Journal of Information Hiding and Privacy Protection
DOI:10.32604/jihpp.2020.010464
Review

Deep Learning for Distinguishing Computer Generated Images and Natural Images: A Survey

Bingtao Hu* and Jinwei Wang

Nanjing University of Information Science and Technology, Nanjing, 210044, China
*Corresponding Author: Bingtao Hu. Email: 20181220007@nuist.edu.cn
Received: 21 October 2020; Accepted: 28 October 2020

Abstract: With the development of computer graphics, realistic computer graphics (CG) have become more and more common in our field of vision. This rendered image is invisible to the naked eye. How to effectively identify CG and natural images (NI) has been become a new issue in the field of digital forensics. In recent years, a series of deep learning network frameworks have shown great advantages in the field of images, which provides a good choice for us to solve this problem. This paper aims to track the latest developments and applications of deep learning in the field of CG and NI forensics in a timely manner. Firstly, it introduces the background of deep learning and the knowledge of convolutional neural networks. The purpose is to understand the basic model structure of deep learning applications in the image field, and then outlines the mainstream framework; secondly, it briefly introduces the application of deep learning in CG and NI forensics, and finally points out the problems of deep learning in this field and the prospects for the future.

Keywords: Deep learning; convolutional neural network; image forensics; computer generated image; natural image

1 Introductions

Natural Images (NI) reflect real-world scenes, and computer graphics tools can now generate virtual but visually trustworthy images. Therefore, the recognition of NI and computer generated CG images has received increasing attention. However, this problem is difficult to solve because the ultimate goal of computer graphics is to make CG images have the same surrealism as NI. The emergence of ultra-realistic CG images has revolutionized the multimedia industry, providing the ability to easily create realistic animations and images. Such effective technical support has brought new possibilities to the games, movies and other industries. Numerous realistic games and movie works have appeared in our field of vision, bringing a new visual experience to players and audiences. People can finally break through the reality and give full play to their imagination in the virtual world. At the same time, if CG images are used in areas such as law, authority, and news [1], it poses a serious security threat to the public. For example, using false CG images to tamper with evidence, confuse audiovisuals, or even frame others, will affect normal analysis and judgment in the judicial realm. Therefore, the recognition of CG and NI has become an important topic in the field of image forensics [2], and has attracted widespread attention in the past decade.

Feature representation is the key to image processing. Traditional feature design needs to be done manually. However, this method is complicated and has high requirements on the designer's technology. Therefore, automatic feature design has become an urgent requirement for efficient image processing. Deep learning is an emerging field of machine learning research. It aims to study how to automatically extract multi-level feature representations from data. The core idea is to extract multiple levels from the

original data through a series of nonlinear transformations in a data-driven manner. Multi-angle features, so that the acquired features have greater generalization and expressive ability, which just meets the needs of efficient image processing. In order to meet the various needs of image processing problems, the deep learning theory represented by convolutional neural network has made breakthroughs. Based on the basic principles of deep learning, this paper summarizes the evolution and innovation of its algorithms, models and even methods in the field of CGNI image forensics. The purpose is to keep track of the latest developments and applications of deep learning in the field of CG and NI forensics.

2 Deep Learning

2.1 Related Background

Back in the 1950s, neural networks were used as a toy project to study their core ideas, but for a long time there was no way to train large neural networks. This changed in the mid-1980s, and many people independently discovered the backpropagation algorithm, a method of training a series of parametric operations chains using gradient descent optimization, and began to apply it to neural networks. For the first time in 1989, Bell Labs successfully implemented the practical application of neural networks. Yann LeCun combined the idea of convolutional neural networks with backpropagation algorithms and applied them to the problem of handwritten digital image classification, which led to LeNet. The network [3,4] was adopted by the US Postal Service in the 1990s to automatically read postal codes on envelopes. Then, due to the limitations of hardware conditions and data, the development of neural networks has once again encountered bottlenecks. What followed was the rapid development of a new machine learning method called nuclear method, one of which is the well-known support vector machine (SVM).

From 1990 to 2010, the speed of non-custom CPUs increased by about 5,000 times; in the first decade of the 1900s, companies such as NVDIA and AMD invested billions of dollars to develop fast massively parallel chips (GPUs). In terms of data, in addition to the exponential growth of storage hardware over the past 20 years, the biggest change has come from the rise of the Internet, which has made it possible to collect and distribute very large datasets for machine learning. The ImageNet dataset [5] has spawned the rise of deep learning, which contains 1.4 million images that have been manually divided into 1000 image categories (one for each image). The special thing about ImageNet is not only the sheer volume, but also the annual competition associated with it. Researchers challenge the common benchmark through competition, which greatly promotes the rise of deep learning in the near future. In addition to hardware and data, we did not have a reliable way to train very deep neural networks until the end of the first decade of the 20th century. Therefore, the neural network is still very shallow, using only one or two presentation layers, and cannot go beyond more precise shallow methods (such as SVM and forest trees). The key issue is the gradient propagation through multiple layers of overlays. As the number of layers increases, the feedback signal used to train the neural network will gradually disappear. This situation changed around 2009–2010, when several simple but important algorithmic improvements occurred: 1) Better activation function; 2) Better weight initialization scheme (weight) -initialization scheme); 3) Better optimization schemes such as RMSProp and Adam. These improvements made it possible to train more than 10 layers of models, and deep learning began to receive widespread attention.

In the past few years, deep learning has demonstrated its powerful performance in various speech, image and video classification and recognition tasks.In the field of multimedia forensics, the performance of median filter detection [6], pattern recognition [7–9], forgery detection [10] and steganographic analysis [11–14] has also been greatly improved under the use of CNN.

2.2 The Basic Structure of the Neural Network

Neurons. Deep learning mainly relies on neural network technology. The basic unit of neural network is neuron, as shown in Fig. 1.

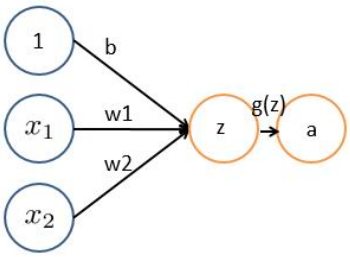

Figure 1: Neuron

x1, x2 represent the input vector, w1, w2 represent the weight, and several inputs mean that there are several weights, that is, each input is given a weight, b is the bias, g(z) is the activation function, a is the output.

Convolution. For the image and the filter matrix to do the inner product (each of the elements is multiplied and then summed). The operation of re-summation is the so-called "convolution" operation, which is also the source of the name of the convolutional neural network.

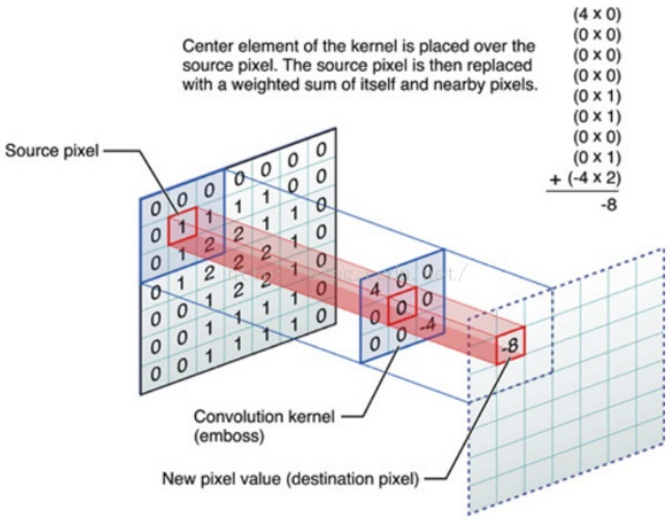

Figure 2: Convolution

The left part of Fig. 2 is the original input data, the middle part of the figure is the filter, and the right side of the figure is the new two-dimensional data output. The intermediate filter and the data window do inner product, and the specific calculation process is: $4 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 1 + 0 \times 1 + 0 \times 0 + 0 \times 1 + -4 \times 2 = -8$.

Activation function. Commonly used nonlinear activation functions are sigmoid [15], tanh, ReLU, etc. The first two sigmoid/tanh are more common in the fully connected layer, the latter relu is common in the convolutional layer. In the actual gradient descent, the sigmoid is easily saturated, causing the termination of the gradient transfer, and there is no zero centralization. The ReLU function does not have these defects, and its graphical representation is as follows:

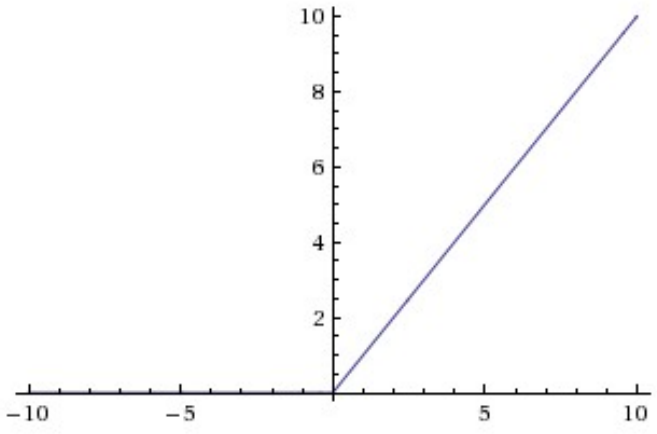

Figure 3: ReLU

Pooling. Pooling, exactly speaking is taking the numerical eigenvalues (maximum, average) of the region, as shown in Fig. 4.

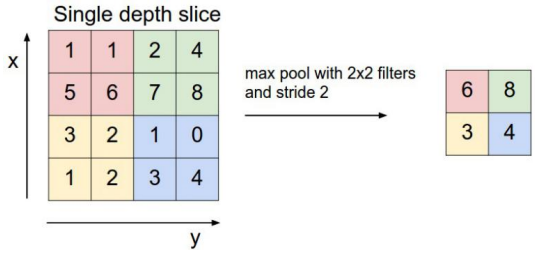

Figure 4: Maximum pooling

In the left part of Fig. 4, the matrix of 2 × 2 in the upper left corner is 6 largest, the matrix of 2 × 2 in the upper right corner is 8 largest, the matrix of 2 × 2 in the lower left corner is 3 largest, and the matrix of 2 × 2 in the lower right corner is the largest, so the result of the right part of the above figure is obtained.

3 The Mainstream CNN Framework

3.1 AlexNet

The literature [16] is considered to be the origin of deep learning. The title of the article is “ImageNet Classification with Deep Convolutional Networks”, which is widely regarded as an article with far-reaching influence in the industry. Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton created a "large-scale, deep convolutional neural network" and used it to win the 2012 ILSVRC Challenge (ImageNet Large-Scale Visual Recognition Challenge). As the Olympics in the field of machine vision, ILSVRC attracts research teams from all over the world every year. They come up with all kinds of competition and use their own machine vision models/algorithms to solve image classification, positioning and detection. In 2012, when CNN first entered the stage, it achieved a good score of 15.4% in the top five test error rate. The score behind it is 26.2%, indicating that CNN has a shocking advantage over other methods, which caused a huge shock in the field of machine vision. It can be said that since then CNN has become a household name in the industry.

Literature [16] mainly discusses the implementation of a network architecture (we call it AlexNet). Compared with the current architecture, the layout structure discussed in [16] is relatively simple, mainly including five convolutional layers, a maximum pooling layer, a dropout layer, and three fully connected layers. The structure is categorized for having 1000 possible image categories.

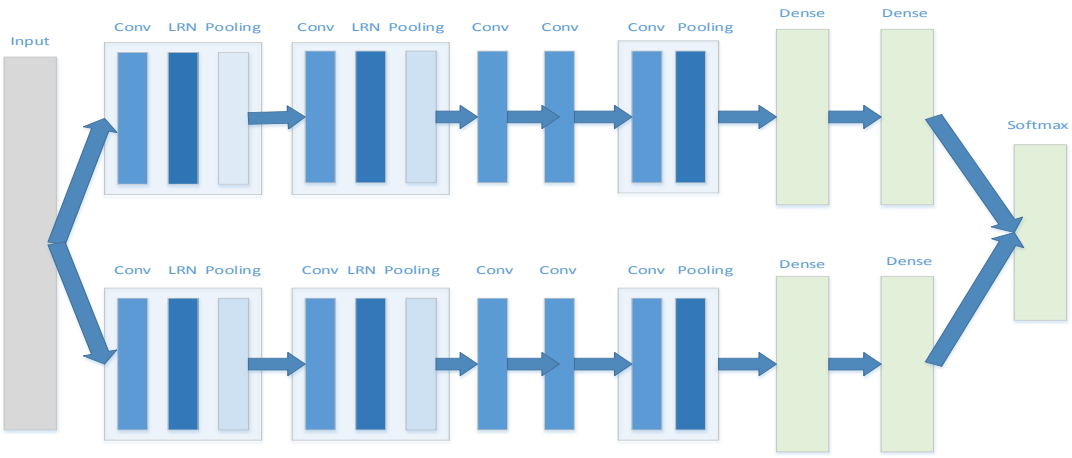

Figure 5: AlexNet

The main points of the literature [16]:

1) Using the ImageNet database for network training, the library contains 22,000 kinds of 15 million tag data;
2) Using a nonlinear function of the linear rectification layer ReLU. (Using the linear rectification layer ReLU, the running speed is several times faster than the traditional hyperbolic tangent function);
3) utilizes data augmentation methods, including image transformation, horizontal reflection, block extraction and other methods;
4) Introduced the dropout layer [17] to solve the over-fitting problem of the training set;
5) Training using batch stochastic gradient descent, setting limits for momentum moment and weight decay weight decay.

3.2 VGGNET

One of the 2014 ILSVRC models relied on a simple increase in depth to achieve a 7.3% error rate (but not the champion of the year), named VGGNET [18]. Oxford's Karen Simonyan and Andrew Zisserman created a 19-layer CNN with only 3 × 3 size filters in the network, with 1 stride and padding, and a pooled layer using 2 × 2 max. Pooled function with a step size of 2.

The main points of the literature [18]:

1) Use only 3 × 3 filters, which is quite different from the previous AlexNet first layer 11 × 11 filter and ZF Net 7 × 7 filter. The reason stated by the author is that two 3*3 convolutional layers combine to produce a valid 5 × 5 perception zone. Therefore, the use of a small-sized filter can maintain the same function as a large size while ensuring the advantage of a small size. One of the advantages is the reduction of the parameters. Another advantage is that we can use one more linear rectification layer ReLU for the two convolution networks. (The more ReLU, the lower the linear performance of the system).
2) Three 3 × 3 convolutional layers are arranged side by side to represent a valid 7 × 7 perception zone.
3) The spatial size of the input image decreases as the number of layers increases (by convolution or pooling of each layer), and its depth increases as the number of filters increases.

3.3 GoogLeNet

Google has abandoned the principle of keeping the network hierarchy simple in its own architecture, Inception Module [19–22]. GoogLeNet [23] is a 22-layer CNN that won the 2014 ILSVRC championship with a 6.7% error rate. This is the first new architecture of CNN that differs from the traditional method in

that the convolutional layer and the pooled layer are simply superimposed to form a sequence structure. The authors emphasize that their new model also pays special attention to the use of memory and computational quantities (this has not been considered before: multi-layer stacking and the use of a large number of filters can consume a lot of computational and storage resources, as well as over-fitting. The chance). The Inception structure is shown in Fig. 6.

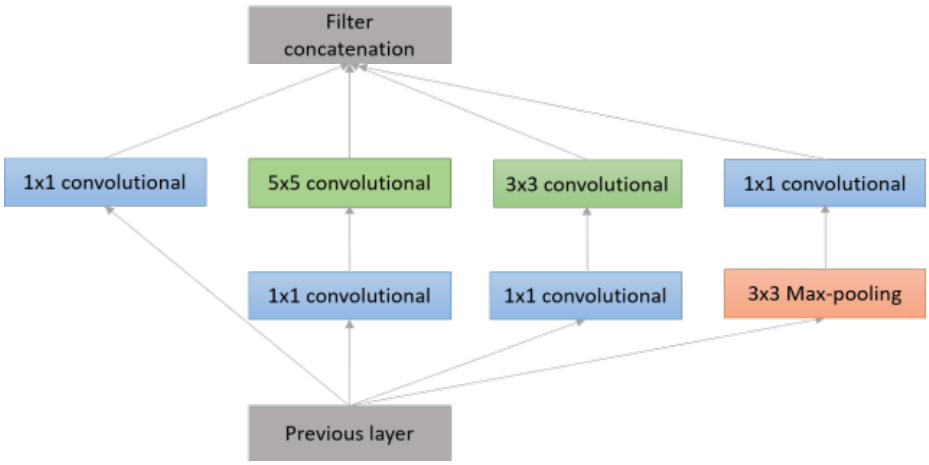

Figure 6: Inception

The main points of the literature [23]:

1) A total of 9 Inception module modules are used in the model, with a total depth of 100 layers.
2) Instead of using the fully connected layer, use an average pooling pool average [24] instead, and reduce the data of 7 × 7 × 1024 to 1 × 1 × 1024. This configuration greatly reduces the number of parameters.
3) 12 times less than AlexNet's parameters.
4) In the test, multiple cropping images of the same input image are used as system inputs, and the results are normalized by the exponential function softmax averaging operation to obtain the final result.
5) Introduced the concept of regional convolutional network (R-CNN) [25] in the model.
6) Inception module is constantly updated.

3.4 Residual Network

ResNet [26] is a 152-layer network architecture that combines classification, detection and translation capabilities. ResNet's own performance broke the record of ILSVRC2015, reaching an incredible 3.6% (usually humans can only achieve 5 to 10% error rate). The residual block concept of the residual block proposed in the article is designed as follows: The result generated by the operation of input x through convolution – linear rectifying – convolution series is set as F(x), which is added to the original input x to obtain H(x) = F(x) + x. In traditional CNN, only H(x) = F(x). And in ResNet you have to add the convolution result F(x) to the input x. The submodule in Fig. 7 shows a calculation process that is equivalent to calculating a small change "delta" for the input x, so that the output H(x) is the superposition of x and the change delta (in traditional CNN, the output F(x) is a completely new expression, which does not contain the information of the input x). The author believes that the residual mapping is easier to optimize than the previously unreferenced mapping.

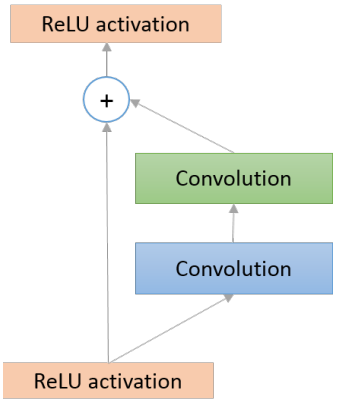

Figure 7: Residual network

The main points of the literature [26]:

1) Depth reaches 152 layers. After the first two layers are processed, the input image space size is compressed from 224 × 224 to 56 × 56.
2) The authors declare that if you increase the number of layers randomly in the plain nets, the training calculation and the error rate will increase.
3) Try to use the 1202 layers network architecture, the accuracy of the result is reduced, the presumed reason is over-fitting.

4 Deep Learning and Forensics of CG and NI

4.1 Relevant Status

The starting point of CG and NI forensics is to extract the relationship between local pixels. Different from image recognition, image recognition is to distinguish the difference in image content, and extract high-level semantic features. CG and NI forensics mainly focus on low-level images, try not to let the neural network extract high-level semantic features, but focus on simple statistical features as features into the classifier. Therefore, the general deep learning model of CG and NI forensics problems is not very competent. Generally dealing with CG and NI forensics, we often divide it into three steps: feature extraction, feature transformation, and incoming classifier classification. The existing deep learning-based CG and NI forensics methods are mainly embodied in the three aspects of feature extraction, feature transformation and migration learning. The literature on deep learning based CG and NI forensics is not rich. I think the main reason for this situation may be the lack of data sets and the difficulty of self-built CG image data sets. Before Nicolas Rahmouni et al. [27] presented their own data sets in 2017, most of the methods were based on the original Colombian dataset [28], the Colombian dataset was proposed in 2004, and the CG images were placed in today's CG. And the lack of persuasiveness on the NI forensics mission. After the new data set was presented, the most advanced methods are now experimenting with Nicolas Rahmouni's datasets, but their datasets also have a significant problem with too little sample size. The CG image and the NI image each contain 1800 sheets and the fidelity of the CG image needs to be improved. Self-built CG image datasets are a very difficult task, because there are fewer channels for acquiring CG images, and video game screenshots are more convenient. There are few games that can meet the requirements of realism, and screenshots are taken. Building a large data set itself has a lot of work. These problems have limited the research and development of CG and NI forensics.

4.2 Application of Deep Learning in Forensics of CG and NI

Method based on image preprocessing. Yao et al. [29] proposed a recognition method based on sensor pattern noise (SPN) and deep learning CG and NI. These images (CG and NI) are clipped into the image patch before being input to the Convolutional Neural Network (CNN) based model. In addition, three high pass filters (HPFs) are used to remove low frequency signals representing image content. These

filters are also used to display residual signals as well as SPNs introduced by digital camera devices. Unlike traditional methods of distinguishing between CGs and NIs, this method uses five layers of CNN to classify input image patches. Based on the classification result of the image patch, a majority ticket scheme is deployed to obtain the classification result of the full-size image. This method achieves nearly 100% verification accuracy on a data set contributed by Nicolas Rahmouni et al.

Method based on feature extraction. Quan et al. [30] made some changes to the deep learning model, as shown in Fig. 8, A convFilter layer is added in front of the traditional neural network structure, which is specifically composed of 32 3D convolution kernels. The performance of this network is better than the traditional method, not only for Google and PRCG data sets, but also for simple data sets, they consider the fine-tuning and structure of CNN. Activate function design, flexibility, visualization, and understanding. After voting on the 240 × 240 patch, it achieved 93.2% accuracy on full-size images.

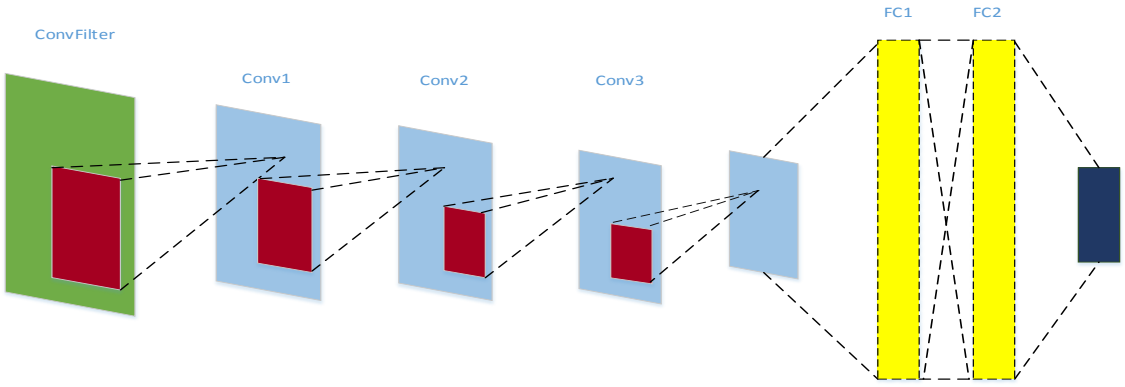

Figure 8: Network structure proposed by Wang et al.

Chawla et al. [31] proposed an improved Convolution layer called New Conv Layer. As shown in Fig. 9, The idea of the improved convolutional layer comes from the fact that the structural relation between some regions in the image has nothing to do with the image content, but exists in the pixel relation. The correlation between pixels in a photographic image is different from that between computer-generated images. Therefore, the classifier should determine the relationship between the pixels and them. The difference between this layer and the ordinary convolutional layer is that the following constraint relationship needs to be satisfied during training.

$$w_k^{(l)}(0,0) = -l \tag{1}$$

$$\sum_{l,\ m\neq 0} w_k^{(l)}(l,m) = l \tag{2}$$

where $w_k^{(l)}(0,0)$ is the weight of the filter center point, and $w_k^{(l)}(l,m)$ is the weight of the filter at point (l,m).This method achieved near 100% accuracy on Nicolas Rahmouni's dataset.

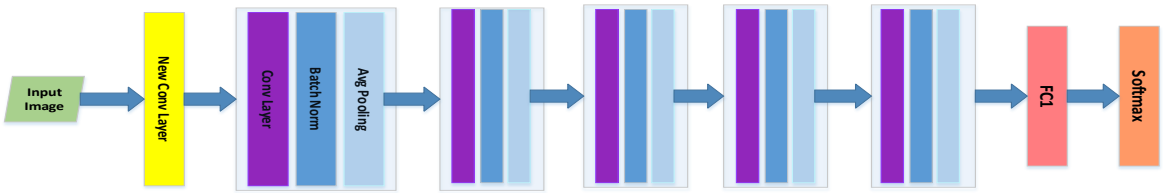

Figure 9: Network structure proposed by Chaitanya Chawla et al.

Method based on feature-conversion. Nicolas Rahmouni et al. [27] pointed out that the commonly used Colombian dataset [28] was created in 2004, and the CG images at the time were incomparable with the current CG images, so they collected and contributed a data set of their own. The CG image of the

dataset is derived from the Level-Design reference database [32], which consists of more than 60,000 high-resolution video game screenshots, and selects 1800 CG images that meet the criteria. Another 1800 NI images are from the RAISE data set [33]. Their method adds a feature conversion layer after the feature extraction of the two layers of convolutional layers, in order to extract the simple statistical properties including the expectation, variance, maximum and minimum values, and then pass it to the classifier. This method achieved an accuracy of 93.2% on their own data set.

Then HuyH.Nguyen et al. [34] combined the idea of migration learning on the basis of Nicolas Rahmouni, using the pre-trained VGG-19 network for feature extraction, and then taking features for feature conversion after each layer is convolved. Finally, each layer of transformed features is passed to the classifier for training, and nearly 100% accuracy is obtained on the new data set contributed by Nicolas Rahmouni et al.

Method based on transfer learning. Edmar R. S. de Rezende∗, Guilherme C. S. Ruppert∗ et al. [35] proposed a new CG image detection method based on ResNet-50 deep convolutional neural network model and transfer learning concept. After simple pre-processing, each image in the dataset is input into the deep CNN model. Therefore, the CNN convolution obtains a 2048-dimensional feature vector, which is called the bottleneck feature. These feature vectors are used to train a machine learning classifier to detect whether an image is generated by a computer graphics method. Applying the t-SNE dimensionality reduction technique to the visualization of high-dimensional features, it can be observed that the bottleneck feature generated by the ResNet-50 transport layer is more separable than the original image input feature, which makes the classification task easier. The method achieved an accuracy of 94.1% on a public dataset containing 9,700 images of different scenes.

Tiago Carvalho∗, Edmar R. S. de Rezende† proposed a CG image detection method based on the preservation of important information in the human eye region. The method determines whether the image is a CG image by determining whether the human eye region is generated by the CG technique. When the highlight of the eye is removed from the previously detected and cropped eye area, the resulting image of the CG image (without reflection) shows more artifacts than the PG image (and no reflection). Since the human eye is a difficult part of the generation in CG images, they believe that these artifacts are caused by defects in the computer graphics technology used to generate the eye. Once each eye in the image is positioned and its highlights are removed, they use the idea of migration learning to remove the fully connected layer from the pre-trained VGG19 architecture in Imagenet as a feature extractor to create a set of bottlenecks feature. Finally, the feature extracted from each eye is input to the classifier to detect whether one eye is generated by the CG, thereby detecting whether the image is generated by the CG.

5 Conclusion

Deep learning is mainly based on convolutional neural network research in the image field, but CNN network training requires large-scale database and powerful computing power, and obtaining labeled samples requires a lot of manpower and material resources, even in some directions, such as for the problem of forensics, there is no human or material force to calibrate the sample, and the corresponding professional knowledge is needed. The data sets needed for the forensic problems of CG and NI are difficult to construct. At present, the related data sets are slow to update, the number of samples is small, and the diversity is poor. Therefore, the trained models have poor generalization and are prone to over-fitting problems. In the future work, building a complete data set is an urgent problem to be solved. Further improvement of the model and enhancement of its generalization are the main objectives.

Funding Statement: This work is supported by National Natural Science Foundation of China (62072250).

Conflicts of Interest: We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

[1] Rocha, W. Scheirer, T. Boult and S. Goldenstein, "Vision of the unseen Current trends and challenges in digital image and video forensics," *ACM Computing Surveys*, vol. 43, no. 4, pp. 26–40, 2011.

[2] Y. Rao and J. Ni, "A deep learning approach to detection of splicing and copy-move forgeries in images," *Workshop on Information Forensics and Security*, pp. 1–6, 2016.

[3] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech and time series," *The Handbook of Brain Theory and Neural Networks*, vol. 3361, no. 10, 1995.

[4] Y. LeCun, B. Boser, J. S. Denker, D. Henderson and R. E. Howard *et al.,* "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp.541–551, 1989.

[5] J. Deng, W. Dong and R. Socher, "ImageNet: A large-scale hierarchical image database," *Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[6] J. Chen, X. Kang, Y. Liu and Z. J. Wang, "Median filtering forensics based on convolutional neural networks," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1849–1853, 2015.

[7] Tuama, F. Comby and M. Chaumont, "Camera model identification with the use of deep convolutional neural networks," in *2016 IEEE Int. Workshop on Information Forensics and Security*, pp. 1–6, 2016.

[8] L. Bondi, L. Baroffio, D. Güera, P. Bestagini, E. J. Delp *et al.,* "First steps toward camera model identification with convolutional neural networks," *IEEE Signal Processing Letters* vol. 24, no. 3, pp. 259–263, 2017.

[9] L. Bondi, D. Güera, L. Baroffio, P. Bestagini, E. J. Delp *et al.,* "A preliminary study on Convolutional Neural Networks for camera model identification," *Electronic Imaging 2017*, pp. 67–76, 2017.

[10] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pp. 5–10, 2016.

[11] Y. Qian, J. Dong, W. Wang and T. Tan, "Deep learning for steganalysis via convolutional neural networks," *Media Watermarking, Security and Forensics*, 2015.

[12] L. Pibre, P. Jérˆ ome, D. Ienco and M. Chaumont, "Deep learning for Steganalysis is better than a Rich Model with an Ensemble Classifier and is natively robust to the cover source-mismatch," arXiv preprint arXiv:1511.04855, 2015.

[13] G. Xu, H. Z. Wu and Y. Q. Shi, "Structural design of Convolutional Neural Networks for Steganalysis," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 708–712, 2016.

[14] V. Sedighi and J. Fridrich, "Histogram layer, moving Convolutional Neural Networks towards feature-based Steganalysis," *Electronic Imaging 2017*, pp. 50–55, 2017.

[15] J. Rafferty, P. Shellito and N. H. Hyman, "Practice parameters for sigmoid diverticulitis," *Diseases of the Colon & Rectum*, vol. 49, no. 7, pp. 939–944, 2006.

[16] Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Neural Information Processing Systems*, pp. 1106–1114, 2012.

[17] G. E. Hinton, N. Srivastava and A. Krizhevsky, "Improving neural networks by preventing co-adaptation of feature detectors," *Computer Science*, vol. 3, no. 4, pp. 212–223, 2012.

[18] K. Simonyan and A. Zisserman, "Deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. on Machine Learning*, pp. 448–456, 2015.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet and S. Reed *et al.,* "Going deeper with convolutions," *Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[21] C. Szegedy, V. Vanhoucke and S. Ioffe, "Rethinking the inception architecture for computer vision," *Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016

[22] C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi Inception-v4, "Inception-ResNet and the impact of residual connections on learning," *Computer Vision and Pattern Recognition*, arXiv: 1602.07261, 2017.

[23] C. Szegedy, W. Liu, Y. Jia and P. Sermanet, "Going deeper with convolutions," *Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[24] M. Lin, Q. Chen and S. Yan, "Network in network," *Computer Science*, 2013.

[25] R. Girshick, J. Donahue and T. Darrell, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[26] K. M. He, X. Y. Zhang, S. Q. Ren and J. Sun, "Deep residual learning for image recognition," *Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[27] N. Rahmouni, V. Nozick and J. Yamagishi, "Distinguishing computer graphics from natural images using convolution neural networks," in *2017 IEEE Workshop on Information Forensics and Security*, pp. 1–6, 2017.

[28] T. T. Ng, S. F. Chang, J. Hsu and M. Pepeljugoski, "Columbia photographic images and photorealistic computer graphics dataset," *Columbia University, ADVENT Technical Report*, pp. 205–2004, 2005.

[29] Y. Yao, W. Hu and W. Zhang, "Distinguishing computer-generated graphics from natural images based on sensor pattern noise and deep learning," *Sensors*, vol. 18, no. 4, pp. 1296, 2018.

[30] W. Quan, K. Wang and D. M. Yan, "Distinguishing between natural and computer-generated images using convolutional neural networks," *IEEE Transactions on Information Forensics and Security*, vol.13, no. 11, pp. 2772–2787, 2018.

[31] C. Chawla, D. Panwar and G. S. Anand, "Classification of computer generated images from photographic images using convolutional neural networks," *2018 Int. Conf. on Advances in Computing, Communication Control and Networking*, pp. 1053–1057, 2018.

[32] M. Piaskiewicz, "Level-design reference database," 2017. [Online]. Available: http://level-design.org/referencedb.

[33] D. T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, "Raise: A raw images dataset for digital image forensics," in *Proc. of the 6th ACM Multimedia Systems Conf.*, pp. 219–224, 2015.

[34] N. H. Nguyen, T. Tieu and H. Q. Nguyen-Son, "Modular convolutional neural network for discriminating between computer-generated images and photographic images," in *Proc. of the 13th Int. Conf. on Availability, Reliability and Security*, pp. 1–10, 2018.

[35] E. R. S. De Rezende, G. C. S. Ruppert and T. Carvalho, "Detecting computer generated images with deep convolutional neural networks," in *30th SIBGRAPI Conf. on Graphics, Patterns and Images*, pp. 71–78, 2017.