Deep Fake Image Detection Based on Pairwise Learning

Chih-Chung Hsu, Yi-Xiu Zhuang, Chia-Yen Lee

Published: 3 January 2020

Baran Deniz Korkmaz

August 2020

Abstract

Conventional image forgery detectors fail to recognize fake images generated by the GAN-based generator since these images are generated and manipulated from the source image. In this paper, a deep learning-based approach for detecting the fake images by using the contrastive loss is proposed.

The proposed method can be divided into the following subparts:

1. Several state-of-the-art GANs are employed to generate the fake–real image pairs.
2. The reduced DenseNet is developed to a two-streamed network structure to allow pairwise information as the input.
3. The proposed common fake feature network is trained using the pairwise learning to distinguish the features between the fake and real images.
4. A classification layer is concatenated to the proposed common fake feature network to detect whether the input image is fake or real.

Conclusion: The authors concludes that the experimental results demonstrated that the proposed method significantly outperformed other state-of-the-art fake image detectors.

Keywords (Author's Selection): forgery detection, GAN, contrastive loss, pairwise learning

1 Introduction

Recently, deep learning-based generative models, such as **variational autoencoders** and **generative adversarial networks** (GANs), have been widely used to synthesize the photo-realistic partial or whole content of an image or a video.

In the traditional image forgery detection approaches, two types of forensics schemes are commonly used:

1. Active Schemes: Externally additive signal (i.e. Watermark)
2. Passive Schemes: Statistical Information Within the Source Image

The passive image forgery detectors cannot be used to identify fake images generated by the GANs because they are synthesized from the low-dimensional random vector. Specifically, the fake images generated by the GANs are not modified from their original images.

Recently, the deep learning-based approached for fake image detection using supervised learning has been studied. In other words, fake image detection has been treated as a binary classification problem (i.e., fake or real image).

1. The convolution neural network (CNN) network was used to develop the fake image detector [9,10]
2. In [11], the performance of the fake face image detection was further improved by adopting the most advanced CNN–Xception network [12].
3. In [13], a manipulated face detection algorithm was proposed based on a hybrid ensemble learning approach.

NOTE: However, none of these studies has investigated the fully generated image, but instead, they have been focused only on partial manipulation of face images; thus, they cannot be used to detect the fully generated fake images.

NOTE: To develop a fake image detector, it is necessary to collect all of the GAN's images as the training set for deep neural networks to achieve the promising performance. However, it is difficult and very time-consuming to collect the training samples generated by all the GANs. In addition, such a supervised learning strategy [9–11] tends to learn the discriminative features of fake images generated by all the GANs, and as a result, the learned (trained) detector may not have a good generalization ability.

To meet the current requirement for the GANs-based generator of fake image detection, we propose a modified network structure, including a pairwise learning approach, called the common fake feature network (CFFN). By using the pairwise learning, the proposed structure overcomes the shortcomings of the supervised learning-based CNNs, such as those presented in [9,11]. To verify the effectiveness of the proposed method, the proposed deep fake detector (DeepFD) is applied to identify the fake face and generic images. The main contributions of this work are as follows.

- A fake face image detector based on the novel CFFN, consisting of an improved DenseNet backbone network and Siamese network architecture, is proposed.
- The cross-layer features are investigated by the proposed CFFN, which can be used to improve the performance.
- The pairwise learning approach is used to improve the generalization property of the proposed DeepFD.

2 Fake Face Image Detection

The proposed two-step learning method that combines the CFF based on pairwise learning strategy and the classifier learning.

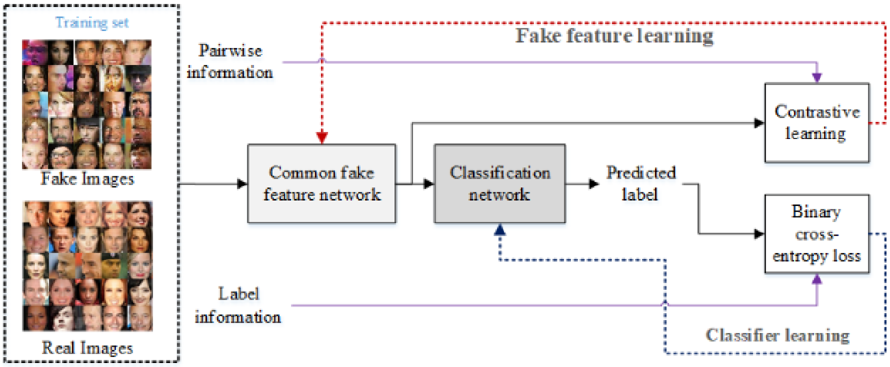

Figure 1: The flowchart of the proposed fake face detector based on the proposed common fake feature network with the two-step learning approach.

The supervised learning strategy in the fake face image detection causes two following problems:

1. Difficult collection of training samples generated by all possible GANs
2. The need to retrain the network to obtain an efficient model for the fake face images generated by a new GAN

To overcome these problems, the fake and real images are paired and follow by using the pairwise information to construct the contrastive loss to learn the discriminative common fake feature (CFF) by the proposed CFFN. Once the discriminative CFF is learned, the classification network captures the discriminative CFF to identify whether the image is real or fake.

The pairwise information is necessary for the training stage so that the CFFN can learn the discriminative CFFs well. Toward this end, the pairwise information can be generated from the training set X and its corresponding label set Y by the permutation combination. Therefore, there are C $(N_T,2)$ pairs P = [$p_{i=0,j=0}, p_{i=0,j=1}, \ldots, p_{i=0,j=N_r}, \ldots, p_{i=N_f,j=N_r}$] generated from the training samples.

2.1 Common Fake Feature Network

Many advanced CNN can be used to learn the fake features from the training set.However, most of these advanced CNNs are trained in a supervised way, so the classification performance depends on the training set. Rather than learn the fake features from all the GANs' images, we seek the CFF over different GANs.

In this way, a suitable backbone network is needed for learning CFFs. However, the traditional CNNs (e.g., the DenseNet [19]) are not designed to learn the discriminative CFF. To overcome this shortcoming, we propose integrating the Siamese network with the DenseNet [19], developing the CFFN to achieve the discriminative CFF learning.

Once the backbone network is trained to have the best feature representation ability, the performance of the fake image recognition can be improved as well. To this end, DenseNet is selected as a backbone network of the proposed CFFN. In addition, the cross-layer features are integrated into the classification layer to improve the fake image recognition performance.

In general, the classification of fake image can be performed by a different classification learning model, such as random forest, SVM, or Bayes classifier. However, the discriminative feature may be further improved by applying the back-propagation algorithm to the end-to-end structure. Therefore, in this work, the convolution and fully connected layers are concatenated into the last convolution layer of the proposed CFFN to obtain the final decision result.

2.1.1 Discriminative Feature Learning

To enhance the performance of the proposed method, we introduce contrastive loss to learn the CFFs by pairwise learning. Therefore, the Siamese network structure [22] is used for allowing the pairwise inputs.

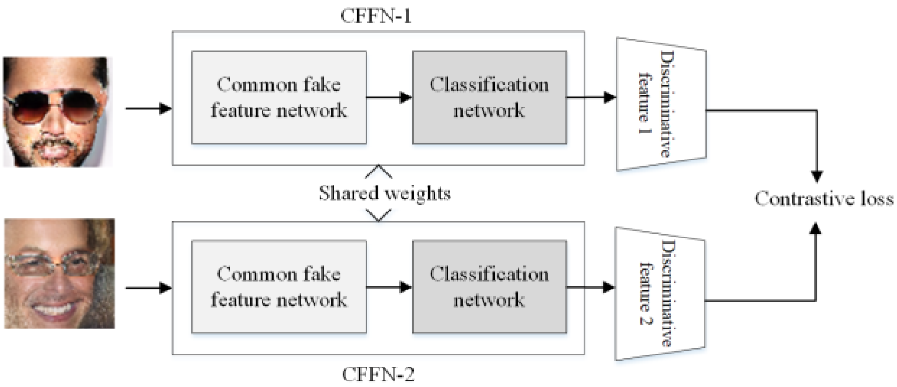

Figure 2: The proposed pairwise learning based on the Siamese network and contrastive loss.

With the aim to make the proposed CFFN learn the discriminative features during the training process, the contrastive loss term is incorporated into the energy function of the traditional loss function for supervised learning (i.e., the cross-entropy loss). Afterward, given the face image pair (x 1 , x 2) and the pairwise label y, where y = 0 indicates an impostor pair, and y = 1 indicates a genuine pair, the energy function between two images is defined as:

$$E_W(x_1, x_2) = ||f_{CFFN}(x_1) - f_{CFFN}(x_2)||_2^2 \tag{1}$$

The contrastive loss is introduced to learn the discriminative feature representation as well as to avoid constant mapping, which can be expressed as:

$$L(W, (P, x_1, x_2)) = 0.5 * (y_{ij} E_w^2) + (1 - y_{ij}) * max(0, (m - E_w)_2^2) \tag{2}$$

where m denotes the predefined threshold value.

2.1.2 Classification Learning

To improve the performance of the fake face image detection, we adopt a subnetwork as a classifier. Thus, the classification learning can be quickly learned by the cross-entropy loss function, which is given by:

$$L_c(x_i, p_i) = -\sum_{i}^{N_T}(f_{CLS}(f_{CFFN}(x_i))log(p_i) \tag{3}$$

The CFFN is first trained by the proposed contrastive loss and follows by training the classifier based on cross-entropy loss.

3 Source

- Deep Fake Image Detection Based on Pairwise Learning; Chih-Chung Hsu, Yi-Xiu Zhuang, Chia-Yen Lee

4 Referenced Sources

1. **Extrinsic Feature Analysis - Watermark** - [6] - Chang, H.T.; Hsu, C.C.; Yeh, C.H.; Shen, D.F. Image authentication with tampering localization based on watermark embedding in wavelet domain. Opt. Eng. 2009, 48, 057002.

2. **Intrinsic Feature Analysis - Statistical Information** - [7] - Hsu, C.C.; Hung, T.Y.; Lin, C.W.; Hsu, C.T. Video forgery detection using correlation of noise residue. In Proceedings of the IEEE Workshop on Multimedia Signal Processing, Cairns, Australia, 8–10 October 2008; pp. 170–174.

3. **Intrinsic Feature Analysis** - [8] - Farid, H. Image forgery detection. IEEE Signal Process. Mag. 2009, 26, 16–25.