Automatic detection of photorealistic images through deep learning techniques

Evangelos Skoulas
Department of Electrical and Computer Engineering
University of Patras

Abstract

Photorealistic image detection is the problem of distinguishing between real life photographs and artificially generated ones. The high quality images that are being produced nowadays through GANs need an automated way to be detected. Therefore, a method was proposed based on a combination of image processing tools and deep learning techniques. A preprocessing stage was designed that used the discrete cosine transform (DCT) alongside a rearrangement of its coefficients to provide a new image format in the frequency domain. The resulting images were used to train a simple convolutional neural network. Through a series of experiments, testing on a variety of datasets and under different conditions a handful of interesting observations and results were derived.

Introduction

The evolution of generative adversarial networks (GANs) has resulted in synthetic images that are as realistic as ever. Nowadays, even the most meticulous observer can be deceived by their realistic nature as distinguishing between photographic and generated images is almost impossible. As a result, the trustworthiness of images as a medium that carries information is jeopardized. Under these circumstances, it is essential to design automated systems that detect photorealistic images successfully. To achieve the highest impact possible it was important to assess the capabilities of humans to carry out the task at hand while identifying the cases in which our optical system is at its' most vulnerable. The aforementioned process concluded that modern machine learning advances can produce incredibly realistic data by replicating with precision the 'building blocks' of reality. Humans have proven particularly unreliable at distinguishing between fabricated and real data.

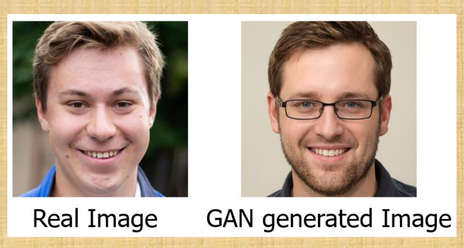

Figure 1. Instances from Lagos et al dataset [1]

Proposed Method

The designed procedure consists of 1)Data collection, 2)Image preprocessing, 3)CNN training, 4)Model testing
The proposed method's core is the preprocessing phase that can be split into 4 stages:

- Application of the blockwise DCT to the images in blocks of size 8x8 same as JPEG compression
- Grouping in different vectors the homonymous DCT coefficients aka the ones that are in the corresponding positions within the different blocks of the image
- Rearrangement of the DCT coefficients of each group to create essentially compressed versions of the initial image using a single coefficient at a time
- Stacking of those 'sub-images' next to one another

Experiments and Results

Firstly, the FFHQ dataset provided 10000 real images of human portraits while the StyleGAN2 architecture produced 10000 synthetic faces. A CNN model was trained using Root Mean Square propagation (RMSprop) algorithm for 30 epoch on 3 different levels of image compression. Specifically, 10, 28 and 64 DCT coefficients were selected out of the 64 available of the 8x8 block of the images. The accuracy achieved on each case can be seen on Table 1. Furthermore, an evaluation was done on the performance of the method on identifying with data from similar GAN architectures. In this case, 3 new models were trained using shrinked versions of the FFHQ and StyleGAN2 images through 3 'downsampling' methods (AverageOf4, Bilinear Interpolation and Nearest Neighbour Interpolation). The best performing models were tested on the Lagos et al dataset that contains 150 real portraits provided by the FFHQ dataset along side 50 synthetic images from StyleGAN2, 50 coming from the StyleGAN architecture (its' predecessor) and 50 from ProGAN. The performances on each GAN architecture can be seen in Table 2.

Table 1. Evaluation on data derived from the same GAN architecture.

DCT Coefficients kept	Testing Accuracy
10	99,7%
28	99,9%
64	99,925%

Table 2. Evaluation on data derived from similar GAN architecture.

Shrinking Method	4-pixel Average	Nearest-Neighbour Interpolation	Bilinear Interpolation
FFHQ	92/150	142/150	128/150
ProGAN	10/50	3/50	1/50
StyleGAN	31/50	33/50	35/50
StyleGAN2	32/50	44/50	43/50
TOTAL	165/300 (55%)	222/300 (74%)	207/300 (69%)

Discussion

The proposed method can be highly effective in GAN generated image detection of the current generation.

1. It scores incredibly high on data that are similar to the training data
2. It generalizes on similar GAN architectures such as StyleGAN
3. The method cannot generalize in totally 'unknown' data without further modifications

In terms of resource management it is important to note that:
- Less training data were required to achieve respectable levels of performance
-Smaller neural networks seemed to perform equally well if not better when dealing with images in the frequency domain compared to the image domain

References

[1] Federica Lago, Cecilia Pasquini, Rainer Boehme, Helene Dumont, Valerie Goffaux, and Giulia Boato. More real than real: A study on human visual perception of synthetic faces [applications corner]. IEEE Signal Processing Magazine, pages 109–116, 2021.
[2] Olivia Holmes, Martin S Banks, and Hany Farid. Assessing and improving the identification of computer-generated portraits. ACM Transactions on Applied Perception (TAP), pages 1–12, 2016.
[3] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In International conference on machine learning, pages 3247–3258. PMLR, 2020

Figure 2. GAN generated images over the years

Contact

Evangelos Skoulas
Department of Electrical and Computer Engineering
Evanskoulas99@gmail.com