Before we start, let’s talk about the problem of current deep fake detection models. The problem is that to train the model, we should generate deep fake image input and to generate deep fake image, we should train new deep fake generation model. This process of training deep fake generation model for just making input, costs lots of time and resources.

To solve this problem, this paper suggested one idea.

Instead of using “real” deep fake images for training, use “fake” deep fake images which are normal images that have property of deep fake images.

And, this property of deep fake images is Distinct artifacts in the images.

Then, why deep fake images have distinct artifacts? The answer is in the deep fake generation algorithm as shown in this figure.

During the first step of generation, deep fake generation model detects the face region in source image.

Then it detects face landmarks from face region

After that, with landmarks from previous step, the model computes the transform matrix that normalize these landmarks into fixed size of square.

Using both inverse of computed transform matrix and another normalized face image for overwriting, we can generate new face region image that can be fitted into source face image.

However, because of the limitation of computation resources and production time, to match the configuration of source’s face, new face image should undergo an affine warping and this leaves distinct artifacts due to resolution inconsistency between warped face area and surrounding context.

So now, we know that deep fake images can have several distinct artifacts.

Then how could this paper can only use normal images as inputs of deep fake detection model?

The main point is that we add some distinct artifacts in normal images, by blurring subregion of face image and then affine warp back it to real image.

This fabricated images with distinct artifacts can be think of as deep-fake image and the model can distinguish deep fake images property by using this fabricated images as inputs.

Actually, there are quite many deep fake generation models that are different in resolution and warping technology which can be think of as post-processing of synthesized images.

To handle all images from different deep fake model, this paper changed color information such as brightness, contrast. The paper also changed the shape of affine warped face area to be robust to post-processing methods of different models.

So how did this paper actually train their model?

Firstly, they choose region of interest of positive, negative examples. This region chosen as rectangle areas that contain both face landmarks and it’s surrounding area. These region of interest resized to 224 x 224 to feed to the CNN models for training. For inference, the paper cropped region of interest of each training sample by 10 times and average predictions of all cropped images as the final fake probability.

With this idea and simple image classifying model like VGGnet or resnet, this paper successfully reached to state of the art of deep fake detection technology at 2018.