This paper provides a deep-learning method for separating computer-generated graphics from real photographic images. The proposed method used a Convolutional Neural Network (CNN) with a custom pooling layer to optimize the current best-performing extraction feature algorithms. Local estimates of category probabilities are calculated and aggregated to determine the overall picture tag. They tested their work on recent photo-realistic computer graphics and proved that it outperforms state-of-the-art methods for both local and total object classification.

Some recent advances in image processing, such as facial re-enactment face2face, show the importance of providing some tools to differentiate computer graphics (CGs) from original photographic images (PGs). Although the difference between CG and PG depends not only on the properties of the object but also on the mental characteristics of viewers, people have an incorrect ability to distinguish between altered and unaltered objects.

Deep-learning approaches, and particularly CNN, have recently become very popular with many computer vision problems: classification, denoising, steganalysis, etc. CNN is not only capable of learning the limits of classification, but also of finding the best representation of the data together. They also noted that the overall structure is tailored to their problem: convolution layers model filters, while densely connected layers can effectively replace other classification methods. Motivated by these observations, they implemented a particular pooling layer to extract statistical features within the CNN framework which was later trained to distinguish CG from PG.

Usually, after convolution layers, maximum local pooling is calculated to reduce the size of the data representation before classification. As far as photo forensics is concerned, specific international numerical quantities are considered to be more useful. A particular pooling layer is therefore developed to adapt neural nets to this particular task.

Their CG was downloaded from the Level-Design Reference Database[28], which contains more than 60,000 proper resolution (1920×1080 pixels) video game screenshots in JPEG format. Only five different video games were deemed photo-realistic enough to reduce the collection to 1800 images. Also, in-game information (dialogs, life bars, mini-maps) has been removed by cropping up images.

Photographic images are high-resolution images (about 4900 × 3200 pixels) taken from the RAISE dataset[29], directly converted from RAW to JPEG. Images from both categories cover a wide range of outdoor and indoor scenes: landscapes, human bodies, and heads, human-made objects (e.g., buildings, cars), etc.

From those 3600 images, they created three databases on which their tests were carried out. First, they selected the green channel for each photo. Of category was divided into learning (70 percent), testing (20 percent), and evaluation (10 percent) to create a full-size server. From this original file, they created a smaller one size by cropping that image to 650×650. Eventually, they randomly extracted 43000 patches, 100 × 100 in length, to train the patch classifier.

They identified a novel method for classifying computer graphics and real photographic images that incorporates a statistical function of extraction into CNN frameworks and finds the best features for an efficient boundary. The boosting approach is used to quantify local tag estimates. To challenge their algorithm with today's computer graphics, they've compiled a photo-realistic video-game collection of screenshots. This work is linked with state-of-the-art methods.