Learning to Detect Fake Face Images in the Wild

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

Published: December 2018

Baran Deniz Korkmaz

August 2020

Abstract

Objective: Developing a deep forgery discriminator (DeepFD) to detect deepfake images.
Technique: Directly learning a binary classifier is relatively tricky since it is hard to find the common discriminative features for judging the fake images generated from different GANs. To address this shortcoming, the authors adopt **contrastive loss** in seeking the typical features of the synthesized images generated by different GANs and follow by concatenating a classifier to detect such computer- generated images.
Conclusion: Experimental results demonstrate that the proposed DeepFD successfully detected 94.7% fake images generated by several state-of-the-art GANs.
Keywords (Author's Selection): forgery detection, GAN, contrastive loss, fully convolutional network

1 Introduction

To address the issue of image forgery detection, there are two different categories in the **traditional** approach:

1. Extrinsic Feature Approach: The first strategy aims to detect fake images by embedding external unique signals into the original images (e.g. digital watermarking)

2. Intrinsic Feature Approach: The second strategy aims to detect fake images by discovering the intrinsic and invariant features from the original images. The forgery image should be able to detected by checking the statistical property of the extracted intrinsic feature from the received image because any tampered operation will make the intrinsic feature changed.
Objective: Finding the unusual statistical properties of images to detect whether it is a forgery or not.
Related Properties: See [2],[3], and [13] in the article for further investigations.

(a) Sensor Pattern Noise

(b) Double Compression Cues

Problem: Traditional forgery detection techniques are hard to detect the generated images by GANs since their image content are made by deep neural network directly.
Idea: A deep neural network called deep forgery discriminator (DeepFD) based strategy to effectively and efficiently detect the generated / fake images synthesized by GANs or other advanced networks.
Another challenge in deepfake detection is that, in general, it is hard to collect training images generated by all possible GANs or image synthesizers. Moreover, there are many new GANs proposed every year. Such strategy needs to re-train the classifier to keep its performance when there is a new GAN proposed.
Solution: To ensure the performance of the proposed DeepFD, we tend to learn the jointly discriminative features from collected training images across different GANs by introducing the contrastive loss into the network learning framework [12]*. Contrastive loss learns such joint features from heterogeneous training images by introducing the pairwise information so that the DeepFD should be able to effectively distinguish any fake image generated by any GAN. Besides, the proposed DeepFD can further localize unrealistic details of the fake image based on a fully convolutional architecture. We can follow such regions to further improve the performance of the proposed DeepFD.

2 The Proposed Deep Forgery Discriminator

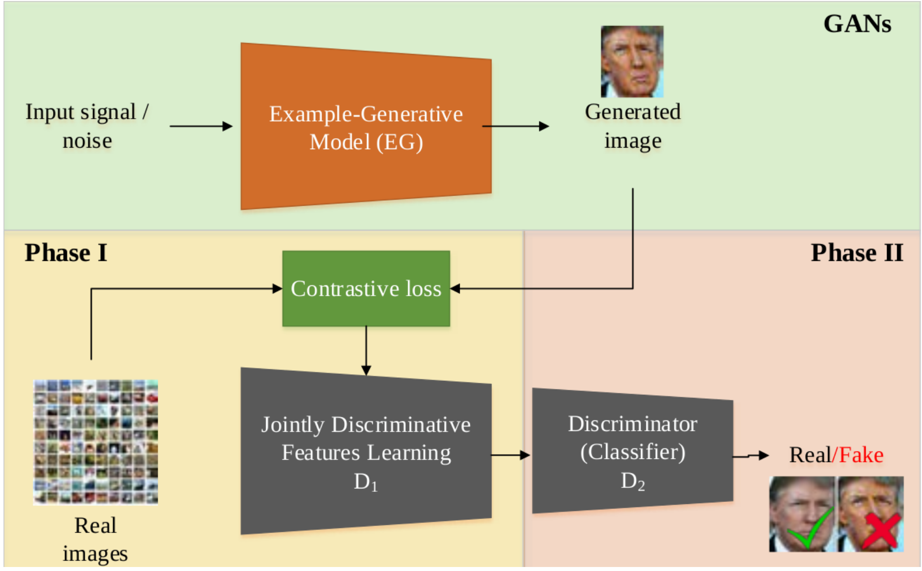
GANs
Input signal / noise
Example-Generative Model (EG)
Generated image
Phase I
Phase II
Contrastive loss
Jointly Discriminative Features Learning D1
Discriminator (Classifier) D2
Real/Fake
Real images

Figure 1: NN Architecture

The proposed method is composed of two stages:

1. We collect a lot of fake images synthesized by several GANs called example-generative model and real images to learn the jointly discriminative features D_1 based on the proposed contrastive loss.
2. A discriminator (classifier) D_2 will be concatenated to the D_1 to further distinguish fake images.

NOTE: The classifier D_2 is directly concatenated to the 4th layer of the jointly discriminative feature learning network D_1.

Please refer to **Table-1** for the details of network architectures of D_1 and D_2.

2.1 Stage 1: Jointly Discriminative Feature Learning

2.2 Stage 2: Classifier Training

3 Source

- Learning to Detect Fake Face Images in the Wild; Chih-Chung Hsu, Yi-Xiu Zhuang, Chia-Yen Lee

4 Referenced Sources

1. **Intrinsic Feature Analysis** - [2] - 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.

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

3. **Contrastive Loss** - [12] - E. Simo-Serra, et al. ”Discriminative learning of deep convolutional feature point descriptors,” Computer Vision (ICCV), 2015 IEEE International Conference on. IEEE, 2015.

4. **Intrinsic Feature Analysis - Double Compression** - [13] - Y.L. Chen and C.T. Hsu. ”Detecting doubly compressed images based on quantization noise model and image restoration,” in Proc. of IEEE International Workshop on. Multimedia Signal Processing, Oct. 2009.