METHOD FOR DETECTION OF ADVERSARIAL ATTACKS ON FACE DETECTION NETWORKS

V.F. Konovalov, E.V. Myasnikov

Attacks

FGSM

$$x^{adv} = x - \epsilon \, \mathrm{sign}\left(\nabla_x J\left(x, y_{true}\right)\right) \quad (Eq.1)$$

Simple and popular attack requiring full knowledge and access to the model. Parameter epsilon controls perturbation strength.

MI-FGSM

$$x_{n+1}^{adv} = x_n^{adv} + \alpha \, \mathrm{sign}\left(g_n\right) \quad (Eq.2)$$

$$g_{n+1} = \mu g_n + \frac{\nabla_{x_n^{adv}} J\left(x_n^{adv}, y_{true}\right)}{\left\|\nabla_{x_n^{adv}} J\left(x_n^{adv}, y_{true}\right)\right\|_1}$$

Attack that improves on FGSM by introducing iterability and momentum. Momentum allows to overcome local optimums while iterability helps avoid jumping over optimal values.

Fig. 1. Comparison between fragments of heavily perturbed and clean images. Grainy, high-frequency structure is visible.

Defences

Baseline algorithm

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{k} w_i \left(x_i - \hat{x}_i\right) \quad (Eq.\ 3)$$

$$\hat{x} = \begin{bmatrix} -0.25 & 0.5 & -0.25 \\ 0.5 & 0 & 0.5 \\ -0.25 & 0.5 & 0.25 \end{bmatrix} * x \qquad w_i = \left(\left(\sigma_{local}\right)^2 + 5\right)^{-1}$$

Baseline algorithm (*Eq. 1*) calculates the weighted sum of difference between the pixels in the neighborhood. Thresholding is used to determine if image was attacked or not.

Proposed algorithm

$$\hat{p} = w_i \left(x_i - \hat{x}_i\right) \quad (Eq.\ 4) \qquad w_i = \left(\left(\sigma_{local}\right)^2 + 5\right)^{-1} \quad (Eq.5)$$

$$\hat{x} = \begin{bmatrix} -0.25 & 0.5 & -0.25 \\ 0.5 & 0 & 0.5 \\ -0.25 & 0.5 & 0.25 \end{bmatrix} * x$$

Transform both images into YcbCr, which removes dependency on luminosity channels. Separate CbCr, compute «approximate noise» using Eq.4. Resulting «approximate noise» is normalized using interquantile algorithm, bringing it to zero mean and unit variance and dropping 25% lowest and highest values. Result is binned into a histogram, ranging [-5.1;5.1] with a bin width of 0.4. Resulting histogram is divided by number of pixels to acquire PDF as shown on Fig.1.

Proposed algorithm visualization

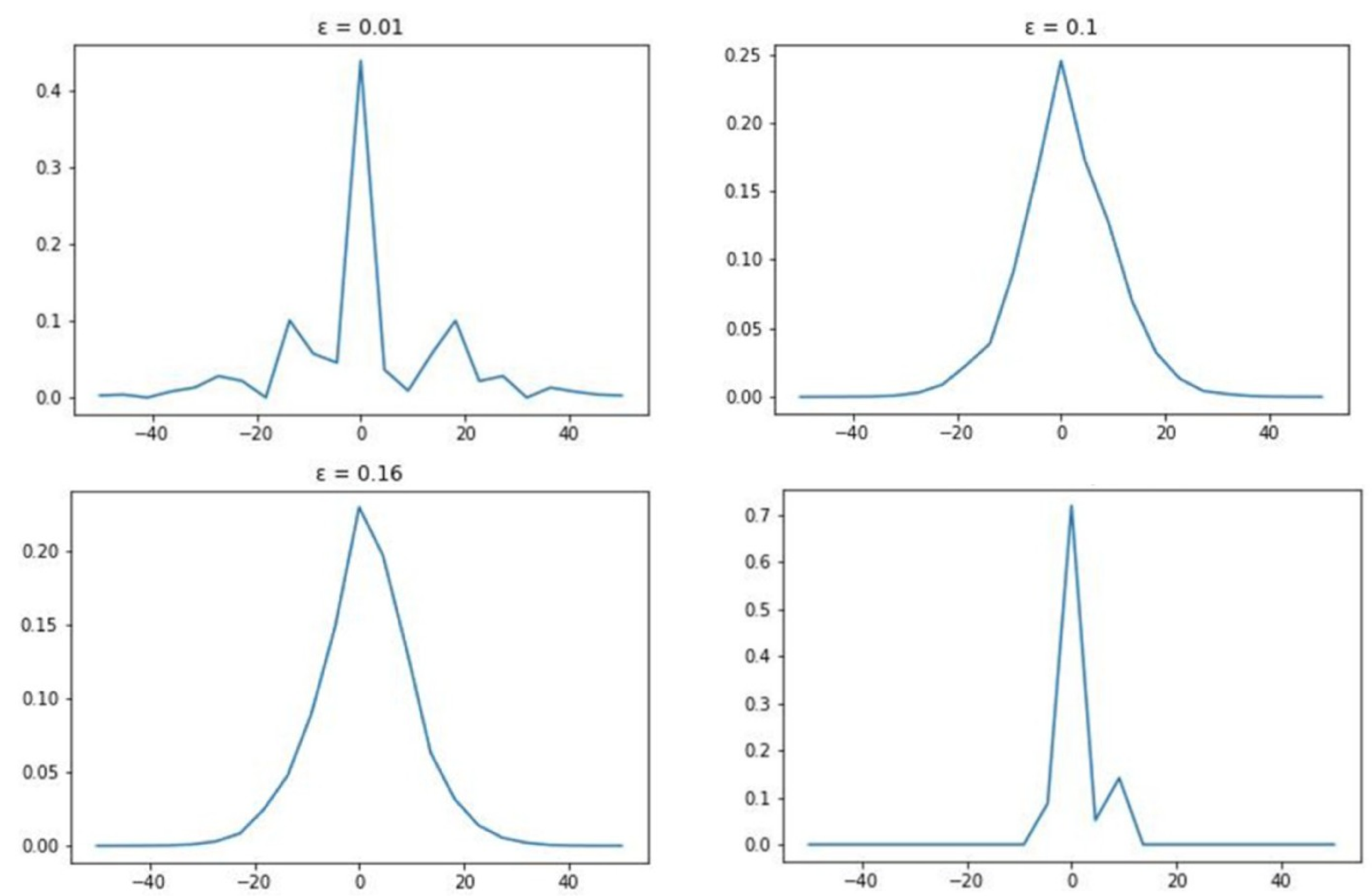

Fig. 2. Distributions of values acquired using proposed algorithm. First three graphs show distributions for cases where original images x are the images perturbed with FGSM under different ε, the last graph — clean images. As it can be seen, distributions for clean and perturbed images are easy to distinguish. By summing them up, *Eq.3* and therefore baseline algorithm loses that additional information. The hypothesis for this difference is, clean, natural images have lower spatial frequencies.

Experiments

Experiments were conducted under two conditions: compressed and uncompressed attacked images, for two feature extraction/detection algorithms — baseline and proposed.
For the FGSM attack, ε was varied from 0.01 to 0.20, for MI-FGSM — from 0.02 to 0.20 with a step of 0.02, weight decay of 0.7, and α=127.5ε, which helped keep MI-FGSM single step value constant for all epsilon.

TABLE I. BASELINE METHOD

	Macro F1	Macro Precision	Macro Recall
FGSM attack, compressed files	0.690	0.724	0.784
FGSM attack, uncompressed files	0.754	0.746	0.825
MI_FGSM attack, compressed files	0.526	0.696	0.572

TABLE II. SUGGESTED METHOD, FGSM ATTACK, COMPRESSED FILES

Name, window type	Macro F1	Macro Precision	Macro Recall
Cross-validation mean			
SVM, window eq. (5)	0.949	0.942	0.961
RF, window eq. (5)	0.961	0.957	0.965
MLP, window eq. (5)	0.951	0.945	0.958
SVM, avg. window	0.976	0.970	0.983
RF, avg. window	0.992	0.991	0.993
MLP, avg. window	0.976	0.971	0.982
Test dataset result			
SVM, window eq. (5)	0.905	0.865	0.965
RF, window eq. (5)	0.930	0.902	0.966
MLP, window eq. (5)	0.910	0.874	0.960
SVM, avg. window	0.957	0.932	0.987
RF, avg. window	0.980	0.969	0.992
MLP, avg. window	0.979	0.968	0.991

TABLE III. SUGGESTED METHOD, FGSM ATTACK, UNCOMPRESSED FILES

Name, window type	Macro F1	Macro Precision	Macro Recall
Cross-validation mean			
SVM, window eq. (5)	0.981	0.976	0.982
RF, window eq. (5)	0.977	0.976	0.979
MLP, window eq. (5)	0.972	0.972	0.973
SVM, avg. window	0.998	0.998	0.999
RF, avg. window	0.999	0.999	0.999
MLP, avg. window	0.998	0.998	0.999
Test dataset result			
SVM, window eq. (5)	0.962	0.943	0.983
RF, window eq. (5)	0.965	0.951	0.980
MLP, window eq. (5)	0.965	0.954	0.976
SVM, avg. window	0.999	0.999	0.999
RF, avg. window	0.999	0.999	0.999
MLP, avg. window	0.997	0.995	0.999

TABLE IV. SUGGESTED METHOD, MI-FGSM ATTACK, COMPRESSED FILES

Name, window type	Macro F1	Macro Precision	Macro Recall
Cross-validation mean			
SVM, window (5)	0.977	0.977	0.977
RF, window (5)	0.978	0.977	0.979
MLP, window (5)	0.975	0.982	0.972
SVM, avg. window	0.977	0.966	0.988
RF, avg. window	0.974	0.975	0.973
MLP, avg. window	0.969	0.977	0.962
Test dataset result			
SVM, window eq. (5)	0.978	0.979	0.978
RF, window eq. (5)	0.966	0.967	0.965
MLP, window eq. (5)	0.975	0.980	0.970
SVM, avg. window	0.977	0.966	0.990
RF, avg. window	0.973	0.972	0.975
MLP, avg. window	0.961	0.972	0.952

Conclusion

A method for detecting adversarial gradient attacks was proposed. Proposed method for feature extraction shows good results when used with any of the machine learning algorithms. Proposed method can be extended to include correction of detected attacked images.

SAMARA UNIVERSITY
34 Moskovskoye shosse, Samara, 443086
+7 (846) 335-18-26, +7 (846) 335-18-36 (fax)
ssau@ssau.ru, www.ssau.ru