International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-020-01157-9

ORIGINAL ARTICLE

Fine-grained pornographic image recognition with multiple feature fusion transfer learning

Xinnan Lin[1] · Feiwei Qin[1] · Yong Peng[1] · Yanli Shao[1]

Received: 27 May 2019 / Accepted: 11 June 2020

Abstract
Image has become a main medium of Internet information dissemination, makes it easy for an Internet visitor to get pornographic images with just few clicks on websites. It is necessary to build pornographic image recognition systems since uncontrolled spreading of adult content could be harm to the adolescents. Previous solutions for pornographic image recognition are usually based on hand-crafted features like human skin color. Hand-crafted feature based methods are straightforward to understand and use but limited in specific situations. In this paper, we propose a deep learning based approach with multiple feature fusion transfer learning strategy. Firstly, we obtain the training data from an open data set called NSFW with 120,000+ images. Images would be classified into different levels according to its content sensitivity. Then we employ data augment methods, train a deep convolutional neural network to extract image features and conduct the classification job, without the need for hand-crafted rules. A pre-trained model is used to initialize the network and help extract the basic features. Furthermore, we propose a fusion method that makes use of multiple transfer learning models in inference, to improve the accuracy on the test set. The experimental results prove that our method achieves high accuracy on the pornographic image recognition and inspection task.

Keywords Pornographic image recognition · Image classification · Multiple feature fusion · Transfer learning

1 Introduction

With the rapid development of Internet and social networks in twenty-first century, the amount of information available online has become overwhelming. Nowadays, image has become a major carrier of data on Internet, disseminates colorful information to our life. However, lack of image sources supervision also makes pornographic images much easier to acquire [1]. It is no secret that people can obtain pornographic images with just a few clicks on websites, which leads to the uncontrollable spread of pornographic images. Internet-enabled devices have allowed people of all ages to encounter sexually explicit contents, which could be harmful to adolescents [2]. Moreover, for adults, watching online porn frequently could reduce working effectiveness, lead to higher divorce rate [3] and even cause social problems [4]. Therefore, building a healthy environment with pornographic image recognition methods for Internet visitor is a vital job.

Researchers have worked on pornographic image recognition task for the past decades. Traditional, recognition is achieved by analyzing the extracted hand-crafted features (most commonly, human skin features) [4]. However, hand-crafted features are always too complicated to build, or their performance could be constrained in specific situations. In recent years, deep convolutional neural networks (CNNs) has shown great power on image analyzing tasks, especially image classification [5–8], deep CNN based methods have achieved high accuracy and robustness on these tasks.

Taking advantage of feature extraction ability of CNNs, researchers have proposed new pornographic image recognition methods [9–13], but we think the abilities to abstract image features of deep CNNs could be better utilized. In this paper, we consider fusing different deep neural network models together to propose a fine-grained pornographic image recognition approach via fusing multiple learned features. Our work can be summarized as follows:

✉ Feiwei Qin
qinfeiwei@hdu.edu.cn

[1] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China

Published online: 06 July 2020

1. Build deep convolutional neural networks based on well-performed network structures, and design a transfer learning strategy to improve the representation power of the extracted features.
2. Train our network with a 5 classes public pornographic image dataset. Others only divided their dataset into 2 classes, porn and non-porn. On the contrary, we tend to train a model that not only distinguishes whether an image is porn or not, but also finely classifies pornographic images according to its content sensitivity.
3. Most deep learning based methods train only a single network to extract features from input images. To further utilize the feature representation ability of CNNs, we propose a method to fuse extracted features from multiple trained models with similar topological structure to further improve the classification accuracy.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 elaborates the dataset used in our work, details of the network structure and the proposed method. Section 4 demonstrates the experiment results and gives discussion. Lastly, conclusion and future work are provided in Sect. 5.

2 Related work

2.1 Hand-crafted feature based methods

During the past decades, researchers have proposed many approaches to recognize pornographic images. Basically, pornographic images recognition task is a kind of image classification task in essence, traditional solutions for this task are usually based on hand-crafted features, such as skin color based feature.

Due to a great bunch of pornographic images contain large area of exposed human skin, it is a straightforward feature. Zhu et al. [14] compare the human skin color and texture detection effect in different color space, use Support Vector Machine (SVM) to classify pornographic images with skin features. Srisaan et al. [15] combine human skin and face detector, set a series of rules like the proportion of skin and face area to recognize pornographic images. Moreira et al. [16] make use of the skin detection results in both RGB and YCbrCr color space, combine with face detection method to extract Region of Interest (ROI) of the images, by figuring the relationships between ROIs to classify bikini and pornographic images.

In addition to skin color based methods, Bag of Visual Words (BoVW) model is another kind of widely used hand-crafted feature based pornographic recognition method. Deselaers et al. [17] extract the scale-invariant feature transform (SIFT) features from images in training set and use k-means algorithm to build a 100 clusters feature vocabulary, then train a SVM to classify pornographic images. Researchers have also proposed many other similar BoVW model based methods to solve pornographic image recognition task. Avila et al. [18] use HueSIFT features to build feature vocabulary while Zhuo et al. [19] make use of Oriented fast and Rotated Brief (ORB) features.

There are still a lot of approaches based on other hand-crafted features [20–23] with good results. However, hand-crafted features are always too complicated to build and their performance may be constrained in specific situations. For example, skin color based classifier will make mistakes when processing pornographic images with few exposed skin areas. BoVW model would ignore the relationship between the features.

2.2 Deep learning based methods

Since AlexNet [5] has achieved great success on ImageNet competition, deep convolutional neural networks (CNNs) have shown its great effect on image analysis problems such as image classification, object detection, semantic segmentation etc [24]. In image classification field, VGGNet [25], GoogLeNet [26], ResNet [7], DenseNet [8] are the representative network architectures which could be used as the backbone of pornographic image recognition tasks. Compared to the hand-crafted feature descriptors, deep learning based feature descriptors are much easier to build, and machine learning can automatically extract features from the training set. With the help of great amount of training images, the trained classifier will be robust enough and achieve high accuracy.

There have been some studies on recognize pornographic images by employing CNNs [9–13], Most of these approaches divide pornographic images into only 2 classes, pornographic and non-pornographic. But we found that CNN can do the job better by dividing the images obtained in the training set into 5 classes. In this way, the approach proposed in this paper can finely recognize images according to its content sensitivity. Furthermore, in the inference time we take the advantage of different feature descriptors by fusing multiple transfer learning models, which achieves higher accuracy than the approach with a single CNN model.

3 Approach

3.1 Overview

The architecture of our approach can be divided into 3 modules, as shown in Fig. 1. From left to right the first module is image feature descriptor, a pretrained model of DenseNet121 [8] trained on ImageNet task is used as the

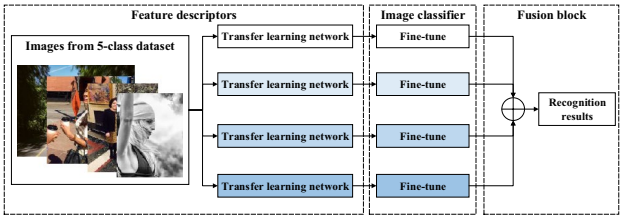

Fig. 1 Overview of the proposed approach

backbone of our network structure. Benefited from pre-trained model, the network gains the ability to extract basic features from images before training. By freezing different denseblock layers, we train 4 transfer learning models with different representation power aiming for diverse features. The second module is an image classifier, we add 4 fully connected layers as the fine-tuning classifier after each of the 4 feature descriptor. By training the entire network using pornographic recognition data set, the feature descriptors will focus on features related to pornographic recognition task and the classifier will gain the ability to classify them. The last module is a fusion block, we have trained 4 models in the former parts and here we fuse their outputs together to make use of different representation power of each model. By taking advantage of multiple transfer learning models, the proposed method can achieve higher classification accuracy than the single model based methods.

3.2 Data preparation

3.2.1 NPDI dataset

NPDI dataset is a well-known dataset for pornographic video and image recognition tasks, collected by Avila et al. [18]. The dataset contains nearly 80 h of 400 pornographic and 400 non-pornographic videos. For the pornographic category, videos are collected from the websites only host such materials. For the non-pornographic category, it can be divided into simple and difficult categories, each category consists of 200 videos chosen from porn video websites. As it is often done in video analysis, a key frame is selected to summarize the shot content into a static image, NPDI dataset also selects the middle frame of each video shot and 16,727 video segments are collected in total. In this task, we focus on recognizing pornographic images so only the video segments are used. We divide the 16,727 key frames into two categories named porn and non-porn, our proposed method is tested on this dataset and compared with other methods to show the efficiency of our method on 2-class dataset.

3.2.2 NSFW dataset

Because of the sensitivity, commonly known pornographic image dataset is relatively small and composed of only 2 classes, porn and non-porn, while NSFW dataset [27] is a public dataset with more than 200,000 images divided into 5 classes: drawings, hentai, neutral, porn and sexy. Examples of the images in NSFW dataset are shown in Fig. 2, the amount of raw images for different classes is highly imbalance, more than 100,000 images are labeled as porn image while the other 4 classes contain only 20,000–40,000 for each. To avoid training a model with great tendency to regard most images as porn, about 70,000 of porn images are randomly discarded from the original dataset to balance the number of instances between 5 classes, the details of the balanced training set is shown in Table 1. After that we randomly select 600 images of each class to build a validation set and 3000 of each class for test set. Then a training set with 120,000+ images, a validation set with 3000 images and a test set with 10,000 images are generated respectively.

3.2.3 Data augmentation

Compared to other common data sets (e.g. ImageNet) in the field of image classification, NSFW data set is relatively small. Since small data sets are prone to reduce the model effect, to maintain the representation power of model and avoid overfitting in some ways, several data augmentation methods are employed. Different from common image labels, pornographic image label is defined by some small parts of images containing sensitive organs such as female breast, male genitals and female genitals. A pornographic image would change into a non-pornographic one if these parts are removed. In other words, cropped pornographic images containing no sensitive organs should be labeled as non-pornographic [11]. Therefore, we don't augment data with crop, translate or scale

Table 1 Details of training data set

Class name	Image amount	Content
Drawings	20,000+	Normal animation, cartoon and cosplay photos
Hentai	30,000+	Animation and cartoon with sensitive content
Neutral	20,000+	Normal photos
Porn	30,000+	Photos with sex behaviors
Sexy	20,000+	Photos with exposed sensitive organ

Fig. 2 Examples of 5 classes in the dataset

operations to avoid changing the original label of images. Instead we randomly rotate and shear the original images between – 10° and 10°, then flip them horizontally or vertically and fill the margin with white pixels. Lastly, to fit our network input, the original images are then resized to 299×299.

3.3 Network construction

Hand-crafted features have been widely used in image classification task, including pornographic image recognition task. Though these hand-crafted feature based methods perform well, they are always too complicated to build and could be constrained in specific situations. With the development of deep learning techniques, deep convolutional neural networks have achieved great success in image analysis task, especially in image classification task because of its ability to extract intricate features from raw data. By training a model with a huge amount of images, the learned parameters network could gain greater representation power and perform better than hand-crafted feature based methods. Hence we decide to build a pornographic images recognizer based on CNNs.

3.3.1 Single network structure

CNN based approaches on pornographic image recognition tasks have been proposed in recent years, most of them train a single neural network model to extract image features [9–13]. After AlexNet achieved great success in the ImageNet competition, the potential of deep CNNs are gradually discovered by researchers, many deep network architectures with well performance on image classification tasks such as VGGNet, GoogLeNet, RensNet, DenseNet and etc are proposed, network based on these well-performed architectures are naturally applied to pornographic image recognition task. Inspired by [28], we build a similar network based on DenseNet121 [8].

As shown in Fig. 3, the network is mainly made up of a feature extractor and an image classifier. For the feature extractor, we build it with a truncated DenseNet121 without the fully connected layers. The backbone of DenseNet121 consists of 4 dense blocks, each dense block is composed of several dense layers. The number of dense layers in each dense block increases while the dense blocks go deeper. The structure of a dense block is shown in Fig. 4.

Different from a normal convolutional neural network, each dense layer's output is fed to all subsequent dense

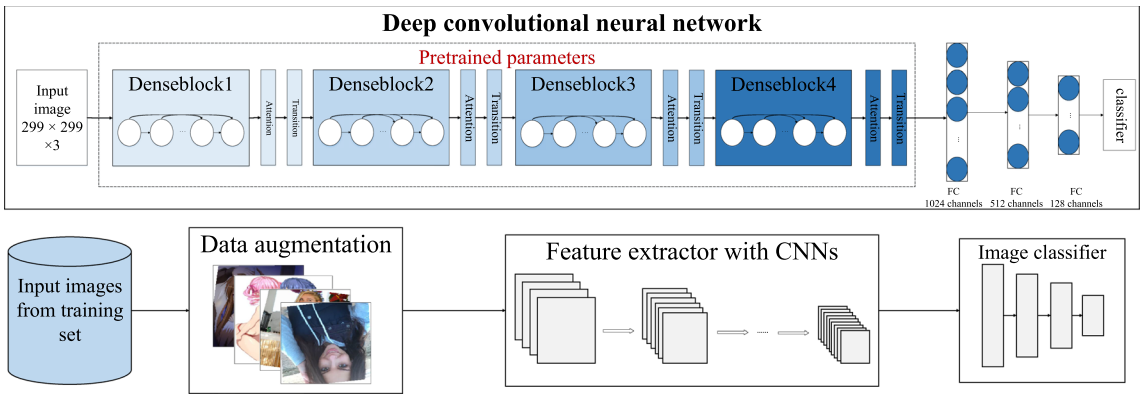

Fig. 3 CNN architecture for vulgar/pornographic image recognition task

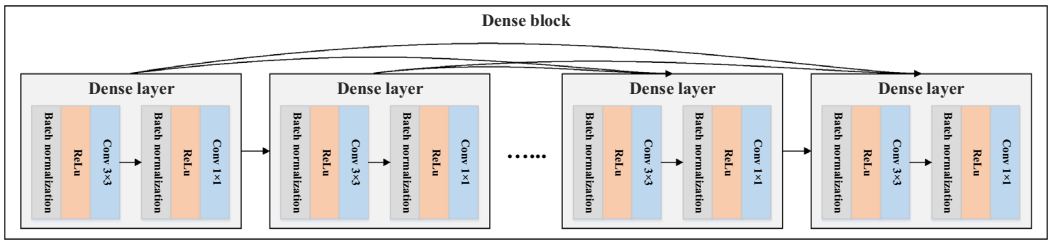

Fig. 4 Structure of the dense block and the dense layers inside

layers in the same block through the densely connections, which is implemented by concatenation operations. In this way, global information can traverse the dense block from beginning to end, each dense layer can acquire extra knowledge from the previous ones. For a dense layer, it includes 2 basic layers, each basic layer is composed of a batch normalization layer [29], a ReLU [30] activation function and 2 convolutional layers. The first convolutional layer with a 1×1 kernel is called bottleneck layer, and it helps to reduce the feature map dimension to improve calculation efficiency. The second layer applies a 3×3 kernel to extract features. Benefit from densely connections, the network can extract more representative features than the normal convolutional network with layers connected in sequence. In addition, an attention block and a transition block are inserted between every two dense blocks. The attention block helps the feature extractors locate the most informative part of the input feature map, details are described in Sect. 3.3.3. The transition block consists of a batch normalization layer and a 1×1 convolutional layer followed by 2×2 max pooling layer. It is set for down sampling the output feature maps of each dense block. While network goes deeper, the size of output feature maps decreases and the dimension of output feature maps increases. An example of the output feature maps after each dense block is shown in Fig. 5.

As for the image classifier, the 1000 dimensional fully connected layer located in the end of DenseNet121 is replaced by 3 new fully connected layers. It is set in this way because of the proposed transfer learning strategy, the details will be described in Sect. 3.3.2.

3.3.2 Transfer learning

Transfer learning aims to transfer knowledge between related source and target domains [31]. It is a useful tool in machine learning, leading to a positive effect on the domains that are difficult to apply because of insufficient training data. Transfer learning methods can be categorized into 4 classes: instances-based, mapping-based, network-based and adversarial-based deep transfer learning [32]. Network-based method is mostly used in convolutional neural networks. It is achieved by transferring the network structure and pretrained parameters of source domain into a part of deep convolutional neural network used in the target domain. In our work, we construct a feature extractor based on a well-performed network DenseNet121 as described in Sect. 3.3.1, then we initialize the network parameters with the parameters of a DenseNet121 model pre-trained on ImageNet dataset.

In general, there are two ways to perform network-based transfer learning: fine-tuning, which consists of using the pre-trained model on source dataset and training all layers in the target dataset; freezing, which consists of leaving the parameters in shallow part of the pretrained model unchanged and training only the rest part of the network, which can make use of the basic feature extracting ability of pretrained model. Freezing layers or not in training phase depends on the number of category and quantity variance between source and target dataset. Compared to ImageNet dataset, the NPDI and NSFW dataset we used are much smaller. Moreover, most images in both datasets are natural images, they are similar in a way, hence we freeze some

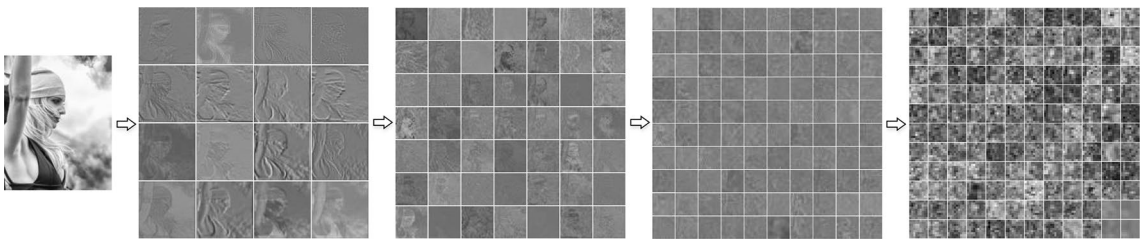

Fig. 5 Feature maps generated by dense blocks in the training phase

of the layers during training. In order to explore the effect of fusing different feature extractors, we try different freezing strategy on the networks, the details are described in Sect. 3.3.4.

In our experiments, an input image is resize to a 299×299×3 tensor, going through the aforementioned network initialized with pre-trained parameters and a 7×7 average pooling layer, output a 1×1 feature in 1024 dimensions. To compensate for the different image statistics of the source and target data, we add 3 fully connected layers (*FC1*, *FC2*, *FC3*) as the adaption layers [28] before the classifier layer. As shown in Fig. 6, *FC1* is composed of 1024 neurons, *FC2* is composed of 512 neurons and *FC3* is composed of 256 neurons. Each fully connected layer is followed by a dropout layer [33], which has proven to be an effective way to prevent overfitting. The dropout rate is set to 0.5, which means that only half of the neurons are connected while the other neurons' output will be discarded during the training phase. Different from the previous convolutional layers, these fully connected layers are initialized randomly and their parameters are non-fixed during the training phase.

3.3.3 Visual attention

Visual attention is a tool that helps feature extractors locate the most informative part of the input image. With the help of visual attention block, the network will concentrate more on the features that contribute most to the recognition results. We adopt Squeeze-and-Excitation (SE) blocks raised by [34] in our network. SE block is a visual attention block composed of 3 parts: squeeze, excitation and scale. Squeeze is implemented with an average pooling layer, in this way the global information for each channel of feature maps is summarized. Then the squeezed information will be fed to the excitation part consisting of 2 fully connected layers, one using sigmoid as activation function and the other using ReLU. After that we can get a series of weights, each channel of weights represents the importance of each input feature map, higher weight means that this feature map is prone to contribute more to the results. Finally, a channel-wise multiplication of excitation weights and original input feature maps are applied, interdependencies between channels are modeled to boost the representation power.

SE block is a flexible block whose size of input and output are the same, it can be embedded into most of CNNs easily. As shown in Fig. 3, we have inserted this visual attention block into our network, 4 attention blocks are inserted between each dense block and its following layer. It can be seen from Table 7 that it can help improve the model performance.

3.3.4 Multiple model fusion

Commonly, most CNN based methods on pornographic image classification task use a single network to extract features, and then feed them into a classifier. However, we found that AGNet [12] did this job in a different way. It tried to explore the potential power of fusion features extracted from different models. Two models are fused to predict the label of input image together, one is based on AlexNet and the other is based on GoogLeNet, that's why it is named AGNet. In the inference phase,the pornographic and non-pornographic image scores of the input image given by two models will be fed to a fusion block, a weighted average with equal weight is employed, final classification result of the input image depends on features extracted from both two models. In this way, AGNet takes advantage of different features with higher accuracy than the single network methods [12].

Inspired by AGNet, we propose an approach to improve the classification accuracy by fusing multiple models with the same structure but trained with different transfer learning strategies. As mentioned in Sect. 3.3.2, during the training phase, parameters in the shallow layers of the network can be fixed to take advantage of the basic feature extracting ability of the pre-trained model. Note that shallower layers tend to extract basic features, deeper layers tends to extract abstract features [35], we can conclude that models with different layers' parameters fixed during the training phase would focus on features with different complexity. At inference time, we fuse the trained models together to make use of the different feature extraction ability of each model.

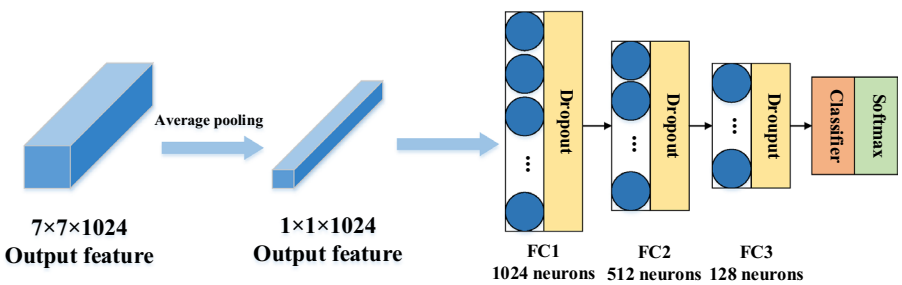

Fig. 6 Structure of the classifier with fully connected layers

Different from AGNet, we train 4 models with the same structure mentioned in Sect. 3.3.1, but freeze different layers of each model while training concurrently. There are two reasons why we select 4 same DenseNet121 based model to be fused. Firstly, we have tested the single model method for this task, among which DenseNet121 with attention block performs best. Secondly, we have tried fusing different trained models such as InceptionV3 [6], ResNet, DenseNet etc together with frozen parameters, the experimental results are demonstrated in Table 8. It can be seen that fusing 4 DenseNet based models is still the better way. We think the reason is that when fusing models in different topological structure, the relationship between features extracted from 4 models is relatively weak and we cannot assure that they are learning different knowledge from training images. Since the structures of different networks varies greatly, models may concentrate on similar features with different performance according to the depth of network. By training 4 similar structure models with different frozen layers, each model extracts features in different levels based on the number of frozen parameters, we can ensure that every model concentrates on different features of the input image. Meanwhile, in the experiment we found the false positive and false negative instances of each model are quite different, which proves that they are concentrating on different kinds of features. Since the 4 models keep a similar network structure, the relationship between features they extracted are stronger, and features can be fused better.

Figure 7 shows how we train 4 models in parallel. A single network consists of a feature extractor and a classifier as shown in Fig. 3. As mentioned in Sect. 3.3.1, the feature descriptor of the network is based on DenseNet121, which can be simply departed into 4 dense blocks. In our work, we train 4 models with parameters in different block frozen in parallel. Each row in Fig. 6 represents one model, the dense block in blue means that the parameters in this block are frozen during the training phase while the red block means that parameters are activated. After that the extracted features will be fed into the classifier composed of 3 fully connected layers, parameters in fully connected layers of all 4 models are initialized randomly and remain activated during the training phase. At inference time, an image is fed to all 4 models, scores of each category are given based on the feature they concentrate. Similar to [12], we fuse the scores of multiple transfer learning models with a weighted sum operation. The weights of 4 models are set manually in the experiments. Taking the advantage of features diversity, the transfer learning models fusion method achieved high accuracy in pornographic image recognition task. Details of experiments and results are described in Sect. 4.

4 Experiments

4.1 Training strategies

Our approach is implemented by the open source framework PyTorch [36]. The pre-trained models for initializing the network parameters are acquired from torchvision [37]. Models are trained and tested on a graphic workstation with an Intel Core i9-7980XE CPU(2.60 GHz, 18 core) and a GeForce GTX 1080Ti GPU. As described in Sect. 3.3, we choose a 2-category dataset NPDI and a 5-category dataset NSFW to train and test our proposed network in the experiments.

In the training phase, we use stochastic gradient decent method with a batch size of 32, a momentum of 0.9 and a weight decay of 0.0005 similar to [11] to optimize our models for 100 epochs. As shown in Fig. 7, we propose a method fusing 4 models with similar structure and freezing different layers called *Freeze1* to *Freeze4* from top to bottom. The accuracy curves on the validation dataset of 4 models in the training phase are shown in Fig. 8. Due to the warmup strategy [38], the accuracy grows slowly in the first 5 epochs. A relatively large batch size reduces the noise in the gradient, and an increased learning rate can make a larger progress along the opposite of the gradient direction [38]. Since our network structure is based on DenseNet121 and the batch size we choose is smaller than only one DenseNet121 neural network model, learning rate is initialized to 0.01 and lowered by 10 times every 20 epochs in the experiments.

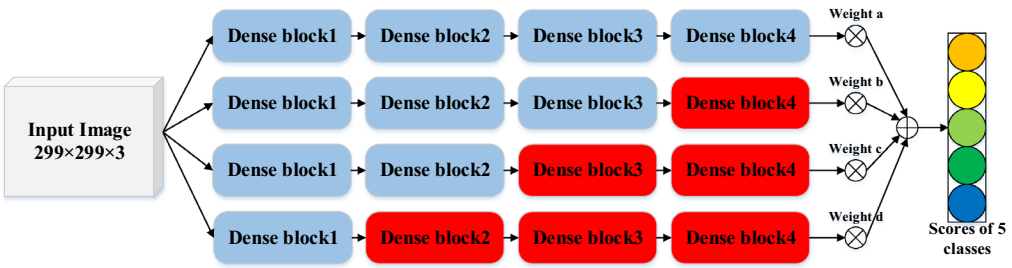

Fig. 7 Fusing multiple transfer learning models

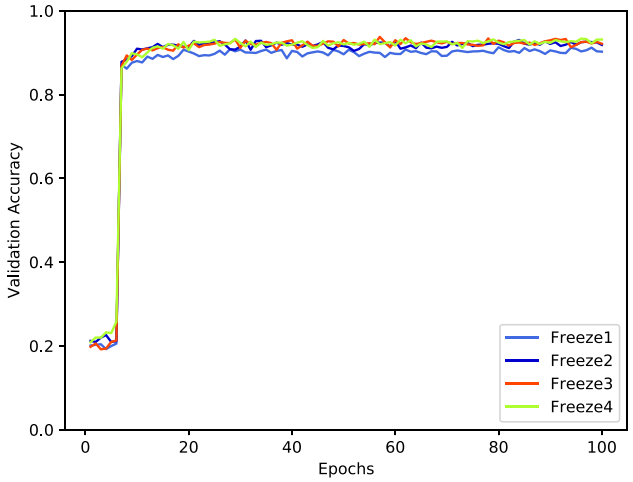

Fig. 8 Validation accuracy curves of 4 models with different frozen parameters during the training phase

Although the parameters of the network were initialized with the model trained by ImageNet at the beginning of the training phase, and the model is far from the final solution, we adopt a warmup strategy [38] by using a very small learning rate 0.00001 in the first 5 epochs and then switch to the initial learning rate when the training process is stable. Besides, for every 5 epochs, we compare the performance between the current model and the best model of previous epochs. If the previous one performs better, the parameters in the former model will be loaded into the current model before the next epoch starts.

4.2 Results on NPDI dataset

For pornographic image recognition, NPDI dataset is a widely used 2-class dataset which collects nearly 80 h of 400 pornographic and 400 non-pornographic videos. It selects several key frames of each video to represent the video content. There are 16,727 frames snapped in total, images are labeled according to their corresponding videos class. In the experiment, the NPDI dataset is divided into a training set with 13,382 images and a test set with 3345 images. Besides, we found noise in the porn category. Images labeled porn are snapshots of porn videos, but in fact not every frame of a porn video contains sexual behaviors or exposed sensitive organs. Although these images origin porn videos, they should not be labeled as porn images. Therefore, we relabeled 823 images of porn category as non-porn.

Table 2 Results of the proposed method on NPDI dataset

Label	Predict	
	Non-porn	*Porn*
Non-porn	2046	22
Porn	191	1087

Table 2 shows the results of our proposed method on the NPDI dataset. It can be inferred that our method performs better in recognizing non-porn images. Since the amount of non-porn images is quite larger than the amount of porn images in the training set and there are some features in common between some non-porn images and porn images such as large area of exposed female skin, network is more likely to label an input image as non-porn image.

We also compared the results on the NPDI dataset with other methods including hand-crafted feature based methods and CNN feature based methods. As shown in Table 3, the proposed method achieved high accuracy close to other methods without obvious advantages on NPDI dataset. Since the NPDI dataset is a 2-class dataset, many images with different details are simply grouped into non-porn and porn. We trained 4 models concentrating on different kind of features by freezing different parameters during training. This method performs better on a fine-grained classification task. We trained and tested the proposed method on NSFW dataset with 5 categories in Sect. 4.3. Our method achieved high accuracy on this task, significantly higher than other methods.

4.3 Results on NSFW dataset

The NSFW dataset contains more than 200,000 images divided into 5 classes called *drawings*, *hentai*, *neutral*, *porn* and *sexy*. After balancing the amount of each class as mentioned in Sect. 3.2, we obtain a train set with 120,000+ images, a validation set with 3000 images to validate the trained model in every epoch and a test set with 10,000 images. Table 1 shows the details of the training set. Different from training with a 2-class dataset, working on NSFW dataset is a fine-grained classification task, and the network should pay more attention to the tiny differences in image features.

Table 3 Performance comparing to different methods on NPDI dataset

Method	Accuracy (%)
SIFT-BoVW [18]	89.5
Binboost16-BoVW [39]	90.9
VGGNet [40]	93.8
AGNet [12]	93.8
Inception [13]	93.7
DenseNet	93.9
DenseNet-attention	94.1
Our	94.3

4.3.1 Statistical analysis of fusion method

In the previous sections we have proposed a method which fuses 4 models with different layers' parameters frozen to improve performance, particularly in Sect. 3.3.4 we mention that different models concentrates on different aspects of input images. To evaluate the diversity of 4 models, we conduct statistical analysis with series of ablation studies.

First we try to fuse any 2 models of 4 models with different frozen parameters and calculate their accuracies on test set comparing to a single model. Details of the results are listed in Table 4. *F1–F4* represent model *Freeze1* to *Freeze4* respectively and & means fusing two models together, for instance, *F1&F2* means fusing model *Freeze1* and *Freeze2*. According to the results, we find that fusing any of 2 models always performs better than a single one for each category. It can be concluded that models with different layers' parameters frozen focus on different features of images in different categories, which leads to the fused network model get better performance.

In addition, when testing each single network's performance, we count the numbers of instances that only one model recognize them wrongly while the other three predict correctly. As shown in columns from 3 to 6 in Table 5, 4 models don't give the same predict result of every image in each category. For each category, there are dozens of images which recognized wrongly by only one model, and the other 3 models can recognize them better. And It can be observed that none of images are wrongly recognized by all the 4 models at the same time, from the last column in Table 5. In other words, the 4 models pay attention to different aspects of the vulgar/pornographic images, and each model has its advantage and drawback. By integrating them and fusing their extracted diversified features together, the ensemble model learns more representative features.

4.3.2 Comparing to single network methods

There have been approaches with single network structures applied on pornographic image recognition task [9–13], well-performed structures like VGGNet, Inception, ResNet and etc are used as the backbones for constructing the networks of pornographic image recognition approaches. To figure out which backbone is most suitable for this task, the experiments are conducted to test the performance of network with different backbones. Figure 9 shows the accuracy of several backbones we tried on validation set during 100 training epochs. In the first 5 epochs, the accuracy grows slowly because of the warmup strategy mentioned in Sect. 4.1, and the learning rate is set to a relatively small value before the training process is stable. We found the accuracy of DenseNet121 is slightly higher than other models in the experiment. Hence it is used as the backbone of our network.

Table 4 Results of fusing any 2 models

Category	Model									
	F1 (%)	*F2* (%)	*F3* (%)	*F4* (%)	*F1&F2* (%)	*F1&F3* (%)	*F1&F4* (%)	*F2&F3* (%)	*F2&F4* (%)	*F3&F4* (%)
Drawings	92.85	92.66	93.98	91.19	94.76	95.35	94.32	94.22	93.73	94.27
Hentai	97.10	96.13	95.78	96.95	96.84	97.25	97.86	96.69	97.31	97.10
Neutral	92.24	93.02	91.07	93.51	93.95	93.02	95.12	92.87	94.73	94
Porn	92.91	92.76	93.12	94.21	94.21	94.31	94.88	93.59	94.88	94.98
Sexy	88.31	88.56	89.36	90.31	89.91	90.95	92.01	90.81	91.15	91.65
Total	92.66	92.61	92.64	93.20	93.92	94.16	94.82	93.62	94.34	94.38

Table 5 Comparison of number of images recognized wrongly by only one model

Category	Total number	Wrong number only recognized by *Freeze1*	Wrong number only recognized by *Freeze2*	Wrong number only recognized by *Freeze3*	Wrong number only recognized by *Freeze4*	Wrong number recognized by all 4 models
Drawings	2044	58	44	31	67	0
Hentai	1968	24	24	29	36	0
Neutral	2050	68	36	61	60	0
Porn	1935	61	38	41	44	0
Sexy	2002	76	48	56	87	0
Total	9999	287	190	218	294	0

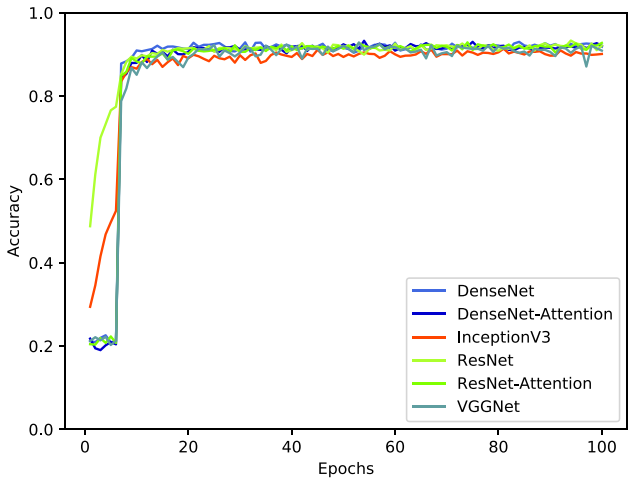

Fig. 9 Accuracy curves of models with different backbone

The proposed transfer learning model fusion method is shown in Fig. 7. 4 models with similar network structure based on DenseNet121 are trained in parallel with different frozen parameters. Each row in Fig. 7 represents a model, dense block in blue means that the parameters in the block are frozen in the training phase while the red one means that the parameters in the block are activated. Fully connected layers of all 4 models are initialized randomly and activated in the training phase. A weighted sum operation is adopted to fuse the scores of 5 classes produced by 4 different transfer learning models. The weight of each model is set manually according to the singles model's accuracy. The accuracy of each single model with different frozen parameters is shown in Table 6. For instance, model *Freeze1* represents the top row in Fig. 7, which is the model with most frozen parameters. In our work, we test different weight for each model according to their accuracies on the validation set, as shown in Fig. 8. Lastly, the weights for 4 models are set to [0.2, 0.2, 0.3, 0.3].

Because the four models (*Freeze1, 2, 3, 4*) are based on the same basic model 'Densnet121-Attention', they share similar topological network structure. That's why their classification accuracies on the test set are very close. However although their classification accuracies are close from Table 6, the wrongly recognized samples from the test set for each model are different, as shown in Fig. 10. For example, a

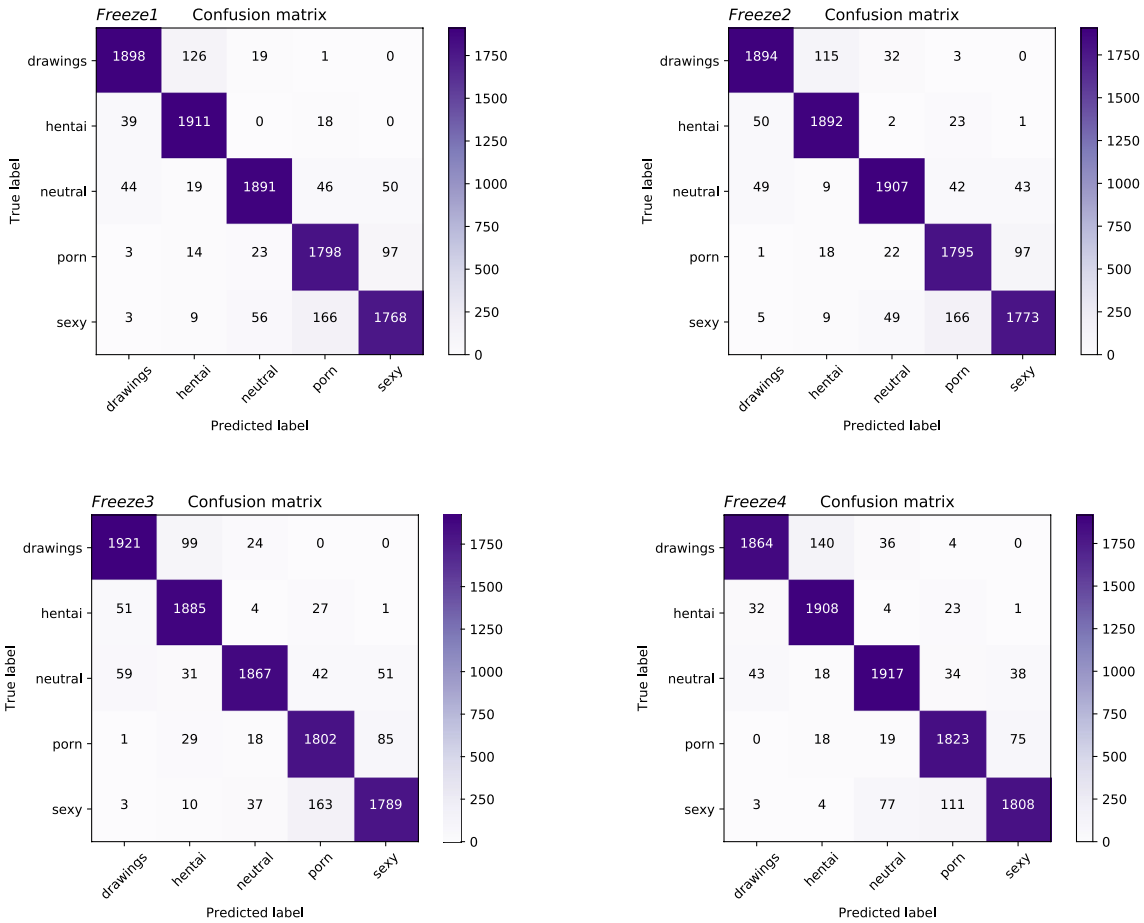

Fig. 10 Confusion matrices produced by the 4 differently frozen models on the test set

Table 6 Accuracies of models with different parameters frozen

Models	Freezed params	Test accuracy (%)
Freeze1	5.91 M	93.05
Freeze2	4.99 M	92.97
Freeze3	2.15 M	92.86
Freeze4	9.5K	93.02
Fusion models	–	94.96

'neutral' labeled image, as shown in Fig. 5, could be wrongly recognized as a 'porn' or 'sexy' image by *Freeze1* and *3*, but recognized correctly by *Freeze2* and *4* at the same time. The extracted features by *Freeze1* pay more attention to some aspect of the original vulgar images, while the extracted features by *Freeze2* pay more attention to other aspect of the original vulgar images. By fusing models together, we can make better use of the multiple feature extractors and get better results than using only a single network. Furthermore, we test other single network based methods on this task. The comparison results are given in Table 7. Our method achieves a accuracy of 94.96%, which is higher than the other single network based methods.

4.3.3 Comparing to different network fusion methods

We have proposed a method of fusing 4 models with same network structure but different frozen parameters. This idea is inspired by AGNet, a method enriching extracted features by combining 2 networks with a weighted sum. However, different from AGNet which fuses 2 trained models in different network structure, our proposed network uses the same. Due to the different topological structure, the relationship between features they extracted is relatively weak and we cannot assure that they have learned knowledge of different aspects from the training images.

In the experiments, we test the methods of fusing models in different network structures and the methods of fusing models in same network structures on NSFW dataset. As shown in Fig. 8, among the 4 models we trained, *Freeze4* achieves the best result on validation set during the training phase. Therefore, we try to fuse it with other models in different network structure such as VGGNet, Inception and ResNet to check whether the fusion models in different network structure helps improve the performance or not. The parameters of these model are initialized by the corresponding pre-trained model on ImageNet and their shallow layers are frozen while training. According to the results listed in Table 8, compared to the single network based method, it has higher accuracy, but fusing 4 DenseNet121 based models still performs better. By training 4 models in same network structure with frozen layers in different depth, we can ensure that every model concentrates on features of the input image in different levels. In this way, the network extracts more rich features and the classifier produces better recognition results.

Table 7 Comparing with the single network based methods on NSFW dataset

Method	Accuracy (%)	Average inference time of each image on test set (ms)
VGGNet based [40]	90.90	5.896
InceptionV3 based [13]	92.15	2.273
ResNet50 based	92.10	3.034
ResNet50-Attention based	92.75	3.418
DenseNet121 based	92.64	3.169
DenseNet121-Attention based	93.02	3.499
Our	94.96	12.575

Figure 11 shows the classification results of each class with confusion matrix. The accuracies of 4 classes are higher than 95%. From the confusion matrixes we can intuitively find that misclassification mainly occurs between class *porn* and *sexy*. False recognition instances are shown in Table 9 to help analyze deficiency of our work. They can be mainly summarized into 2 types of faults, the normal *drawings* are wrongly predicted as *hentai* images and the *neutral* images or *sexy* images are wrongly predicted as *porn* images. Because the features are learned from the training set and there are some overlapping image parts between classes, which contain a large area of exposed human skin, these images are more likely be wrongly classified.

5 Conclusion

In this paper, we construct a deep convolutional neural network for pornographic images recognition task. Different from the data sets consisting of only pornographic and non-pornographic images, we design our model for a data set of 5 classes which are divided by the sensitivity of image content, intending to propose an image recognition approach like a rating system and get good results. We propose a multiple model fusion transfer learning method to further improve the classification accuracy. Taking advantage of multiple models' representation power, we achieve higher accuracy than merely a single model could reach. In the experiments, we find that classifying images based on CNN relies heavily on

Table 8 Comparing performance of fusing models with different network structure

Fusing models	Accuracy (%)	Average inference time of each image on test set (ms)
VGGNet+DenseNet	93.96	9.172
Inception+DenseNet	94.14	5.309
Resnet+DenseNet	94.44	6.182
VGGNet+Inception+ResNet	94.21	11.204
VGGNet+Inception+ResNet+DenseNet	94.56	14.352
Our	94.96	12.358

Table 9 Instances of false recognition results for each category

Image	Label	Scores				
		drawings	*hentai*	*neutral*	*porn*	*sexy*
	*drawings*	10.32%	89.67%	0%	0.001%	0%
	*drawings*	44.53%	55.41%	0.009%	0.036%	0.005%
	*sexy*	0.006%	0.053%	0.021%	71.73%	28.18%
	*sexy*	0%	0%	0.023%	99.38%	0.59%
	*neutral*	0.006%	0.02%	0.61%	1.01%	98.34%

the quantity and breadth of the training data set. Our network is still not efficient enough to cover every situation of Internet vulgar image inspection. In the future, we plan to collect a larger data set with more detailed labels and build a more effective architecture to further improve the performance of the pornographic image recognition.

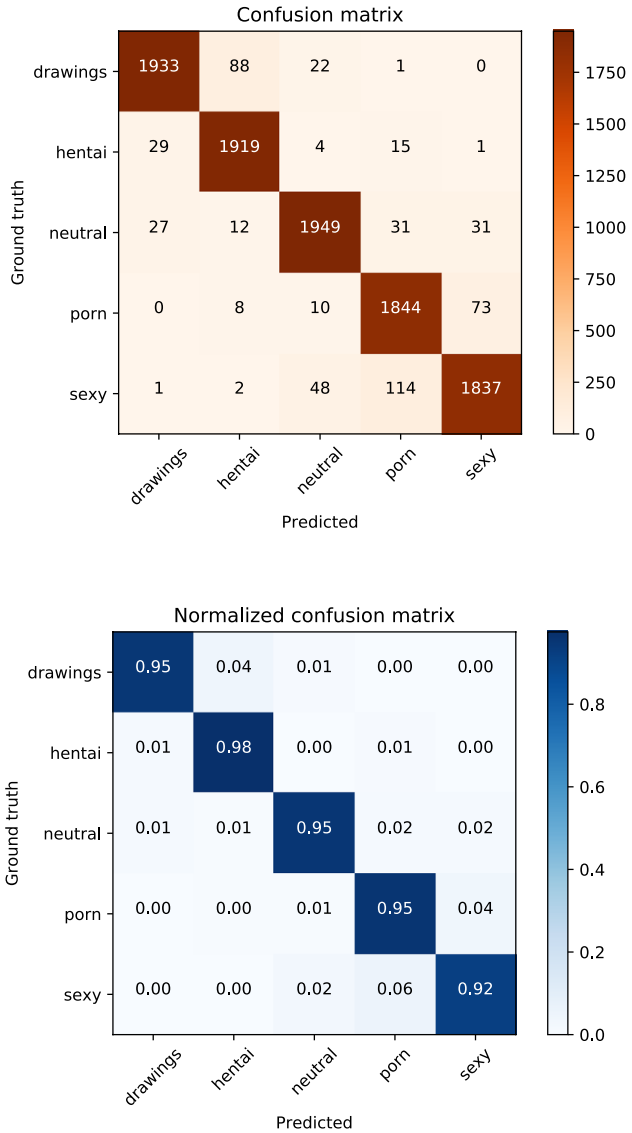

Fig. 11 Confusion matrices produced by the fusion model on the test set

Acknowledgements This work was supported in part by Key Research & Development Program of Zhejiang Province (No.2019C03127), National Natural Science Foundation of China (Nos. 61972121, 61602140, 61702517), and the open fund of Engineering Research Center of Cognitive Healthcare of Zhejiang Province, Sir Run Run Shaw Hospital (No. 2018KFJJ05). The authors would like to thank the reviewers in advance for their comments and suggestions.

References

1. Short MB, Black L, Smith AH, Wetterneck CT, Wells DE (2012) A review of internet pornography use research: methodology and content from the past 10 years. Cyberpsychol Behav Soc Netw 15(1):13–23. https://doi.org/10.1089/cyber.2010.0477
2. Owens EW, Behun RJ, Manning JC, Reid RC (2012) The impact of internet pornography on adolescents: a review of the research. Sex Addict Compuls 19(1–2):99–122. https://doi.org/10.1080/10720162.2012.660431
3. Manning JC (2006) The impact of internet pornography on marriage and the family: a review of the research. Sex Addict Compuls 13(2–3):131–165. https://doi.org/10.1080/10720160600870711
4. Zaidan A, Karim HA, Ahmad N, Zaidan B, Sali A (2013) An automated anti-pornography system using a skin detector based on artificial intelligence: a review. Int J Pattern Recognit Artif Intell 27(04):1350012. https://doi.org/10.1142/S0218001413500122
5. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
6. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
7. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 770–778
8. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 4700–4708
9. Wang X, Cheng F, Wang S, Sun H, Liu G, Zhou C (2018) Adult image classification by a local-context aware network. In: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, pp 2989–2993, https://doi.org/10.1109/ICIP.2018.8451366
10. Zhu R, Wu X, Zhu B, Song L (2018) Application of pornographic images recognition based on depth learning. In: Proceedings of the 2018 International Conference on Information Science and System, ACM, pp 152–155, https://doi.org/10.1145/3209914.3209946
11. Nian F, Li T, Wang Y, Xu M, Wu J (2016) Pornographic image detection utilizing deep convolutional neural networks. Neurocomputing 210:283–293. https://doi.org/10.1016/j.neucom.2015.09.135
12. Moustafa M (2015) Applying deep learning to classify pornographic images and videos. arXiv preprint arXiv:151108899
13. Vitorino P, Avila S, Perez M, Rocha A (2018) Leveraging deep neural networks to fight child pornography in the age of social media. J Vis Commun Image Represent 50:303–313. https://doi.org/10.1016/j.jvcir.2017.12.005
14. Zhu H, Zhou S, Wang J, Yin Z (2007) An algorithm of pornographic image detection. In: Fourth International Conference on Image and Graphics (ICIG 2007), IEEE, pp 801–804, https://doi.org/10.1109/ICIG.2007.29
15. Srisaan C (2016) A classification of internet pornographic images. Int J Electron Commerce Stud 7(1):95–104. https://doi.org/10.7903/ijecs.1408
16. Moreira DC, Fechine JM (2018) A machine learning-based forensic discriminator of pornographic and bikini images. In: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, pp 1–8, https://doi.org/10.1109/IJCNN.2018.8489100
17. Deselaers T, Pimenidis L, Ney H (2008) Bag-of-visual-words models for adult image classification and filtering. In: 2008 19th International Conference on pattern recognition, IEEE, pp 1–4, https://doi.org/10.1109/ICPR.2008.4761366
18. Avila S, Thome N, Cord M, Valle E, AraúJo ADA (2013) Pooling in image representation: the visual codeword point of view. Comput Vis Image Underst 117(5):453–465. https://doi.org/10.1016/j.cviu.2012.09.007
19. Zhuo L, Geng Z, Zhang J, Guang Li X (2016) ORB feature based web pornographic image recognition. Neurocomputing 173:511–517. https://doi.org/10.1016/j.neucom.2015.06.055
20. Liu Y, Gu X, Huang L, Ouyang J, Liao M, Wu L (2019) Analyzing periodicity and saliency for adult video detection. arXiv preprint arXiv:190103462
21. Tang S, Li J, Zhang Y, Xie C, Li M, Liu Y, Hua X, Zheng YT, Tang J, Chua TS (2009) Pornprobe: an lda-svm based pornography detection system. In: Proceedings of the 17th ACM International Conference on Multimedia, ACM, pp 1003–1004, https://doi.org/10.1145/1631272.1631490
22. Liu Y, Xie H (2009) Constructing surf visual-words for pornographic images detection. In: 2009 12th International Conference on computers and information technology, IEEE, pp 404–407, https://doi.org/10.1109/ICCIT.2009.5407272
23. Yizhi L, Shouxun L, Sheng T, Yongdong Z (2010) Adult image detection combining bovw based on region of interest and color moments. In: International Conference on intelligent information processing, Springer, pp 316–325, https://doi.org/10.1007/978-3-642-16327-2_38
24. Zhang D, Zou L, Zhou X, He F (2018) Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer. IEEE Access 6:28936–28944. https://doi.org/10.1109/ACCESS.2018.2837654
25. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556
26. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 1–9
27. Kim A (2019) NSFW dataset. https://github.com/alexkimxyz/nsfw_data_scraper. Accessed 1 Apr 2019
28. Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 1717–1724
29. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167
30. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: Proceedings of the fourteenth International Conference on artificial intelligence and statistics, pp 315–323
31. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191
32. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) A survey on deep transfer learning. In: International Conference on artificial neural networks, Springer, pp 270–279, https://doi.org/10.1007/978-3-030-01424-7_27

33. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
34. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 7132–7141
35. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European Conference on computer vision, Springer, pp 818–833, https://doi.org/10.1007/978-3-319-10590-1_53
36. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in PyTorch. In: NIPS 2017 autodiff workshop: the future of gradient-based machine learning software and techniques
37. Paszke A, Suhan A, Meurer A, Gross S (2019) Pretrained models from torchvision. https://github.com/pytorch/vision/tree/master/torchvision. Accessed 3 Apr 2019
38. He T, Zhang Z, Zhang H, Zhang Z, Xie J, Li M (2019) Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 558–567
39. Caetano C, Avila S, Guimaraes S, Araújo AdA (2014) Pornography detection using bossanova video descriptor. In: 2014 22nd European Signal Processing Conference (EUSIPCO), IEEE, pp 1681–1685
40. Agastya IMA, Setyanto A, Handayani DOD, et al. (2018) Convolutional neural network for pornographic images classification. In: 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), IEEE, pp 1–5, https://doi.org/10.1109/ICACCAF.2018.8776843