such as Squeezenet [21], Shufflenet [22,23], Mobilenet series model [24,25,13], GhostNet [7], etc. Their design idea is different from the traditional model structure block stacking to improve the accuracy, mainly through adopting a more efficient method for model design, so as to achieve the requirements of reducing the number of parameters and model compression, and improving the speed of model recognition.

The Mobilenet series model is a lightweight neural network proposed by Andrew G. of Google, whose main idea is first to replace the traditional standard convolution in neural networks with the depthwise separable convolution. Later, the authors proposed the inverse

residual network in version V2, followed by the addition of the SE attention mechanism in version V3, which continued to optimize the network using the NAS search parameter [26] and other methods. ShuffleNet is a channel shuffling operation using group convolution to fuse the features of each group, which reduces the computational effort while can maintaining the accuracy and improves the efficiency of the model while reducing the network running time. GhostNet adopts the Ghost module architecture, which is based on the ordinary convolution operation, adding simple linear transformations to produce a large number of ghost feature mappings, and this approach is able to reveal deeper information of intrinsic features. It can acquire the features of each channel with a very small amount of convolution and stitch them with the original feature layer, gaining the same number of features using a smaller amount of computation.

2.3 EfficientNet Model

In the development of deep neural networks, increasing the depth and width of CNN or increasing the resolution of input images are common methods used by researchers to improve the accuracy of their models. This method is highly subjective and therefore it is difficult to design the model, which may cause the model to be too large and cannot ensure the accuracy of the model.

In 2019, the researchers of Google Brain, Tan et al., proposed a method of uniformly scaling the depth, width and input resolution of network models in literature [9] to ensure the balance between model accuracy and model size. Based on this approach, Google Brain proposed the EfficientNet family of models with powerful performance, convenient deployment, easy training, and high accuracy.

However, this series of models, like the traditional CNN, chooses to reduce the resolution of the image using a convolutional kernel with a stride of 2 in the first module. This operation will directly convolute and scale the image resolution, so when detecting sensitive images, a large amount of original image information will be lost due to the complexity of the image. Also, the model makes extensive use of the Swish activation function. The advantage of the Swish [27] activation function is that, although it is very similar to the ReLU graph, it performs slightly better in that it does not change abruptly at some point, which makes it easier to converge during training. However, the authors of the paper [10] found that the Swish activation function only works in deep networks. And the high computational cost is also a major drawback of Swish due to its own computational volume, which will greatly reduce the detection efficiency of the model when detecting large batches of multi-classified sensitive images. Therefore, this model is not better for the task of detecting sensitive images in the current network environment where huge amounts of data are generated every moment of the day.

3. Approach

In order to further improve the detection effect of the EfficientNet model, this paper introduces the Ghost Module with the attention mechanism of the SE channel (namely, SE-Ghost Module) to further improve it, so as to enhance the processing of the original image information by the model and enrich the information in the convolutional feature map. Meanwhile, the Hard-Swish activation function and model pruning are introduced to improve the detection rate of the model.

Accordingly, the model with the SE-Ghost Module added in this paper is named GhostEfficientNet, and the lightweight convolutional neural model with Hard-Swish activation function introduced and pruned on this basis is named GhostEfficientNet_s.

3.1 Ghost Module

The Ghost Module was first proposed in the GhostNet model. As a lightweight CNN model, which uses a novel end-side neural network architecture, it proceeds in two steps, as shown in **Fig. 1**.

The Ghost module first generates feature maps using a small number of convolution kernels, then uses these feature maps to obtain Ghost feature maps by group convolution, finally superimposing them to achieve the effect of increasing the number of channels. In this way, the Ghost module is able to increase the number of channels with a small number of parameters and very low computational complexity.

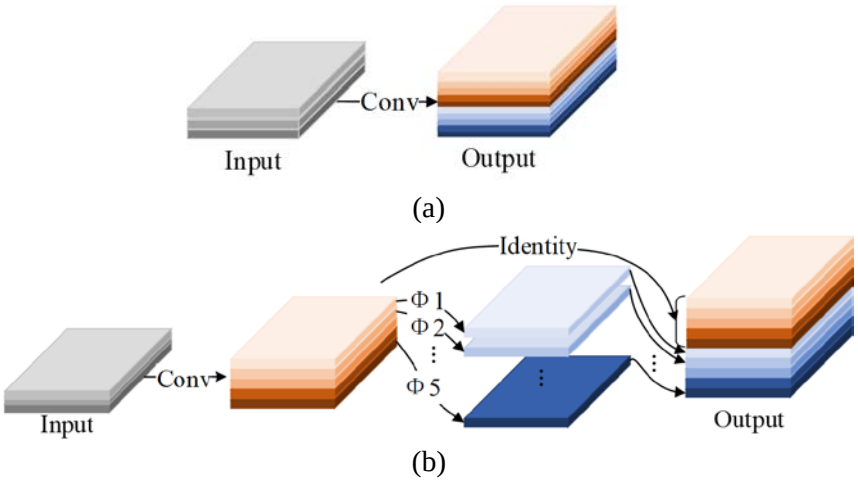


Fig. 1. An illustration of the convolution layer and the introduced Ghost module for outputting the same number of feature maps. Φ represents the group convolution operation.

3.2 SE Channel Attention Mechanism

For convolution operations, a necessary operation to improve performance is to increase the perceptual field of the model, as in the multi-branch structure of the Inception [28] network to fuse more features in space. For the convolution operation, the default practice is to fuse the feature maps obtained from each convolution kernel directly. However, the capability of the SE channel attention mechanism lies in the ability to enable the model to focus on the weight information between channels and to automatically learn the importance of different channel features through training.

The process of an SE channel attention mechanism is divided into two steps: Squeeze and Excitation, which can be applied to all mappings in the convolution operation: $\mathrm{F}_{conv}: I \to O, I \in R^{H' \times W' \times C'}, O \in R^{H \times W \times C}$, Where $I = [i^1, i^2, \cdots, i^{C'}]$ stands for input and $O = [o_1, o_2, \cdots, o_c]$ stands for output. Set the set of kernels to $V = [v_1, v_2, \cdots, v_c]$, and v_c represents the nth kernel, for a convolution operation:

$$o_c = v_c * i = \sum_{s=1}^{C'} v_c^s * i^s \tag{1}$$

Where * represents a convolution operation and represents the kernel of an s-channel. The process of adding the SE channel attention mechanism to it is as follows:

Squeeze: Compressing the current feature map from H × W × C to 1 × 1 × C by performing global average pooling on the feature map layer can help the lower layers of the network utilize the information from the global sensory field:

$$z_c = F_{sq}(o_c) = \frac{1}{H \times W} \sum_{j=1}^{H} \sum_{k=1}^{W} o_c(j,k), z \in R^C \tag{2}$$

Excitation: Obtaining the weights of each channel in the feature map through a bottleneck structure consisting of two fully connected layers:

$$s = F_{ex}(z,W) = \sigma(g(z,W)) = \sigma(W_2 \delta(W_1, z)) \tag{3}$$

where $W_1 \in R^{\frac{C}{r} \times C}, W_2 \in R^{C \times \frac{C}{r}}$, σ represents the Sigmoid activation function. This paper considers the optimization of it and replaces it with the Hard-Swish activation function, while δ represents the ReLU6 activation function. The role of the first fully connected layer is to reduce the dimension, where r is a hyperparameter with a default of 4. After activation by the Hard-Swish activation function, a second fully connected layer is used to restore it to its original dimensionality.

Finally, the learned activation values of each channel are multiplied with the original features O, making the model more discriminative by assigning weights to the features of each channel. The formula is as follows:

$$\tilde{i}_c = F_{scale}(o_c, s_c) = s_c o_c \tag{4}$$

3.3 Hard-Swish Activation Function

The original EfficientNet network model improves the accuracy of the neural network by using a large number of Swish activation functions in the model, which is defined as:

$$Swish(x) = x \cdot \sigma(x) \tag{5}$$

Although this nonlinearity function improves accuracy, the sigmoid function is much more computationally expensive on mobile devices because it is made up of exponentials. Sigmoid activation functions can be fitted with Hard-Sigmoid segmented linear functions:

$$Hardsigmoid(x) = \begin{cases} 0, & x \le -3 \\ 1, & x \ge 3 \\ \frac{x}{6} + \frac{1}{2} & otherwise \end{cases} \tag{6}$$

Therefore, replacing the sigmoid function with the hard-sigmoid function can greatly reduce the operation cost. Hard-Swish was born with the following formula:

$$HardSwish = \begin{cases} 0, & x \le -3 \\ x, & x \ge 3 \\ x \cdot (\frac{x}{6} + \frac{1}{2}) & otherwise \end{cases} \tag{7}$$

The Hard-Swish function has almost no significant performance difference from Swish

but has multiple advantages from a deployment perspective. The segmentation function can reduce the number of memory accesses, thus significantly reducing latency costs.

3.4 Improve EfficientNet Network Structure

The schematic diagram of the improved GhostEfficientNet_s network structure is shown in **Fig. 2**.

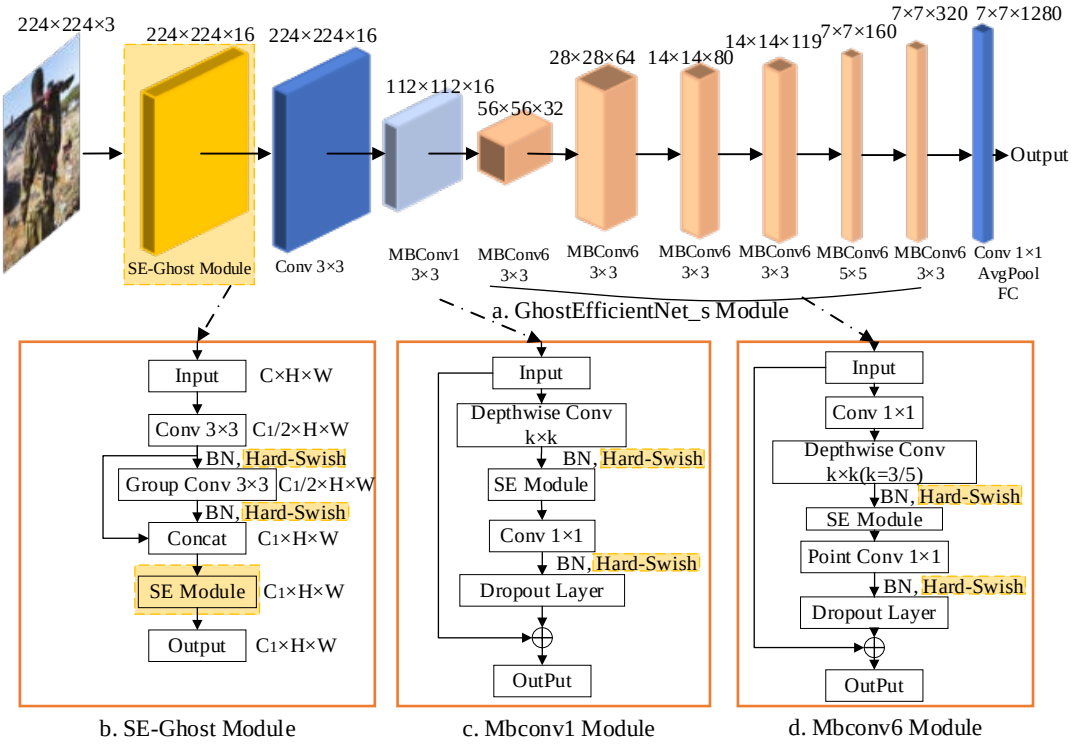


Fig. 2. Network architecture of GhostEfficientNet_s. Conv represents a convolutional kernel; MBConv represents a light flip bottleneck convolution kernel; 1 or 6 represents the number of amplification channels in the middle layer of the linear bottleneck block; AvgPool represents the adaptive average pooling; FC represents the fully connected layer; BN represents the batch normalization; Hard-Swish represents the segmented linear activation function; C represents the number of input channels; C1 represents the number of output channels; H and W represent the height and width of the feature map, respectively; Concat represents stacking the feature maps obtained by normal convolution and the feature map obtained by direct connection through group convolution for channel stacking; ⊕ represents the channel-level summation, and each direct connection branch represents the fusion of the input feature map with the feature map obtained through convolution and other operations.

As shown in **Fig. 2**, the process of multi-category sensitive image detection and recognition based on this model is as follows:

First, the input sensitive image is converted into 224 pixels × 224 pixels × 3 channels by preprocessing operations such as data enhancement and input to GhostEfficientNet_s. Then, a small number of convolution kernels are used in the SE-Ghost Module to generate feature maps, which are then stacked with these feature maps by group convolution to obtain more

image features. Subsequently, these feature maps are assigned weights by the SE channel attention mechanism, followed by a reduced-dimensional convolution with seven lightly flipped bottleneck convolution layers using the Hard-Swish activation function to extract the high-level features of sensitive images, resulting in a 7 pixels × 7 pixels × 1280 channels feature map. Finally, adaptive average pooling of feature maps and mapping of sensitive image features to specific dimensions using a fully connected layer are performed to achieve classification.

On the basis of the EfficientNet network model, this study makes the following main improvements to it:

1) Before the first module of the EfficientNet model, this paper uses the Ghost Module to perform feature extraction first, which uses a small amount of convolutional computation to generate feature maps for the RGB three channels of the original image to obtain more image features so that the model can extract more features from the original image with a simple operation. When dealing with complex multi-classification sensitive images, it is able to classify them more accurately.

2) Since Ghost Module eventually stacks features in a simple stacking manner, the information between channels is ignored. Therefore, this paper proposes to improve the Ghost Module module by using the SE channel attention mechanism to assign weights to the channels after they are stacked to filter out the more important information from the feature map. Comparative experiments demonstrate that the model with this structure added has the best accuracy, i.e., GhostEfficientNet model mentioned above.

3) Although the original EfficientNet network model improves the detection speed of the model as much as possible while ensuring its accuracy, it still has a relative disadvantage compared with the current mainstream lightweight CNN model. Therefore, this paper introduces the Hard-Swish multi-terminal linear activation function to activate the SE channel attention mechanism and the convolutional features, followed by ablation and pruning of the model to improve the detection speed of the model while ensuring the detection accuracy. The **Table 1** shows the architecture of GhostEfficientNet_s.

Table 1. The architecture of GhostEfficientNet_s. Each row describes a stage with n layers, input resolution, and output channels.

Stage	Module Name	Resolution	Channels	Layer
1	SE-Ghost Module	224	16	1
2	Conv, k=3	224	16	1
3	MBConv1, k=3	112	16	1
4	MBConv6, k=3	112	32	1
5	MBConv6, k=5	56	64	2
6	MBConv6, k=3	28	80	2
7	MBConv6, k=5	14	119	2
8	MBConv6, k=5	14	160	1
9	MBConv6, k=3	7	320	1
10	Conv1×1&Pooling&FC	7	1280	1

4. Experiments

In order to verify the effectiveness of GhostEfficientNet_s, this paper evaluates the practical application of the designed method by comparing the analysis of experimental results.

4.1 Experimental Models

The control group of the experiment is a relatively popular image recognition model in the current research field of lightweight CNN. The experimental group is the GhostEfficientNet model with the addition of the Ghost module with the SE channel attention mechanism, and the GhostEfficientNet_s with model pruning and Hard-Swish activation function replacement based on it.

4.2 Experimental Environment

The experimental environment: The Python version is Python 3.6. On this basis, the open-source Python machine learning library Python 1.10.0 is used with Cuda 11.3 for training. All experiments were trained using the AdamW [29] optimizer with the same number of experimental batches and epochs. The initial value of the learning rate was chosen as 0.005, and the amplification and continuous decay were performed in the first round using the Warmup method, and the loss function was adopted as the Cross-Entropy Loss Function.

Hardware conditions of the training environment: the GPU model is RTX 3080; GPU memory is 10GB; Memory 16G; The CPU is 12 core Intel (R) Xeon (R) Platinum 8255.

Hardware conditions of the testing environment: the GPU model is GTX1650; GPU memory is 4GB; Memory 8G; The CPU is Intel core i5-9300H.

4.3 Dataset Collection and Enhancement

Since there is no unified public dataset for work related to sensitive image detection, this study searched the NSFM public dataset and several Kaggle datasets, and then used a Python program to extract relevant images from them to enrich the dataset of this study. which includes 3016 pornographic images, 2433 images related to politics, 3325 terrorism images and 5801 normal images. The total number of images is 14575.

In practical applications, considering that the images posted by users will be influenced by factors such as pose, position, and orientation, this paper uses the torchvision framework for image processing by random inversion and normalization during training.

4.4 AdamW Optimizer

The traditional Stochastic Gradient Descent (SGD) method [30] was used in the original EfficientNet network, which draws several small batch samples according to the data generation distribution, calculates their gradient means, updates the parameters, and randomly disrupts the samples at each iteration. However, the hyperparameter setting of SGD is very sensitive where the L2 norm is highly coupled with the initial learning rate.

The Adam optimizer, on the other hand, differs from the traditional SGD optimizer in that it utilizes an adaptive learning rate computed independently for each parameter. Compared with SGD, Adam will bring lower training and testing errors with better generalization performance to the model. However, since Adam does not require artificial adjustment of the learning rate, it causes the situation that it can easily oscillate at local minima and there exists a sudden increase in the learning rate under special data sets, resulting in non-convergence. The AdamW optimizer is an improvement of Adam. AdamW with Weightdecay improves the regularization in Adam by decoupling the weight decay from the gradient-based

update, thereby achieving the decoupling from the initial learning rate parameter. Therefore, AdamW is used as an optimizer in this paper.

Parameter update method of AdamW:

$$\begin{cases} t = t+1 \\ g_t = \nabla f_t\left(\theta_{t-1}\right) \\ m_t = \beta_1 m_{t-1} + \left(1-\beta_1\right) g_t \\ v_t = \beta_2 v_{t-1} + \left(1-\beta_2\right) g_t^2 \\ \tilde{m}_t = m_t / 1-\beta_1^t \\ \tilde{v}_t = v_t / 1-\beta_1^t \\ \theta_t = \theta_{t-1} - \left(\alpha \tilde{m}_t / \left(\sqrt{\tilde{v}_t} + \xi\right) + \omega\theta_{t-1}\right) \end{cases} \tag{8}$$

α : learning rate, this study chooses 5e-4 and dynamically decay it to 5e-10 depending on the number of learning epochs. m_t , v_t : the first moment estimation and second moment estimation of gradient. $\beta_1, \beta_2 \in [0,1)$: exponential decay coefficients, generally taken as 0.9 and 0.999. ξ : prevent the molecule from dividing by 0, generally taking 10e-8. $f(\theta)$: loss function under the learnable parameters θ . Initialize the parameter vector m_0 , v_0 = 0.

5. Result

5.1 Dataset Results for Models

In order to ensure the representativeness of the experimental results, multiple experiments were conducted for all models in this paper, and the effectiveness of the experiments was fully demonstrated by controlling the learning rate, loss function, optimizer, and other variables kept constant. The experimental results are shown in **Fig. 4**.

In order to comprehensively compare the effects of the models in all aspects, this study compares them in terms of accuracy, F1 values, params, and time.

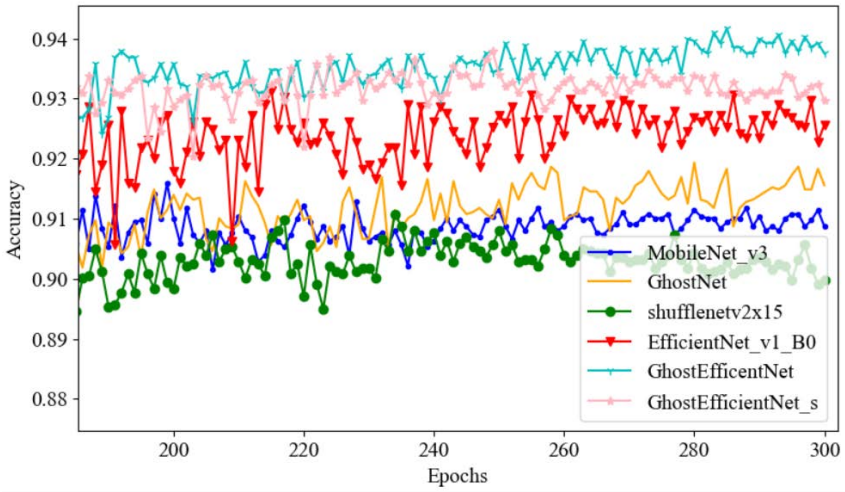


Fig. 3. Comparison of the accuracy of each model on the dataset of this paper

The formula for calculating accuracy in this paper is defined as follows:

$$Accuracy = \frac{P_C}{P_{ALL}} \times 100\% \tag{9}$$

where P_C represents the count of images correctly predicted by the model and P_{ALL} represents the count of all images in the validation set.

In order to evaluate the advantages and disadvantages of different algorithms, the experiments introduce the concept of F1 value to evaluate the correctness and recall of these models with an overall evaluation. which is defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$recall = \frac{TP}{TP + FN} \tag{11}$$

$$\text{F1}=2\frac{Precision \cdot recall}{Precision + recall} \tag{12}$$

Where TP represents the number of predicted positive samples that are predicted correctly, FP represents the number of predicted incorrectly, and FN represents the number of positive samples that are predicted by the model as negative class. Combining the above three formulas, the F1 score can show the prediction effect of the model in a more comprehensive way.

Model speed detection was selected from 4980 sensitive images, and the model detection batch size was set to 1 to test the time consumption of the model during single image detection. In order to exclude interference factors, the first detection speed of this test is not involved in the calculation, and the rest of the batch detection speed is summed and then averaged, the calculation formula is as follows:

$$\text{Time} = \frac{1}{n-1}\sum_{i=2}^{n}(t_{i,end} - t_{i,start}) \tag{13}$$

Where n represents the number of prediction sensitive images, $t_{i,start}$ and $t_{i,end}$ represent the start and end time of the ith image, and the time unit is s.

This study chooses to compare the more excellent lightweight neural network models in recent years with the model in this paper, and the detailed data are shown in **Table 2**.

Table 2. Comparison results on accuracy, F1 score, Params and Time of each model

Model Name	Accuracy	F1Score	Params(M)	Time(s)
squeezenet1_1[21]	90.53%	90.96%	**0.708**	**2.64**
MobileNetV3[10]	91.42%	91.99%	4.204	9.62
GhostNet[7]	92.10%	92.52%	4.207	11.33
ShuffleNetv2[23]	90.97%	91.48%	2.511	7.97
EfficientNet_b0[9]	93.13%	93.59%	4.013	8.88
GhostEfficientNet_s	**93.79%**	**94.07%**	1.862	6.37

It can be concluded from **Fig. 3** and **Table 2** that GhostEfficientNet can achieve the best accuracy in the same task of sensitive image recognition. The GhostEfficientNet_s model with Hard-Swish activation function and pruning also achieves the best accuracy and detection speed after squeezenet1_1 in the classification task of sensitive images, but is significantly higher in accuracy and F1 score than the squeezenet1_1 network model which is composed of a very small number of parameters.

5.2 Results on Confusion Matrix

Under the classification model in this paper, there are sixteen different combinations between the predicted results and the correct markers. The confusion matrix is formed by aggregating the records in the dataset according to the true categories and the judgment categories made by the classification model in a matrix form. This study can use the confusion matrix to evaluate the performance of supervised learning algorithms visually.

Fig. 4 shows the confusion matrix of each comparison model, from which it can be seen that GhostEfficient_s achieves more than 93% accuracy for all kinds of image detection tasks, which is more suitable for deployment in practical applications than other models.

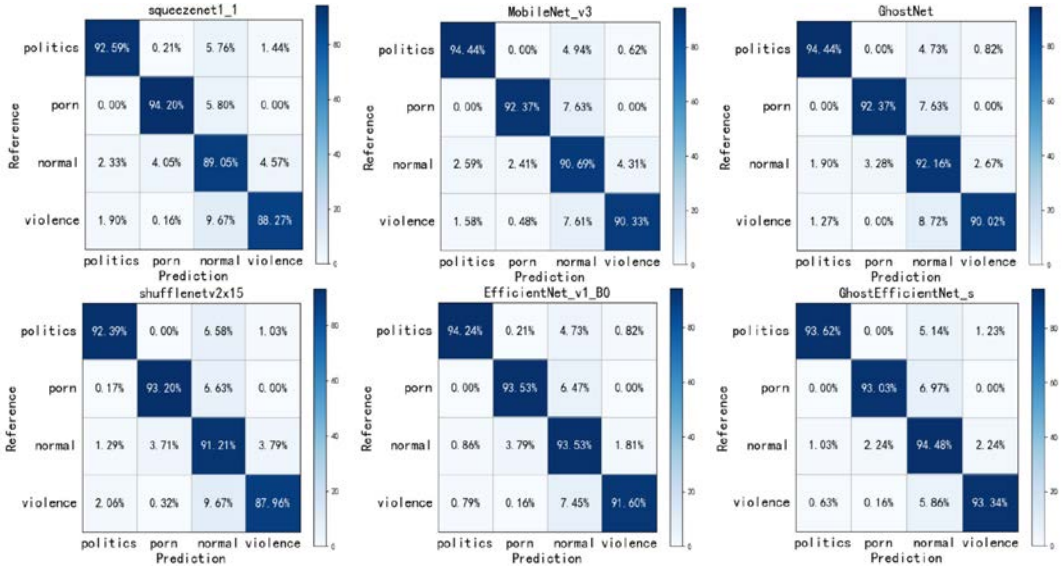


Fig. 4. Results on Confusison Matrix

5.3 GhostEfficient_s Ablation Experiment

To verify the validity of the GhostEfficient_s model proposed in this study, the following five ablation test schemes were used in this experiment:

1) Only EfficientNet baseline model;
2) Adding the Ghost module to 1) for pre-extraction of sensitive image features allows the model to increase the nonlinear operation of the feature extraction layer with a very small number of parameters by this operation, thus reducing the loss of the original image vector features in training. This combined feature extraction enhancement strategy is marked GM (Ghost Module).
3) Based on 2), the Ghost Module is added to the SE channel attention mechanism to assign weights to features extracted from sensitive images through the Ghost Module, which increases the nonlinear relationship between feature channels in the model. This combined feature is marked SE.
4) The swish activation function in the MBconv module of the original model and in the attention mechanism of the SE channel is replaced by the Hard-Swish activation function on the basis of 3), which reduces the linear operation in feature activation and speeds up the model detection. The combination is marked HS;

5) The model is pruned on the basis of 4) to further reduce the model parameters. Thus, the GhostEfficientNet_s proposed in this paper is obtained. the results obtained from the above five schemes are shown in **Table 3**.

Table 3. Results on GhostEfficient_s ablation experiment

No.	Improvement strategy	Accuracy	F1 score	Params(M)	Time(s)
1	EfficientNet	93.13%	93.59%	4.013M	8.88
2	EfficientNet+GM	93.89%	94.17%	4.022M	9.06
3	EfficientNet+GM+SE (GhostEfficientNet)	**94.17%**	**94.46%**	4.022M	9.30
4	EfficientNet+GM+SE+ (Swish->Hard-Swish)	94.03%	94.33%	4.022M	8.90
5	GhostEfficientNet_s	93.79%	94.07%	**1.862M**	**6.37**

By comparing schemes 1 and 2, it can be found that the recognition accuracy of the model is improved by 0.76% and the F1 score is improved by 0.58% by using the Ghost module for feature extraction network strategy. This shows that using the Ghost module for feature extraction can effectively improve the recognition of sensitive images.

Comparing Scheme 2 and Scheme 3, it is shown that the model is able to achieve better accuracy on the detection task of sensitive information by combining the SE channel attention mechanism with the Ghost module.

From the experimental results of schemes 3 and 4, it is clear that the adoption of the Hard-Swish activation function instead of the Swish activation function can provide an improvement in the detection speed of the model while ensuring its accuracy. It has better adaptability and detection capability on sensitive image detection tasks in real scenarios.

And in Scheme 5, this paper performs a further model pruning operation on the model to eliminate the layers with less accuracy improvement in the model, which greatly improves the detection speed of the model.

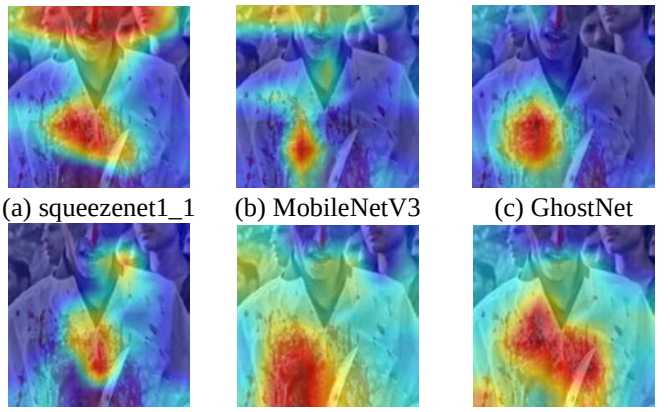
5.4 Visual Analysis of Class Activation Maps

To make it more intuitive to show the effect of the model, this paper uses the GradCAM [31] neural network visualization method to exhibit the actual prediction effect of the model.

The GradCAM method is able to generate a heat map of category activations. By combining this map with the original image, a two-dimensional grid can be obtained that is distributed over the entire image and associated with a specific output category. In each grid, the GradCAM method uses a different color to mark the model's focus on that region.

The heat map, as shown in **Fig. 5**, can be learned that for the same original bloody image, except for GhostNet, the baseline network EfficientNet selected in this paper, the other models focus on the region that does not involve blood in the sensitive image detection task.

However, GhostNet only focuses on a small portion of the pixels in the image, and the localization is not accurate enough. Although EfficientNet can also focus on most of the blood-related regions and concentrate on the blood-stained areas of the clothes in the image, it is accompanied by some background interference information.



(a) squeezenet1_1 (b) MobileNetV3 (c) GhostNet

(d) ShuffleNetV2 (e)EfficientNet_b0 (f) Proposed model

Fig. 5. Results on visualization of attention for each model using the GradCAM method

The GhostEfficientNet_s model can accurately locate blood targets in the image, while having less attention to other regions that are not bloody. It can be seen that the SE-Ghost module proposed in this paper can extract the features of regions containing sensitive objects better and can effectively reduce the interference from background regions. Therefore, this paper's method of incorporating this structure to optimize the feature extraction capability of the model is very effective, and it enables the model to have better accuracy in the detection task of sensitive information.

6. Conclusion

A multi-classification sensitive image detection method based on lightweight CNN is proposed in this paper. In terms of model size, GhostEfficientNet_s uses the lightweight EfficientNet as the baseline model for modification and innovation. EfficientNet greatly reduces the number of parameters of the model through NAS search parameters, so that both the speed and accuracy of the model can be taken into the account in the use of the model. At the same time, the model is pruned layer by layer to reduce the layers that contribute less to the accuracy of the model, and the model parameters and calculation burden are greatly reduced. In terms of the accuracy of the model, GhostEfficientNet_s is enhanced by introducing Ghost Module that incorporates the attention mechanism of the SE channel for feature extraction, so that the model can better reduce the feature loss of the model on sensitive images. In terms of the detection speed of the model, this paper improves the detection efficiency of the model for sensitive images by using model pruning operation to reduce the model parameters and introducing the Hard-Swish activation function for performance optimization.

By comparing several lightweight models to each other, the experimental results demonstrate that the model in this paper is able to obtain excellent detection speed in the task of detecting sensitive images with higher accuracy than similar lightweight network models. GhostEfficientNet_s is mainly applied to the detection task of sensitive images. Through the improvement of model accuracy and detection speed, it can more adapt to the current chaotic and complex network environment. In addition, due to the particularity of sensitive images, the number of the dataset in this paper is still insufficient, so it may not achieve the training effect in the actual production and application. Therefore, the following work of this paper will continue to improve and adjust the model for the problem of data enhancement to better adapt to the actual production and application, so as to achieve better use effect.

References

[1] Liu Y, Xiao C, Wang Z, et al., "An one-class classification approach to detecting porn image," in *Proc. of the 27th Conference on Image and Vision Computing New Zealand*, pp. 284-289, 2012. Article (CrossRef Link)

[2] Ulges A, Stahl A, "Automatic detection of child pornography using color visual words," in *Proc. of the 2011 IEEE international conference on multimedia and expo*, pp. 1-6, 2011. Article (CrossRef Link)

[3] Yang J, Fu Z, Tan T, et al., "A novel approach to detecting adult images," in *Proc. of the 17th International Conference on Pattern Recognition (ICPR)*, pp. 479-482, 2004. Article (CrossRef Link)

[4] Balouchian P, Safaei M, Foroosh H, "LUCFER: A large-scale context-sensitive image dataset for deep learning of visual emotions," in *Proc. of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1645-1654, 2019. Article (CrossRef Link)

[5] Tran K, Nguyen D, Nguyen H, "Toward Deep and Handcraft Features for Detecting Violent Behaviors," in *Proc. of the 2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, pp. 422-427, 2019. Article (CrossRef Link)

[6] Xu W, Parvin H, Izadparast H, "Deep learning neural network for unconventional images classification," *Neural Processing Letters*, val. 52, no 1, pp. 169-185, 2020. Article (CrossRef Link)

[7] Han K, Wang Y, Tian Q, et al., "Ghostnet: More features from cheap operations," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1580-1589, 2020. Article (CrossRef Link)

[8] Hu J, Shen L, Sun G, "Squeeze-and-excitation networks," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 7132-7141, 2018. Article (CrossRef Link)

[9] Tan M, Le Q, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. of the Machine Learning Research (PMLR)*, pp. 6105-6114, 2019. Article (CrossRef Link)

[10] Howard A, Sandler M, Chu G, et al., "Searching for mobilenetv3," in *Proc. of the IEEE international conference on computer vision (ICCV)*, pp. 1314-1324, 2019. Article (CrossRef Link)

[11] Haiming Yin, Xiaodong Xu, Lihua Ye, "Big Skin Regions Detection for Adult Image Identification," in *Proc. of the Workshop on Digital Media & Digital Content Management*, pp. 242-247, 2011. Article (CrossRef Link)

[12] Basilio J A M, Torres G A, Gabriel Sánchez Pérez, et al., "Explicit image detection using YCbCr space color model as skin detection," in *Proc. of the Applications of Mathematics and Computer Engineering*, pp. 123-128, 2011. Article (CrossRef Link)

[13] Nuraisha S, Pratama F I, Budianita A, et al., "Implementation of K-NN based on histogram at image recognition for pornography detection," in *Proc. of the 2017 International Seminar on Application for Technology of Information and Communication*, pp. 5-10, 2017. Article (CrossRef Link)

[14] Lv L, Zhao C, Lv H, et al., "Pornographic images detection using high-level semantic features," in *Proc. of the 2011 Seventh International Conference on Natural Computation*, pp. 1015-1018, 2011. Article (CrossRef Link)

[15] Lopes A P B, de Avila S E F, Peixoto A N A, et al., "A bag-of-features approach based on hue-sift descriptor for nude detection," in *Proc. of the 17th European Signal Processing Conference*, pp. 1552-1556, 2009. Article (CrossRef Link)

[16] Won D, Steinert-Threlkeld Z C, Joo J, "Protest activity detection and perceived violence estimation from social media images," in *Proc. of the 25th ACM international conference on Multimedia*, pp. 786-794, 2017. Article (CrossRef Link)

[17] Connie T, Al-Shabi M, Goh M, "Smart content recognition from images using a mixture of convolutional neural networks," *IT Convergence and Security 2017*, Singapore: Springer Singapore, vol. 1, pp. 11-18, 2018. Article (CrossRef Link)

[18] X. Lin, F. Qin, Y. Peng, Y. Shao, "Fine-grained pornographic image recognition with multiple feature fusion transfer learning," in *Proc. of the International Journal of Machine Learning and Cybernetics*, vol. 12, pp. 73-86, 2021. Article (CrossRef Link)
[19] Huang G, Liu Z, Van Der Maaten L, et al. "Densely connected convolutional networks," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4700-4708, 2018. Article (CrossRef Link)
[20] Surinta O, Khamket T, "Recognizing Pornographic Images using Deep Convolutional Neural Networks," in *Proc. of the 4th International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*, pp. 150-154, 2019. Article (CrossRef Link)
[21] Iandola F N, Han S, Moskewicz M W, et al., "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," *arXiv paper*, 2016. Article (CrossRef Link)
[22] Zhang X, Zhou X, Lin M, et al., "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 6848-6856, 2018. Article (CrossRef Link)
[23] Ma N, Zhang X, Zheng H T, et al., "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proc. of the European conference on computer vision (ECCV)*, pp. 116-131 2018. Article (CrossRef Link)
[24] Howard A G, Zhu M, Chen B, et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv paper*, 2017. Article (CrossRef Link)
[25] Sandler M, Howard A, Zhu M, et al., "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4510-4520, 2018. Article (CrossRef Link)
[26] Zoph B, Le Q V, "Neural architecture search with reinforcement learning," *arXiv paper*, 2016. Article (CrossRef Link)
[27] Ramachandran P, Zoph B, Le Q V, "Searching for activation functions," *arXiv paper*, 2017. Article (CrossRef Link)
[28] Szegedy C, Liu W, Jia Y, et al., "Going deeper with convolutions," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1-9, 2015. Article (CrossRef Link)
[29] Loshchilov I, Hutter F, "Decoupled weight decay regularization," *arXiv paper*, 2017. Article (CrossRef Link)
[30] Bottou L, "Stochastic gradient descent tricks," *Neural networks: Tricks of the trade. Springer, Berlin, Heidelberg*, pp. 421-436, 2012. Article (CrossRef Link)
[31] Selvaraju R R, Cogswell M, Das A, et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. of the IEEE international conference on computer vision (ICCV)*, pp. 618-626, 2017. Article (CrossRef Link)

Yueheng Mao received the B.S. degree in Internet of Things Engineering from Henan University of Science and Technology, Luoyang, China, in 2020. He is currently pursuing the M.Eng. degree in computer technology at Henan University of Science and Technology, Luoyang, China. Her research interest computer vision technology, big data processing.

Bin Song received the B.Eng., M.Eng., and Ph.D. degree in computer science from Zhengzhou University, China University of Geosciences, and Korea University, respectively. He is currently a Research Fellow with the Henan International Joint Laboratory of Cyberspace Security Applications. He is also a Associate Professor with the Henan University of Science and Technology. His research interests include artificial intelligence, image processing, and computer vision for security applications.

Zhiyong Zhang (Senior Member, IEEE), received received his Master, Ph.D. degrees in Computer Science from Dalian University of Technology and Xidian University, P. R. China, respectively. He was ever post-doctoral fellowship at School of Management, Xi'an Jiaotong University, China. Nowadays, he is Director of Henan International Joint Laboratory of Cyberspace Security Applications, Vice-Dean of College of Information Engineering, and full-time Henan Province Distinguished Professor at Henan University of Science and Technology, China. He is also a visiting professor of Computer Science Department of Iowa State University. His research interests include cyber security and frontier computing, human-centric CPS security, intelligent manufacture and industrial internet security. Recent years, he has published over 150 scientific papers and edited 6 books in the above research fields, and also holds 20 authorized patents. He is Chair of IEEE MMTC DRMIG, IEEE Systems, Man, Cybernetics Society Technical Committee on Soft Computing, World Federation on Soft Computing Young Researchers Committee, Committeeman of China National Audio, Video, Multimedia System and Device Standardization Technologies Committee. And also, he is editorial board member and associate editor of Multimedia Tools and Applications (Springer), Human-centric Computing and Information Sciences (Springer), IEEE Access (IEEE), Neural Network World, EURASIP Journal on Information Security (Springer), leading guest editor or co-guest Editor of Applied Soft Computing (Elsevier), Computer Journal (Oxford) and Future Generation Computer Systems (Elsevier). And also, he is Chair/Co-Chair and TPC Member for numerous international conferences/ workshops on cyber security, privacy and frontier computing.

Wenhou Yang received his MBA from Tsinghua University. Currently, he is the CEO of Sunnetech Technology Co.,Ltd. (Quzhou).

Yu Lan received his M.S. in Computer Science and Ph.D. from the Massachusetts Institute of Technology (MIT). He is currently the Technical Director and Principal Consultant of Sunnetech Technology Co.,Ltd. (Quzhou).