Shayan Taheri's Revisions:

The following revision items are provided based on the comments of reviewers as well as the proofread suggestions:

1) **System Description and Contributions**: We propose a novel defense system for detection of unseen and unknown adversarial examples. Our system includes eight major elements including, (1) original security dataset, (2) a victim neural network connected to three attack engines for generation of adversarial examples, (3) reference adversarial security dataset, (4) an image-to-image translation unit (called Pix2Pix) connected to the victim neural network, (5) synthetic adversarial security dataset, (6) a retraining mechanism performed on the victim neural network connected to the attack engines, (7) a weight extraction mechanism from the same victim neural network, and (8) an weight update mechanism applied on the victim neural network connected to Pix2Pix. The same architecture is used for the mentioned victim neural networks and the attack engines are Fast Gradient Sign Attack (FGSM), DeepFool, and projected gradient descent (PGD). All these elements are linked to each other with a certain flow in order to accomplish the tasks of attack and defense execution. One of the challenges during the system assembly and implementation was connection of the victim neural network to the Pix2Pix unit. The connection and transmission of data between these two occur through inclusion of the loss from the victim neural network (called adversarial loss) in the system loss function along with the generative adversarial network (GAN) loss and the L1 loss. We apply a specific weight to each of these losses for tuning the training procedure. There is another change on the Pix2Pix software as well that introduces the weight for the victim neural network during the training. Another part of the system to describe is evaluation of the generated data by Pix2Pix from both attack and defense perspective. In order to evaluate the attack strength of the synthetic data, the Pix2Pix tests the victim neural network trained on the original adversarial data on the generated data. With respect to the defense assessment, the trained victim neural network on the Pix2Pix data is tested on the original adversarial data. Linking all these elements to each other and performing system performance measurement is another accomplished objective in this work. The developed system can be incorporated in diverse Cyber-physical and IoT systems.

2) **Title:** AEG-Pix2Pix: Protecting Classifiers Against Adversarial Attacks Using Pix2Pix

   Note: Adversarial Example Generation (AEG)

3) **System Generalization**: Now, we discuss how the AEG-Pix2Pix system is generalizable. All the major elements in the system architecture are subject to change. For future work, we can examine different types of data with various number of sample for system evaluation. Our initial findings in development of our system showed changing the number of samples in the training and the testing datasets can significantly vary the system behavior. In order to further assess the system, other well-known and newly introduced attacks can be introduced into the system. Also, we plan to design our own attacks for fooling the victim neural network. One possible option for this plan is introducing different types of noises (such as Perlin noise) randomly and with minor adjustment into images from multiple domains and observe their malicious payload. In addition, manually crafted mathematical units can be included in the design with the goal of

enhancing the existing attacks. Mimicking the distribution of the original data can be considered in crafting these units. Many of the existing mathematical approaches such as gradient descent and L-norm optimization can be combined in an attack unit. The ultimate goal is maximizing the difference between the original images and the original or the newly generated adversarial images in every iteration. Meanwhile, applying regularization on our adversarial samples helps in transferring them on different network models. The architecture for victim neural network can be changed. We used a shallow neural network in this work, while Pix2Pix may incorporate deep neural networks for generation of possible stronger attacks. The base version of Pix2Pix was used in our work, which can be changed to other ongoing changes to the unit architecture. The retraining, the weight extraction, and the weight update mechanisms may be enhanced with advanced algorithms instead of being direct and simple. The defense ability of the system can be improved by adding noise removal units before feeding the data during the testing classification.

4) Tensorflow. tensorflow/cleverhans, 2020: https://github.com/tensorflow/cleverhans or http://www.cleverhans.io.