**Review on: Detection of GAN-generated Fake Images over Social Network**

Introduction and Related work:

Fake images are all over social media and often it is impossible for human beings to detect them let alone tell which parts have been doctored. Image-to-image translation is quite a general problem where some domain are mapped to corresponding images of another domain. GAN have recently shown massive results in this task. The GAN network is trained with large number of paired images. However, in some cases these correspondences often unknown, and no pairs of images are usually there for training. Adversarial training paradigm have two actors: the image-to-image network (generator) (1), and a support network (discriminator) (2). The discriminator tries to tell apart real images from those created by the generator, while the Generator is trained to deceive the discriminator [23] (1).

Methodology:

In order to train and validate the detectors under comparison, the authors built a dataset of samples of different categories from image-to-image translation [10] using the code available on-line

(https://github.com/junyanz/CycleGAN). For each category, the dataset includes both the real and the fake images. The photos are divided into further two groups where one group has the translations of natural images and the second has photos of buildings, landscapes etc.

To assess the performance in real life scenario, a leave-one-manipulation-out (LOMO) procedure was adopted. Therefore, at each iteration, all images referring to a certain category were set aside for validation while the remaining ones were used for training. Classifier did not adapt to features of a specific class of translations, but learnt patterns that are shared by all images generated by the procedure mentioned.

Scenario one was on the original uncompressed dataset where the highest average accuracy was shown by the by Cozzolino2017. This shallow network provides near-perfect classification for all manipulations except winter2summer, for an average of 95.07%. Scenario two was on twitter-like compressed images with training mismatch where XceptionNet provided an accuracy of 87.17%. Scenario three was on twitter-like compressed images where XceptionNet provided the best performance, reaching an accuracy of 89.03%.

Baseline/ Testing:

The authors have compared a number of methods for the detection of image-to-image translation such as GAN discriminator (3), Steganalysis (4), Cozzolino2017 (2), Bayar2016 (5), Rahmouni2017 (6), DenseNet (7), InceptionNet v3 (8), and XceptionNet (9).

Dataset and Accuracy:

It was carried out on a dataset of 36302 images, detection accuracy was up to 95% was achieved by both conventional and deep learning detectors, but only deep learning detectors provided a high accuracy, up to 89%, on compressed data.

Robustness is better preserved by deep learning networks, especially XceptionNet. Which works well even in the presence of training-test mismatching.

Reference:

Detection of GAN-generated Fake Images over Social Networks by Francesco Marra, Diego Gragnaniello, Davide Cozzolino, Luisa Verdoliva

D. Cozzolino, G. Poggi, and L. Verdoliva, “Recasting residual-based local descriptors as convolutional neural networks:an application to image forgery detection,” in ACM Workshop on Information Hiding and Multimedia Security, 2017, pp. 1–6.

J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in IEEE ICCV, 2017.

J. Fridrich and J. Kodovsk´y, “Rich Models for Steganalysis of Digital Images,” IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, pp. 868–882, Jun. 2012.

B. Bayar and M. Stamm, “A deep learning approach to universal image manipulation detection using a new convolutional layer,” in ACM Workshop on Information Hiding andMultimedia Security, 2016, pp. 5–10.

N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizeny, “Distinguishing computer graphics from natural images using convolution neural networks,” in IEEE WIFS, 2017, pp. 1–6.

G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger, “Densely connected convolutional networks,” in IEEE CVPR 2017, pp. 4700–1708.

C. Szegedy,W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in IEEE CVPR, 2015.

F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in IEEE CVPR, 2017, pp. 1800–1807.

J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in IEEE ICCV, 2017.

References:

1. Detection of GAN-generated Fake Images over Social Networks by Francesco Marra, Diego Gragnaniello, Davide Cozzolino, Luisa Verdoliva Dep. of Electrical Engineering and Information Technology University Federico II of Naples