Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/40/20230663

The recognition and analysis of fake images based on deep neural network

Qihang Luo

The National Mathematics and Science College, Warwickshire, CV4 8JB, the United Kingdom

Qihang.Luo@natmatsci.ac.uk

Abstract. The generation of fake images and the process of technology is developing quickly, which has generated great concern about its impact on society. To effectively identify the fake image, this article introduces a deep learning-based identification method. This article uses a convolutional neural network (CNN) neural network to construct a model for machine learning image recognition and lets it learn a dataset of hundreds of real and fake pictures to let it complete tasks that are difficult for humans to complete. the use of deep learning to solve this problem has the following advantages: the results can be observed after data input, which is conservative and fast, and there is no need to manually design rules as Deep learning methods can optimise the loss function to learn the rules, mining potential features of data, with strong representation ability. Using this method, the accuracy result is over 60%. This paper proves a machine can learn and recognise real and fake pictures, which is specifically inspiring that learning-based methods can also solve the challenging problem of images.

Keywords: Fake Image, Deep Learning, Convolutional Neural Network (CNN).

1. Introduction

Nowadays, the usage of picture editing software is continuously increasing. One report records that Photoshop, one of the most popular software in the category, has revenue in 2023 Q2. On the internet, much information is tampered with by software, which could spread rumours and create fake news. Even if a simple material is reversed, the hero on the battlefield can instantly become a deserter visually. Both of them provide the audience with a new perspective of observation, but because the latter involves specific people and specific scenes, this editing method not only changes the fate of the characters in virtual time and space but also very likely changes it in real time and space. The fate of real people. Once the edited virtual space-time becomes a reality, the offense and harm to the parties are predictable. The problem is even worse if the face is replaced. For example, fake speech videos of famous politicians can be easily generated. It could also be used as a tool that threatens the safety of others [1]. In general, this may lead to the dissemination of false content, causing many harms such as trampling on the personality rights of editing objects, creating or exacerbating social conflicts, and destroying the credibility of mass media organizations. Therefore, it is crucial to accomplish effective true-false image recognition.

As deep learning becomes more widespread and is the most mature in processing video and pictures, Generative adversarial networks (GAN) occur. The Deepfake is built on it. It can easily replace one

DOI: 10.54254/2755-2721/40/20230663

person's face with another person in the picture [2]. It was also the first face-changing program, and Zao was also born for face manipulation. Several years ago, the image quality was improved by Karras et al. using the GANs' incremental growth, generating better-quality human pictures [3]. There are many traditional ways to recognise the truth of pictures and videos. One of the ways is to embed an extra signal into the source image without visual artefacts, after that, it is extracted to be restored. The source image which is extracted can be detected for tampered areas [4]. Unfortunately, for the picture that is generated by GANs, there isn't a source image. Other ways that are used commonly including Image Splicing, Copy-Move, and Object Removal are not working very well, due as after generation after generation of updates, images of GANs start to include detailed features such as hair and wrinkles [5]. The deep learning method is frequently used in the recognition area these days. The convolution neural network (CNN) shows that the performance is improved [6, 7]. To develop a such detector, images produced by GANs are necessary for training and validating. However, it's hard and time-consuming. Additionally, after using this learning strategy, the machine tends to recite all parts of the image, which means it could have a bad generalization ability [6,8,9].

To effectively recognize true and false images, this study introduces deep learning techniques to create a false image detector. Specifically, first, the generalization ability is improved by some degree of rotation, flipping and other operations in the preprocessing stage. Second, CNN is introduced to build an analytical model. CNN can represent the input effectively without a priori. Multi-layer convolutional operations are coupled with pooling layers for hierarchical feature representation to improve recognition performance. Third, a large number of experiments are conducted to analyze the dataset. By analyzing and comparing the prediction performance of different optimizers, the model that is used in this paper has reached a high accuracy rate, which can effectively identify the authenticity of images. This research can promote the development of the Internet industry and provide ideas and valuable insights for researchers.

2. Methodology

2.1. Dataset description and preprocessing

The dataset used is called real and fake face detection from Kaggle [10]. The dataset includes 2 file that contains hundreds of fake and true images respectively. Each image is in size of 600*600*3, with a person's face inside. Every image is separated into 3 difficulties for recognition: easy, mid and hard. The author mentioned that: These groups are separated subjectively, so it is not recommended to use them as explicit categories. In this case, only two labels are set to detect whether the picture is fictitious. So only the hard group of fake images is used. Examples of both types of images are listed in Figure 1. Before training, preprocessing is needed. Every image is inserted into a list and then converted to a float-type array, and a list of labels associated with the images is turned into one hot code. The position of labels and their associated images are the same in the list. Then the picture is transformed in different ways. Finally, the pictures are separated into 8:2 in the training set and testing set.

Figure 1. The real (left) and fake (right) images from the dataset.

DOI: 10.54254/2755-2721/40/20230663

2.2. Proposed approach

2.2.1. Model Construction. In this study, CNNs are introduced to create false image detectors for real and false image discrimination. Specifically, the model used includes 4 convolution layers, 2 max-pooling layers, a flattened layer, and a dense layer, the activation function used is Relu. One feature and advantage of this: is local awareness. Traditional neural networks have a large number of weights and are challenging to train because each neuron in the network must be connected to every pixel in the image. The number of weights in a neural network with a convolutional layer is equal to the size of the convolution kernel, which means that each neuron in the network is only connected to the pixels in the corresponding area of the image. As a result, there are significantly fewer weights. With one convolution kernel operation, not all the features may be got, so to get more feature information on the image, multi-core convolution is introduced. Use multiple convolution kernels to learn more different features of the image (each convolution kernel learns different weights).

The retrieved feature information's dimensionality can be decreased via the pooling layer. It uses feature compression to extract the key features while also reducing the size of the feature map, streamlining the network's computational cost and to some extent preventing over-fitting. Two factors illustrate the importance of the max pooling layer in computer vision: The computational complexity of the upper hidden layer is firstly reduced. Second, because these pooling units are translation invariant, the retrieved characteristics will not change even if the image has a slight displacement. The reason for using the activation function is that in the classification of two types of data, there are still relatively few cases where these data can be truly linearly separable. At this time, if the data is not linearly separable. At this time, using a straight line to divide the data can no longer work well. At this time, nonlinear factors are needed to classify the data, which is also the role of the activation function. The addition of the activation function can help the neural network more effectively address more challenging problems, increasing the expressiveness of the model because the linear model's limited expressive capacity is insufficient. The role of the Relu activation function is if the input is negative, it returns a zero and if the input is positive, it gives the same element (i.e., without any modification). The flattened layer is used to transfer the array into 1 dimension array, so the dense layer can modify. Finally, in the fully dense layer, all features will be connected, and the output value will be sent to the classifier, and then converted into a probability between 0 and 1 by the softmax function.

2.2.2. Model optimization. Adaptive Moment Estimation (Adam) is the optimizer used because it offers the following benefits: brevity, efficient computing, less memory requirement, invariance to diagonal scaling of gradients, suitability for problems with non-stationary targets, suitability for problems with very noisy or sparse gradients, and intuitive interpretation of hyperparameters with little tuning required. Adam realizes the benefits of the Adaptive Gradient Algorithm (AdaGrad) and RMSProp. Adam also employs the average of the second instant of the gradient (non-central variance), in contrast to Root Mean Square Propagation (RMSProp), where the parameter learning rate is based on averaging the first instant (mean). The AdaGrad maintains a configurable learning rate that enhances performance on issues involving sparse gradients (such as issues involving computer vision). Additionally, RMSProp keeps track of a learning rate for each parameter that is modified based on the average (for example, rate of change) of recent weight gradients. This indicates that the approach works effectively for issues that are both stationary and non-stationary, such as noise.

2.2.3. Loss function. In this paper, cross-entropy loss is used in the process of model optimization. The essence of cross-entropy loss is to measure the distance between two probability distributions. The real label distribution makes up one of the probability distributions, while the model's anticipated probability distribution makes up the other. The prediction of the model is more accurate when the two probability distributions are closer together and the cross-entropy loss is less. It has the two benefits of being able to quantify subtleties and being a convex optimization function, which makes it easier to employ gradient descent to find the best solution. The formula for cross-entropy loss is as follows:

DOI: 10.54254/2755-2721/40/20230663

$$CE(p,q) = -\sum_{i=1}^{C} pi \log(qi) \quad (1)$$

where pi represents the i-th element of the true label, and qi represents the probability that the model predicts that p belongs to the i-th category.

2.3. Implemented details

Two studies with different numbers of pictures are done. The learning rate is set to 1e-4, the epoch is 10 and the batch size is 8 but only 20 images in the dataset are input the first time. The second time, the learning rate is set to 1e-3, the epoch is set to 30 and the batch size is 32 as 400 images are used to learn and test. In both studies, the rate of decay is set to the learning rate divided by the number of epochs that are currently running.

3. Result and discussion

Figure 2 is the learning result for a small batch of pictures, while Figure 3 is a relatively large batch. It can be seen from this that after only about an hour of learning, the highest accuracy rate of this model can reach more than 80%. For a large number of pictures, the accuracy rate can reach nearly 70%. These matrices visually represent the distribution of true and predicted labels, providing insights into the strengths and weaknesses of the model. The image is not very smooth, indicating that there may be an overfitting tendency. Future work can consider exploring alternative combined network architectures, increasing the number of picture input or dropout layers to further optimize the overfitting tendency. In terms of results, the use of deep learning to solve this problem has the following advantages: the results can be observed after data input, which is conservative and fast, and there is no need to manually design rules as deep learning can maximize the loss function as much as feasible. Data mining with powerful representational capabilities of prospective features.

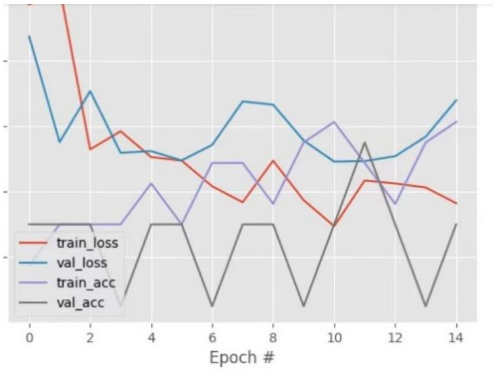

Figure 2. The performance curve of the small batch of pictures. The train_loss and val_loss mean the loss in training and valuation while the train_acc and val_acc mean the accuracy in training and valuation.

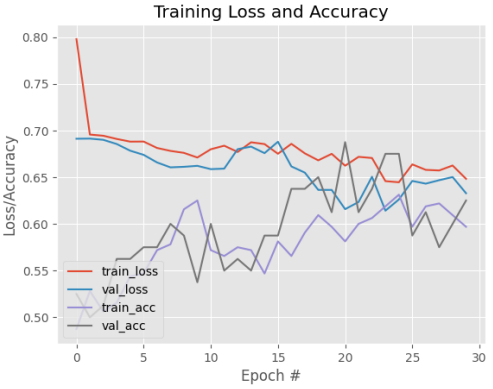

Figure 3. The curve of loss and accuracy. The train_loss and val_loss mean the loss in training and valuation while the train_acc and val_acc mean the accuracy in training and valuation.

DOI: 10.54254/2755-2721/40/20230663

4. Conclusion

This study uses deep learning to learn between real and fake images to determine their authenticity. By using CNN neural networks, the machine can predict the correct results. The purpose is to determine the images produced by new GANs on the internet to prevent the spread of false information. In the future, deep learning authenticity discrimination models for audio should also be created to identify potential forged audio. This article shows that even though cutting-edge face image processing techniques produce beautiful visual effects, trained counterfeit detectors are capable of spotting them. It is especially exciting that learning-based approaches may address the complex issue of pictures.

References

[1] Khodabakhsh A et al 2018 Fake face detection methods: Can they be generalized? 2018 international conference of the biometrics special interest group (BIOSIG) IEEE
[2] Agarwal Shruti 2020 Detecting deep-fake videos from appearance and behaviour 2020 IEEE international workshop on information forensics and security (WIFS) IEEE
[3] Karras Tero 2017 Progressive growing of gans for improved quality, stability, and variation arXiv preprint
[4] Chang H T et al 2009 Image authentication with tampering localization based on watermark embedding in the wavelet domain Optical Engineering 48(5):pp 057002-057002
[5] Zheng Y and Vrizlynn L L 2017 Automated face swapping and its detection 2017 IEEE 2nd international conference on signal and image processing (ICSIP) IEEE
[6] Marra Francesco et al 2018 Detection of gan-generated fake images over social networks 2018 IEEE conference on multimedia information processing and retrieval (MIPR) IEEE
[7] Chollet François 2017 Xception: Deep learning with depthwise separable convolutions Proceedings of the IEEE conference on computer vision and pattern recognition
[8] Mo H Chen B and Luo W 2018 Fake faces identification via convolutional neural network Proceedings of the 6th ACM workshop on information hiding and multimedia security
[9] Dang L Minh, et al 2018 Deep learning based computer generated face identification using convolutional neural network." Applied Sciences 8(12): p 2610
[10] Dataset from https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection