ECE 595: Homework 5

Yi Qiao, Class ID 187

(Spring 2019)

Exercise 1: Adversarial Attacks on Gaussian Classifier

(a) minimum-norm attacks

(i) minimum l_2 and l_∞ attack

Since we only have 2 classes, the question becomes

$$\begin{aligned} &\underset{\boldsymbol{x}}{minimize}\ ||\boldsymbol{x} - \boldsymbol{x}_0|| \\ &subject\ to\ \boldsymbol{w}^T\boldsymbol{x} + w_0 = 0 \end{aligned} \tag{1}$$

using l_2 norm
the problem is the same as

$$\begin{aligned} &\underset{\boldsymbol{x}}{minimize}\ \frac{1}{2}||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 \\ &subject\ to\ \boldsymbol{w}^T\boldsymbol{x} + w_0 = 0 \end{aligned} \tag{2}$$

The lagrangian is

$$\mathcal{L}(\boldsymbol{x}, \lambda) = \frac{1}{2}||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 + \lambda(\boldsymbol{w}^T\boldsymbol{x} + w_0) \tag{3}$$

Taking the derivative

$$\begin{aligned} \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}, \lambda) &= \boldsymbol{x} - \boldsymbol{x}_0 + \lambda\boldsymbol{w} = 0 \\ \frac{\partial}{\partial\lambda}\mathcal{L}(\boldsymbol{x}, \lambda) &= \boldsymbol{w}^T\boldsymbol{x} + w_0 = 0 \end{aligned} \tag{4}$$

$$\begin{aligned} \lambda\boldsymbol{w} &= \boldsymbol{x}_0 - \boldsymbol{x} \\ \lambda\boldsymbol{w}^T\boldsymbol{w} &= \boldsymbol{w}^T\boldsymbol{x}_0 - \boldsymbol{w}^T\boldsymbol{x} \\ \lambda &= \left(\boldsymbol{w}^T\boldsymbol{w}\right)^{-1}(\boldsymbol{w}^T\boldsymbol{x}_0 + w_0) \end{aligned} \tag{5}$$

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{x}_0 - \lambda\boldsymbol{w} \\ &= \boldsymbol{x}_0 - \frac{\boldsymbol{w}(\boldsymbol{w}^T\boldsymbol{x}_0 + w_0)}{||\boldsymbol{w}||_2^2} \end{aligned} \tag{6}$$

using l_∞ norm

$$\begin{aligned} &\underset{\boldsymbol{x}}{minimize}\ ||\boldsymbol{x} - \boldsymbol{x}_0||_\infty \\ &subject\ to\ \boldsymbol{w}^T\boldsymbol{x} + w_0 = 0 \end{aligned} \tag{7}$$

Let $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{x}_0$, $b_0 = -(\boldsymbol{w}^T\boldsymbol{x}_0 + w_0)$, the problem becomes:

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmin}||\boldsymbol{x} - \boldsymbol{x}_0||_\infty \\ &subject\ to\ \boldsymbol{w}^T\boldsymbol{r} = b_0 \end{aligned} \tag{8}$$

The lagrangian is

$$\mathcal{L}(\boldsymbol{r}, \lambda) = ||\boldsymbol{r}||_\infty + \lambda(b_0 - \boldsymbol{w}^T\boldsymbol{r}) \tag{9}$$

Taking derivative,

$$\frac{\partial}{\partial\lambda}\mathcal{L}(\boldsymbol{r}, \lambda) = b_0 - \boldsymbol{w}^T\boldsymbol{r} = 0 \tag{10}$$

By Holder's Inequality:

$$\begin{gathered} |b_0| = |\boldsymbol{w}^T\boldsymbol{r}| \leq ||\boldsymbol{w}||_1||\boldsymbol{r}||_\infty \\ ||\boldsymbol{r}||_\infty \geq \frac{|b_0|}{||\boldsymbol{w}||_1} \end{gathered} \tag{11}$$

Consider $\boldsymbol{r} = \eta \cdot sign(\boldsymbol{w})$, for some constant η tbd. We can show that

$$||\boldsymbol{r}||_\infty = \underset{i}{argmax}\ |\eta \cdot sign(w_i)| = |\eta| \tag{12}$$

let $\eta = \frac{b_0}{||\boldsymbol{w}||_1} \cdot sign(\boldsymbol{w})$, then we have,

$$||\boldsymbol{r}||_\infty = |\eta| = \frac{b_0}{||\boldsymbol{w}||_1} \tag{13}$$

Lower bound is achieved, thus the solution is,

$$\boldsymbol{r} = \frac{|b_0|}{||\boldsymbol{w}_1||} \cdot sign(\boldsymbol{w}) \tag{14}$$

(ii) DeepFool attack

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmin}\ ||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 \\ &subject\ to\ g(\boldsymbol{x}) = 0 \end{aligned} \tag{15}$$

First order approximation

$$g(\boldsymbol{x}) \approx g(\boldsymbol{x}^{(k)}) + \nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)})^T(\boldsymbol{x} - \boldsymbol{x}^{(k)}) \tag{16}$$

Then the problem can be approximate by

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmin}\ ||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 \\ &subject\ to\ g(\boldsymbol{x}^{(k)}) + \nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)})^T(\boldsymbol{x} - \boldsymbol{x}^{(k)}) = 0 \end{aligned} \tag{17}$$

Let $\boldsymbol{w}^{(k)} = \nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)})$ and $w_0^{(k)} = g(\boldsymbol{x}^{(k)}) - \nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)})^T\boldsymbol{x}^{(k)}$
Then the problem is equivalent to

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmin}\ ||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 \\ &subject\ to\ (\boldsymbol{w}^{(k)})^T\boldsymbol{x} + w_0^{(k)} = 0 \end{aligned} \tag{18}$$

This is the same problem as minimum l_2 norm attack, Thus the solution will be,

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \frac{((\boldsymbol{w}^{(k)})^T x^{(k)} + w_0^{(k)})\boldsymbol{w}^{(k)}}{||\boldsymbol{w}^{(k)}||_2^2} \tag{19}$$

substitute $\boldsymbol{w}$ and w_0 back, we get

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \left(\frac{g(\boldsymbol{x}^{(k)})}{||\nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)})||^2}\right)\nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)}) \tag{20}$$

(iii) An example DeepFool never converge

something

(b) maximum-allowable attack

(i) l_∞ attack in the linear case

The problem is,

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmin}\ \boldsymbol{w}^T\boldsymbol{x} + w_0 \\ &subject\ to\ ||\boldsymbol{x} - \boldsymbol{x}_0||_\infty < \eta \end{aligned} \tag{21}$$

let $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{r}$, $b_0 = (\boldsymbol{w}^T\boldsymbol{x}_0 + w_0)$, the problem becomes,

$$\begin{aligned} &\underset{\boldsymbol{r}}{argmin}\ \boldsymbol{w}^T\boldsymbol{r} + b_0 \\ &subject\ to\ ||\boldsymbol{r}||_\infty < \eta \end{aligned} \tag{22}$$

by Holder's inequality,

$$\boldsymbol{w}^T\boldsymbol{r} \geq -||\boldsymbol{r}||_\infty||\boldsymbol{w}||_1 \geq -\eta||\boldsymbol{w}||_1 \tag{23}$$

as shown in the lecture note, the solution

$$\boldsymbol{r} = -\eta \cdot sign(\boldsymbol{w}) \tag{24}$$

(ii) FGSM attack

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmax}\ J(\boldsymbol{x}, \boldsymbol{w}) \\ &subject\ to\ ||\boldsymbol{x} - \boldsymbol{x}_0||_\infty \leq \eta \end{aligned} \tag{25}$$

Approximately, $J(\boldsymbol{x}, \boldsymbol{w}) = J(\boldsymbol{x}_0 + \boldsymbol{r}, \boldsymbol{w}) \approx J(\boldsymbol{x}_0, \boldsymbol{w}) + \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0, \boldsymbol{w})^T\boldsymbol{r}$
Then, the problem becomes

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmin}\ \ -J(\boldsymbol{x}_0, \boldsymbol{w}) - \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0, \boldsymbol{w})^T\boldsymbol{r}_{\qquad in} \\ &subject\ to\ ||\boldsymbol{r}||_\infty \leq \eta \end{aligned} \tag{26}$$

of which the solution is given by

$$\boldsymbol{x} = \boldsymbol{x}_0 + \eta \cdot (\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0, \boldsymbol{w})) \tag{27}$$

In the problem setup, we get $J(\boldsymbol{x}) = -g(\boldsymbol{x})$, thus the solution is

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{x}_0 - \eta \cdot (\nabla_{\boldsymbol{x}} g(\boldsymbol{x}_0)) \\ &= \boldsymbol{x}_0 - \eta \cdot ((\boldsymbol{W}_j - \boldsymbol{W}_t)\boldsymbol{x}_0 + (\boldsymbol{w}_j - \boldsymbol{w}_t)) \\ &= (\eta(\boldsymbol{W}_t - \boldsymbol{W}_j) + 1)\boldsymbol{x}_0 + \eta(\boldsymbol{w}_t - \boldsymbol{w}_j) \end{aligned} \tag{28}$$

(iii) I-FGSM attack

$$\begin{aligned} J(\boldsymbol{x}, \boldsymbol{w}) &= J(\boldsymbol{x}_0 + \boldsymbol{r}, \boldsymbol{w}) \\ &\approx J(\boldsymbol{x}_0, \boldsymbol{w}) + \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0, \boldsymbol{w})^T \boldsymbol{r} \\ &= J(\boldsymbol{x}_0, \boldsymbol{w}) + \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0, \boldsymbol{w})^T (\boldsymbol{x} - \boldsymbol{x}_0) \\ &= J(\boldsymbol{x}_0, \boldsymbol{w}) + \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0, \boldsymbol{w})^T \boldsymbol{x} - \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0, \boldsymbol{w})^T \boldsymbol{x}_0 \end{aligned} \tag{29}$$

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= \underset{0 \le \boldsymbol{x} \le 1}{argmax}\ J(\boldsymbol{x}^{(k)}, \boldsymbol{w})\ subject\ to\ ||\boldsymbol{x} - \boldsymbol{x}_0|| \le \eta \\ &= \underset{0 \le \boldsymbol{x} \le 1}{argmax}\ \nabla_{\boldsymbol{x}} J(\boldsymbol{x}^{(k)}, \boldsymbol{w})^T \boldsymbol{x}\ subject\ to\ ||\boldsymbol{x} - \boldsymbol{x}_0|| \le \eta \\ &= \mathcal{P}\left\{\boldsymbol{x}^{(k)} + \eta \cdot sign\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}^{(k)}, \boldsymbol{w})\right)\right\} \end{aligned} \tag{30}$$

(c) Regularization based attack

(i) linear case

$$\underset{\boldsymbol{x}}{argmin}\ \frac{1}{2}||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 + \lambda(\boldsymbol{w}^T \boldsymbol{x} + w_0) \tag{31}$$

Taking derivative

$$\begin{aligned} &\nabla_{\boldsymbol{x}}\ \frac{1}{2}||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 + \lambda(\boldsymbol{w}^T \boldsymbol{x} + w_0) \\ &= \boldsymbol{x} - \boldsymbol{x}_0 + \lambda \boldsymbol{w} = 0 \end{aligned} \tag{32}$$

Solve for $\boldsymbol{x}$,

$$\boldsymbol{x} = \boldsymbol{x}_0 - \lambda \boldsymbol{w} \tag{33}$$

(ii)

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmin}\ \varphi(\boldsymbol{x}), where \\ &\varphi(\boldsymbol{x}) = ||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 + \lambda\zeta(g_j(\boldsymbol{x}) - g_t(\boldsymbol{x})), \\ &\zeta(y) = max(y, 0),\ and\ j \neq t \end{aligned} \tag{34}$$

Taking the derivative,

$$\nabla\varphi(\boldsymbol{x}) = 2(\boldsymbol{x} - \boldsymbol{x}_0) + \lambda\mathbb{I}\{g_j(\boldsymbol{x}) - g_t(\boldsymbol{x}) > 0\} \cdot (\nabla_{\boldsymbol{x}} g_j(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} g_t(\boldsymbol{x})) \tag{35}$$

Using my favorite gradient descent, we can tell,

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha\nabla_{\boldsymbol{x}}\varphi(\boldsymbol{x}^{(k)}) \tag{36}$$

substituting in $g(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T(\boldsymbol{W}_j - \boldsymbol{W}_t)\boldsymbol{x} + (\boldsymbol{w}_j - \boldsymbol{w}_t)^T\boldsymbol{x} + (w_{j,0} - w_{t,0})$, we get

$$\begin{aligned} \nabla_{\boldsymbol{x}} g(\boldsymbol{x}) &= (\boldsymbol{W}_j - \boldsymbol{W}_t)\boldsymbol{x} + (\boldsymbol{w}_j - \boldsymbol{w}_t) \\ \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} - 2\alpha(\boldsymbol{x}^{(k)} - \boldsymbol{x}_0) - \alpha\lambda\mathbb{I}\{g(\boldsymbol{x}) > 0\} \cdot (\nabla_{\boldsymbol{x}} g(\boldsymbol{x})) \\ &= \boldsymbol{x}^{(k)} - 2\alpha(\boldsymbol{x}^{(k)} - \boldsymbol{x}_0) - \alpha\lambda\mathbb{I}\{g(\boldsymbol{x}) > 0\} \cdot ((\boldsymbol{W}_j - \boldsymbol{W}_t)\boldsymbol{x} + (\boldsymbol{w}_j - \boldsymbol{w}_t)) \end{aligned} \tag{37}$$

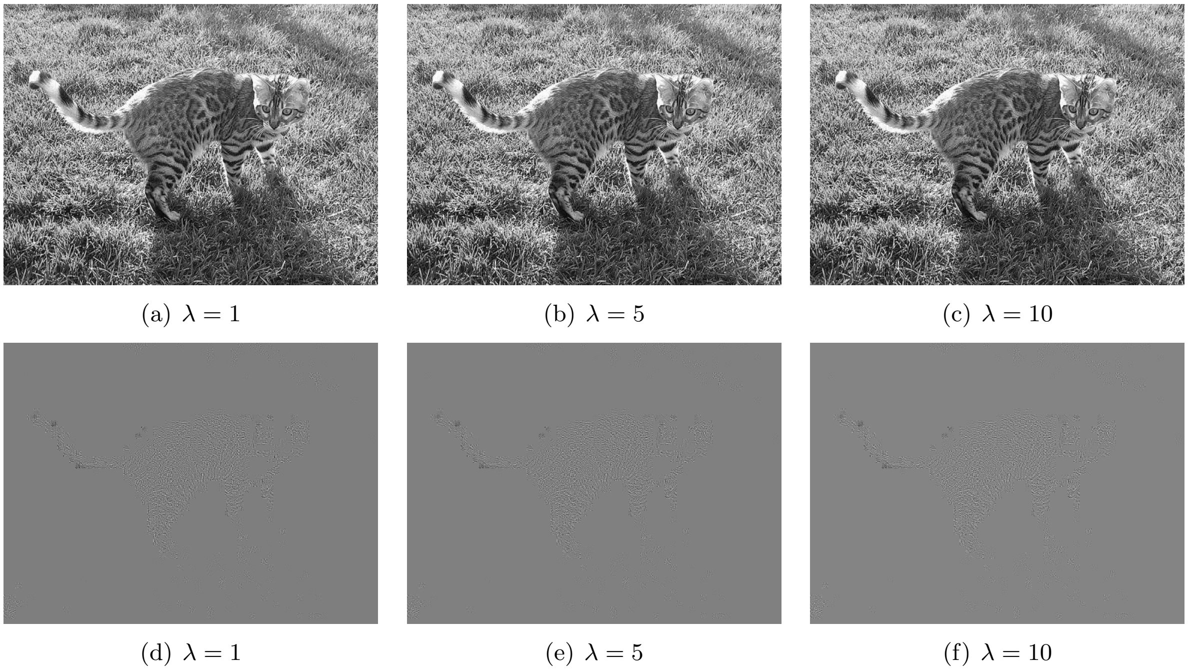
Exercise 2: CW-Attack on Gaussian Classifier - Non-overlapping Patches

(a)&(b) Implementation

refer to code in the back

(c) Results

(i) & (ii) The final perturbed images & their perturbation respectively



(a) $\lambda = 1$ (b) $\lambda = 5$ (c) $\lambda = 10$

(d) $\lambda = 1$ (e) $\lambda = 5$ (f) $\lambda = 10$

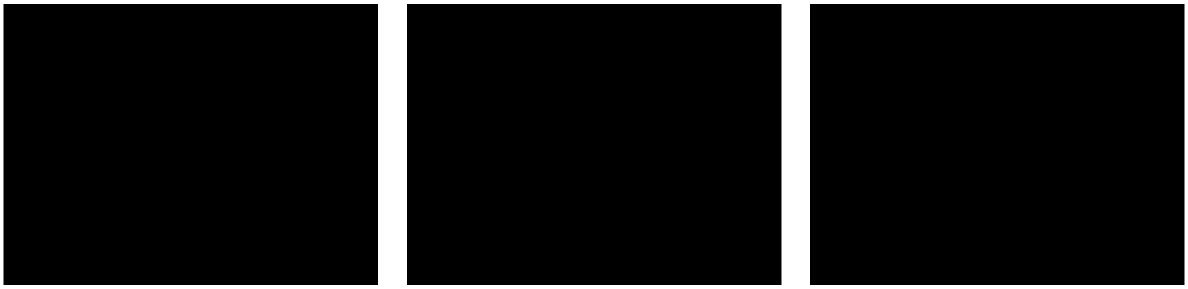
Figure 1: perturbed images and their perturbations with $\lambda = 1, 5, 10$

(iii) Frobenius norm of the perturbation

λ	1	5	10
Frobenius norm	8.253352243144604	8.928742803780679	9.697938770929634

Table 1: Frobenius norm vs. λ for CW-Attack on Gaussian classifier with non-overlapping patches

(iv) The classifiers output

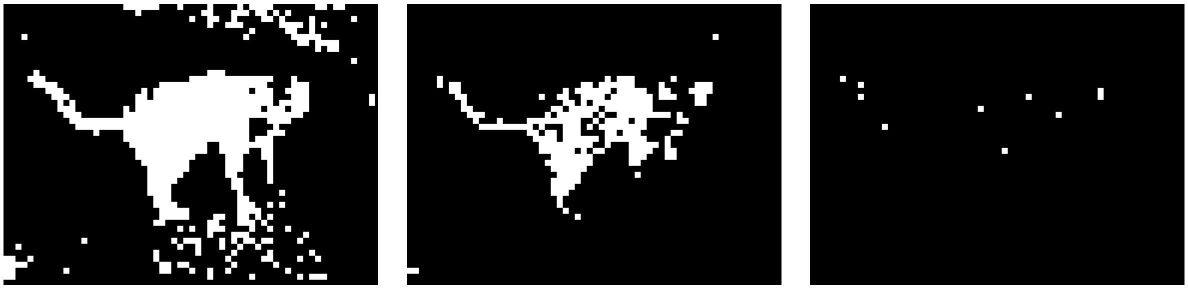


(a) $\lambda = 1$ (b) $\lambda = 5$ (c) $\lambda = 10$

Figure 2: classified perturbed image with $\lambda = 1, 5, 10$

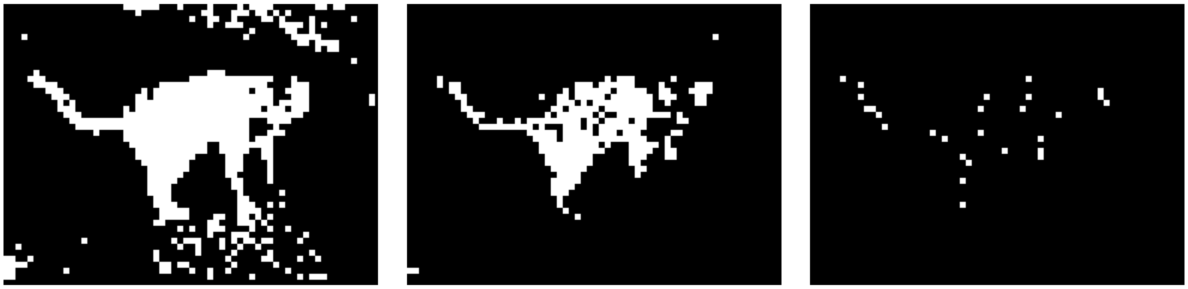
For all three cases, all patches are classified as grass after applying perturbation.

(v) Plots during gradient descent



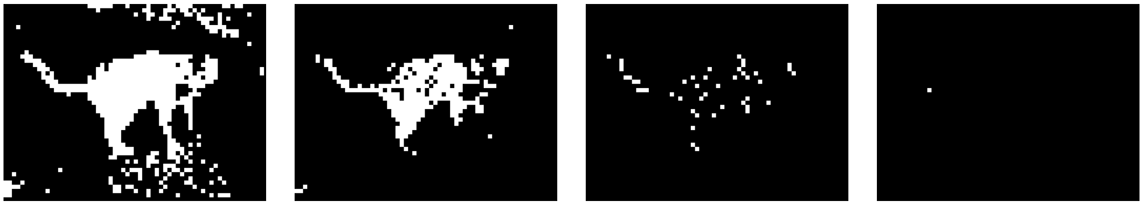
(a) after 0 rounds (b) after 50 rounds (c) after 100 rounds

Figure 3: classified perturbed image during gradient descent with $\lambda = 1$



(a) after 0 rounds (b) after 10 rounds (c) after 20 rounds

Figure 4: classified perturbed image during gradient descent with $\lambda = 5$



(a) after 0 rounds (b) after 5 rounds (c) after 10 rounds (d) after 15 rounds

Figure 5: classified perturbed image during gradient descent with $\lambda = 10$

Comments

From the above plots, we can obviously see that increase λ will speed up the attack. However, as λ goes up, the quality of the attack goes down since the Frobenius norm also goes up while achieving the same result.

Exercise 3: CW Attack on Gaussian Classifier - Overlapping Patches

$$\begin{aligned}\boldsymbol{X}^* &= \underset{\boldsymbol{X}}{argmin}\sum_{i=1}^{L}\{||\boldsymbol{P}_i(\boldsymbol{X}-\boldsymbol{X}_0)||_2^2 + \lambda max(g_j(\boldsymbol{P}_i\boldsymbol{X}) - g_t(\boldsymbol{P}_i\boldsymbol{X}), 0)\} \\ &= \underset{\boldsymbol{X}}{argmin}||\boldsymbol{X}-\boldsymbol{X}_0||_2^2 + \lambda\sum_{i=1}^{L} max(g_j(\boldsymbol{P}_i\boldsymbol{X}) - g_t(\boldsymbol{P}_i\boldsymbol{X}), 0)\end{aligned} \tag{38}$$

(a) Theory

the gradient of the above function is,

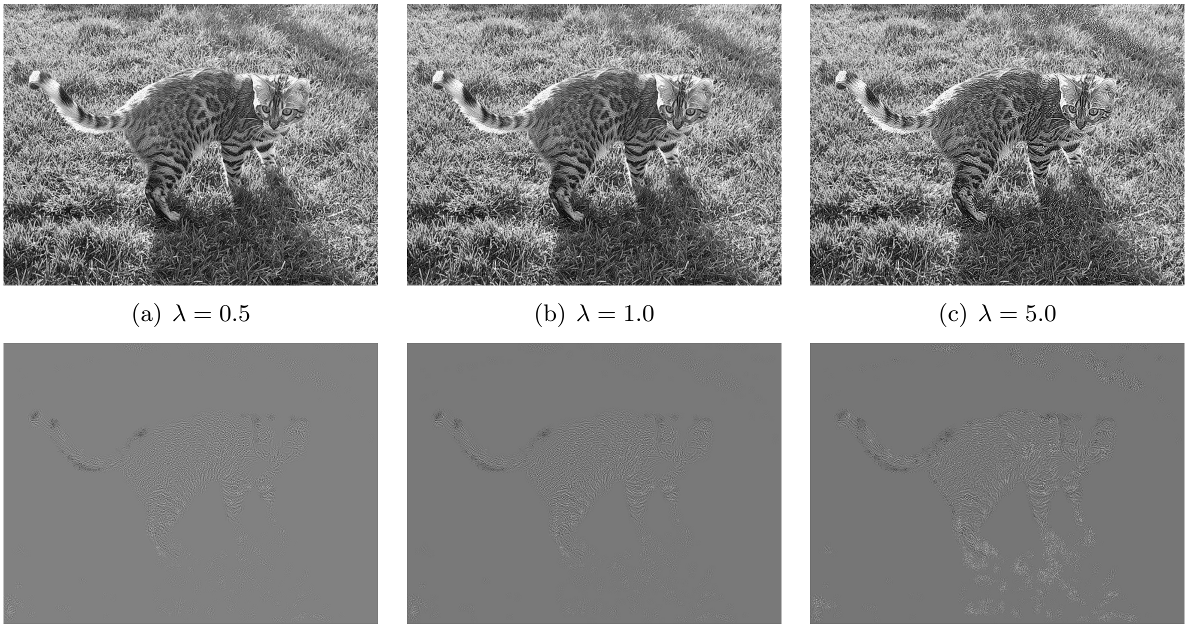
$$\nabla_{\boldsymbol{X}} = 2(\boldsymbol{X}-\boldsymbol{X}_0) + \lambda\mathbb{I}\{g_j(\boldsymbol{P}_i\boldsymbol{X}) - g_t(\boldsymbol{P}_i\boldsymbol{X}) > 0\} \times (\nabla_{\boldsymbol{X}}(g_j(\boldsymbol{P}_i\boldsymbol{X}) - g_t(\boldsymbol{P}_i\boldsymbol{X}))) \tag{39}$$

(b) Implementation

Refer to code in the back.

(c) Results

(i) & (ii) The final perturbed images & their perturbation respectively



(a) $\lambda = 0.5$

(b) $\lambda = 1.0$

(c) $\lambda = 5.0$

(d) $\lambda = 0.5$

(e) $\lambda = 1.0$

(f) $\lambda = 5.0$

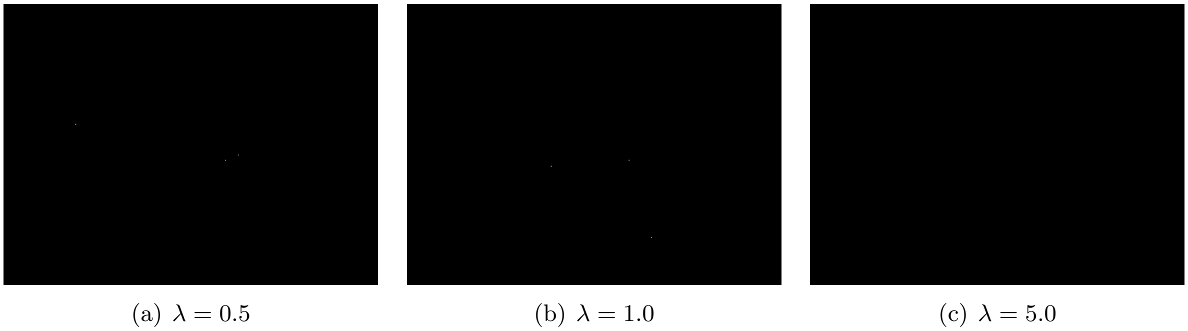
Figure 6: perturbed images and their perturbations with $\lambda = 0.5, 1.0, 5.0$

(iii) Frobenius norm of the perturbation

λ	#grass patches	#cat patches	Frobenius norm
0.5	180561	3	14.059260687697552
1.0	180561	3	16.158563208747122
5.0	180564	0	33.617113348842395

Table 2: testing results for CW-Attack on Gaussian classifier with overlapping patches

(iv) The classifiers output



(a) $\lambda = 0.5$

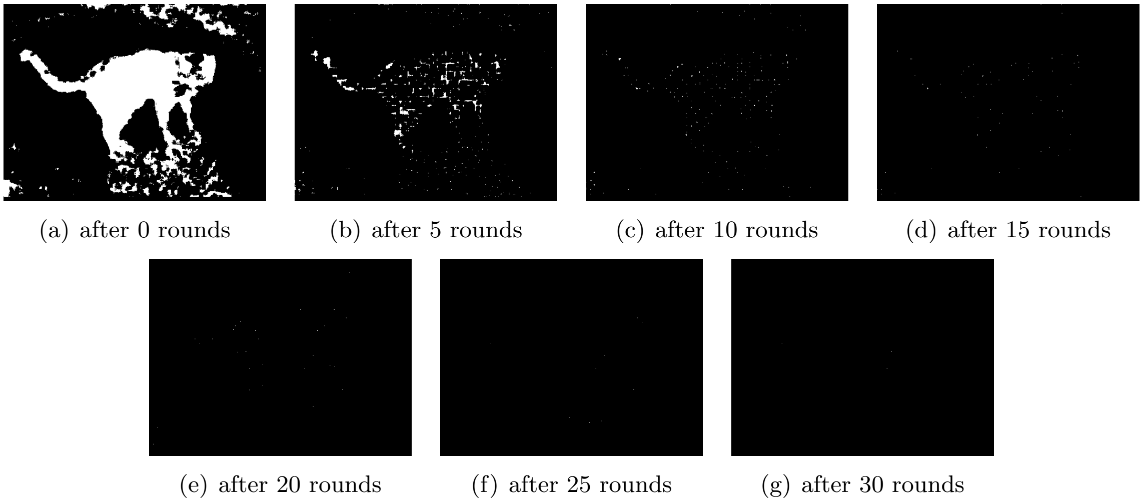
(b) $\lambda = 1.0$

(c) $\lambda = 5.0$

Figure 7: classified perturbed image with $\lambda = 0.5, 1.0, 5.0$

The precise result is in the table above, however, from observation, almost all pixels are classified as grass.

(v) Plots during gradient descent



(a) after 0 rounds

(b) after 5 rounds

(c) after 10 rounds

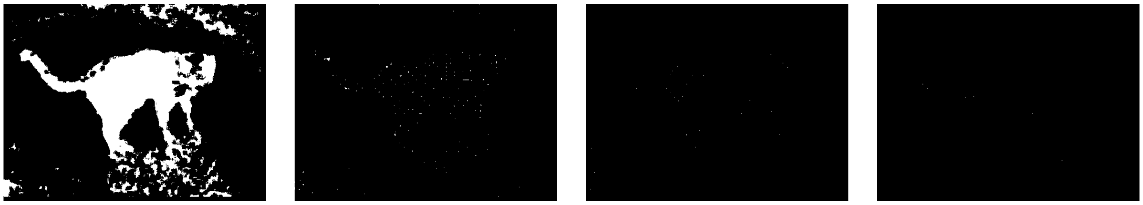
(d) after 15 rounds

(e) after 20 rounds

(f) after 25 rounds

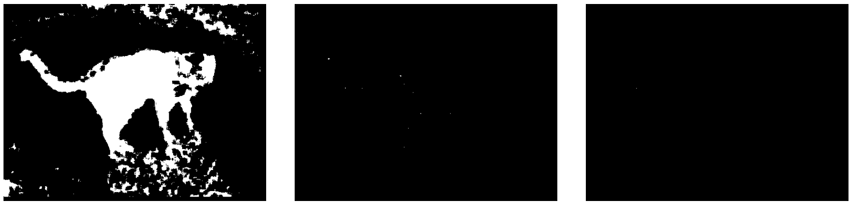
(g) after 30 rounds

Figure 8: classified perturbed image during gradient descent with $\lambda = 0.5$



(a) after 0 rounds (b) after 5 rounds (c) after 10 rounds (d) after 15 rounds

Figure 9: classified perturbed image during gradient descent with $\lambda = 1.0$



(a) after 0 rounds (b) after 2 rounds (c) after 4 rounds

Figure 10: classified perturbed image during gradient descent with $\lambda = 5$

Comments

In general, comparing to the non-overlapping case, the Frobenius norm is significantly larger. However, since the overlapping is a much harder problem, the result is acceptable. As *lambda* goes up, the strength of the attack is approximately the same, less rounds are needed to attack the classifier while sacrificing some quality of the attack.