Summary and Analysis of: 'Fighting Gradients with Gradients: Dynamic Defenses against Adversarial Attacks'

As we have learned throughout this project, deep networks can be extremely vulnerable to what are called adversarial attacks. This is where an adversary utilizes an imperceptible amount of noise. More technically a perturbation constrained by some norm threshold. This noise fools the classifier into incorrectly classify the given input. However, many of our current defenses are what is called "static", meaning they do not change after training time. This leads to a lack of adaptivity for our deep network leads to trouble if the adversarial attacks evolve. This paper proposes a solution to this problem in a "dynamic" defense, meaning it updates during testing, called *Defensive entropy minimization*, or dent. What this means, is that any adversary never faces the same defense, it is constantly changing, making it more effective. The reason being, an adversary can find a hole in the static defense and continually exploit it and our defense can do nothing to prevent this. The paper discusses three main contributions, the weakness of static vs strengths of dynamic approaches, the specifics of dent – the dynamic model proposed, and the effectiveness of dent. The basic idea of the paper is most effectively put on page 3 at the start of section 3 "Adversarial attacks optimize against defenses at test time, so defenses should fight back, and counter-optimize against attacks." Dent not only evolves during test-time, it only evolves during test-time, and does not change the training in any way. So while the adversary is optimizing the perturbation δ, dent is optimizing both its model (Δ) and the input (Σ) , which both depend on the testing data. It updates on batches based on:

$$\text{argmin}_{\Sigma,\Delta}\, H(f(g(x + \delta; \Sigma); \theta + \Delta))$$

Where H is the Shannon Entropy:

$$H(y_i) = \sum_i - p(y_i) \log p(y_i)$$

It allows its defense parameters to reset between the batches. This allows the defense to move with the adversary as an attack of $x + \delta^t$ can be met with Δ^t which means the defense will always get the last move, which is an advantage. The model and input parameters are differentiable which is how dent 'fights gradients with gradients.' They tested a static method versus dent versus dent + (which is dent but adapting

sample-wise rather than batch-wise). We find that dent and dent + are way more effective in a basic example than the static method at improving even state-of-the-art defenses against adversaries or models with no defense at all, while still maintaining natural (no adversary) accuracy. Maintaining natural accuracy is extremely important, as often as a side effect, adversarial defenses can negatively impact natural accuracy. Dent is also robust to many different attack types, improving the best defenses by up to 20 points against certain attacks. Additionally, it is effective within many different architectures and different types of datasets. Ultimately although Dent has some limitations, such as depending on batches to evolve since it utilizes the entropy function as its update function, it seems to be way more effective and robust than even the best static defense in a big variety of situations.

EECS 127 Project

4b. $g(\vec{v}_2, \vec{v}_3) = \min_{\vec{z}} L(\vec{z}, \vec{v})$

$$= \min_{\vec{z}} \vec{c}^T \vec{z}_3 + \vec{v}_3^T \vec{z}_3 + \mathbf{1}_{Bc(x)}(\vec{z}_1) - \vec{v}_2^T W_1 \vec{z}_1 + \left(\sum_{j=1}^{\hat{n}_2} \mathbf{1}_{z_j}(z_{2j}, \hat{z}_{2j}) - \vec{v}_3^T W_2 \vec{z}_2 + \vec{v}_2^T \vec{z}_2\right) - \sum_{i=1}^{k} \vec{v}_{i+1}^T \vec{b}_i$$

since these terms are all only in terms of at most one z_i we can split our minimization

$$= \min_{\vec{z}_3} \left((\vec{c}^T + v_3)^T z_3\right) + \min_{\vec{z}_1} \left(\mathbf{1}_{Bc(x)}(z_1) - \vec{v}_2^T W_1 z_1\right) + \left(\sum_{j=1}^{\hat{n}_2} \left(\min_{z_{2j}, \hat{z}_{2j}} \left(\mathbf{1}_{z_j}(z_{2j}, \hat{z}_{2j}) - \vec{v}_3^T (W_2)_j z_{2j} + v_{2j} \hat{z}_{2j}\right)\right)\right) - \underbrace{\sum_{i=1}^{k} v_{in}^T b_i}_{\text{not dependent on } z}$$

As we wanted to show

4c. $d^*(\vec{x}, \vec{c}) = \max_{\vec{v}} g(\vec{v}_2, \vec{v}_3)$

So we can first find the minima of each

$\frac{\partial}{\partial z_3}\left((c + v_3)^T z_3\right) = 0$

$v_3 = -c$, since this is not in terms of z, we can make this a constraint

$\frac{\partial}{z_1}\left(\mathbf{1}_{x \in C(x)}(\vec{z}_1) - \vec{v}_2^T W_1 \vec{z}_1\right) = 0$

$\mathbf{1}_{x \in C(x)}(-v_1^T \vec{v}_2) = 0$

which will happen at the optimum of our $\mathbf{1}_{x \in C(x)}$ function

$-\mathbf{1}^*_{x \in C(x)}(W_1^T \vec{v}_2)$

$\frac{\partial}{z_2}\sum_{j=1}^{\hat{n}_2}\left(\mathbf{1}_{z_j}(z_{2j}, \hat{z}_{2j}) - v_3^T (W_2)_j z_{2j} + v_{2j} \hat{z}_{2j}\right) = 0$

$\sum_{j=1}^{\hat{n}_2} \mathbf{1}_{z_j}(-v_3^T (W_2)_j + v_{2j})) = 0$

similarly at optimum at characteristic function this is 0

$\sum_{j=1}^{\hat{n}_2} -\mathbf{1}^*_{z_j}(\vec{v}_3^T (W_2)_j - v_{2j})$

And everything else is fixed so we have

$d^*(\vec{x}, \vec{c}) = \max_{\vec{v}} -\mathbf{1}^*_{x \in C(x)}(W_1^T \vec{v}_2) + \sum_{j=1}^{\hat{n}_2} -\mathbf{1}^*_{z_j}(\hat{v}_3^T (W_2)_j - v_{2j}) - \sum_{i=1}^{k} \vec{v}_{i+1}^T \vec{b}_i$

s.t $\vec{v}_3 = -\vec{c}$ which is what we wanted to show

And using the definition as given we have:

$d^*(\vec{x}, \vec{c}) = \max_{\vec{v}} -\mathbf{1}^*_{x \in C(x)}(\vec{v}_1) + \sum_{j=1}^{\hat{n}_2} -\mathbf{1}^*_{z_j}(\hat{v}_{2j} - v_{2j}) - \sum_{i=1}^{k} \vec{v}_{i+1}^T \vec{b}_i$

s.t $\vec{v}_3 = -\vec{c}$

$\vec{\hat{v}}_2 = W_2^T v_3$

$\vec{\hat{v}}_1 = W_1^T v_2$

9.

$$\max_{\|x-x'\|_\infty \le \epsilon} L(F_\theta(x'), y_{true}) = \max_{z_K \in Z_\epsilon(x)} L(\hat{z}_K, y)$$

since $L(x,y) \le L(x - a\mathbf{1}, y) \quad \forall a$

$$\max_{\hat{z}_K \in Z_\epsilon(x)} L(\hat{z}_K, y) \le \max_{\hat{z}_K \in Z_\epsilon(x)} L(\hat{z}_K - (\hat{z}_K)_y \mathbf{1}, y)$$

$$= \max_{\hat{z}_K \in Z_\epsilon(x)} L((I - e_y \mathbf{1}^T)\hat{z}_K, y)$$

$$= \max_{\hat{z}_K \in Z_\epsilon(x)} L(C\hat{z}_K, y)$$

Letting C be the matrix $(I - e_y \mathbf{1}^T)$

by bounding the elementwise maximum over $C\hat{z}_K$ and the elementwise minimum for $c = y$

$$\max_{\hat{z}_K \in Z_\epsilon(x)} L(C\hat{z}_K, y) \le L(h(\hat{z}_K))$$

where i being representative of the rows of C we define it as

$$h(z_K)_i = \max_{\hat{z}_K \in Z_\epsilon(x)} C_i \hat{z}_K$$

which gives us that

$$J_\epsilon(x, y_\theta(-C_i)) \le \min_{z_K \in Z_\epsilon(x)} -C_i^T \hat{z}_K$$

$$-J_\epsilon(x, y_\theta(-C_i)) \ge \max_{z_K \in Z_\epsilon(x)} C_i^T \hat{z}_K$$

$$h(z_K)_i \le -J_\epsilon(x, y_\theta(-C_i))$$

This all gives us that indeed

$$\max_{\|x-x'\|_\epsilon \le \epsilon} L(F_\theta(x'), y) \le L(-J_\epsilon(x, y_\theta(e_y \mathbf{1}^T - I)), y)$$

1. $\underset{\vec{x}'}{\text{argmax}} \; L(f_\theta(\vec{x}), \vec{y}_{true}) + \nabla_{\vec{x}} L(f_\theta(\vec{x}), y_{true})^T \vec{x}'$

s.t. $\|\vec{x} - \vec{x}'\|_\infty \leq \varepsilon$

Let $\vec{x}' = \vec{x} + \varepsilon \vec{u}$ $\rightarrow$

$$\underset{\vec{u}}{\text{argmax}} \; \nabla_{\vec{x}} L(f_\theta(\vec{x}), \vec{y}_{true})^T (x + \varepsilon u)$$

s.t. $\|-\varepsilon u\| \leq \varepsilon \rightarrow \|u\| \leq 1$

$$= \underset{\vec{u}}{\text{argmax}} \; \varepsilon \nabla_{\vec{x}} L(f_\theta(\vec{x}), \vec{y}_{true})^T u$$

s.t. $\|u\|_\infty \leq 1$

The max is $\varepsilon \| \nabla_{\vec{x}} L(f_\theta(\vec{x}), \vec{y}_{true} \|_1$,

thus u must be

$\text{sgn}(\nabla_x L(f_\theta(\vec{x}), y_{true}))$, thus

$$\underset{x'}{\text{argmax}} = \vec{x} + \varepsilon \, \text{sgn}(\nabla_x L(f_\theta(\vec{x}), y_{true}))$$

2a) $f^*(y) = \sup_s \{ys - |s|\}$

We can partition into casework.

Case 1: $y < -1$. Then $g(s) = ys - |s|$,

$$\begin{cases} g(s) \to -\infty & s > 0 \\ g(s) \to \infty & s < 0 \end{cases}$$

Thus $f^*(y)$ s.t. $y < -1 = \infty$

Case 2: $-1 < y < 1$. Then $g(s)$ follows the graph of $-|s|$ closely, thus

$$\begin{cases} g(s) \to -\infty & s > 0 \\ g(s) \to -\infty & s < 0 \end{cases}$$

with an apex at $(0, 0)$. thus $f^*(y) = 0$.

Case 3: $y > 1$. Then the negative of $y < -1$ is true. Thus

$$f^*(y) = \begin{cases} \infty & y > 1 \text{ or } y < -1 \\ 0 & -1 \le y \le 1 \end{cases}$$

b) $f^*(y) = \sup_s (y^T s - \|s\|_1)$

$= \sup_s \sum y_i s_i - \sum |s_i|$.

$= \sup_s \sum y_i s_i - |s_i|$

$= \sum \sup_{s_i} y_i s_i - |s_i|$

$= \sum f^*(y_i)$. Thus

$$f^*(y) = \begin{cases} \infty & \text{iff } \exists y_i < -1 \vee y_i > 1 \\ 0 & \text{iff } \forall y_i \ -1 \le y_i \le 1 \end{cases}$$

3. $\rho^*(\vec{x}, \vec{c}) = c^T \hat{z}_3$

s.t. $\|z_1 - x\|_\infty \le \epsilon$

$(z_{2j}, \hat{z}_{2j}) \in Z_j$

$\vec{z}_2 = W_1 \vec{z}_1 + \vec{b}_1$

$\hat{z}_3 = W_2 \vec{z}_2 + \vec{b}_2$

$$\mathbb{I}_{B_\epsilon(x)}(z) := \begin{cases} 0 \text{ iff } z \in B_\epsilon(x) \\ \infty \text{ otherwise} \end{cases}$$

We can reinforce the constraints by adding $\mathbb{I}_{B_\epsilon(x)}(\vec{z}_1)$ and $\mathbb{I}_{Z_j}(z_{2j}, \hat{z}_{2j})$ since if the constraints are violated then the problem becomes infeasible.

Thus

$$= \min_z \; c^T \hat{z}_3 + \mathbb{I}_{B_\epsilon(x)}(\vec{z}_1) + \sum \mathbb{I}_{Z_j}(z_{2j}, \hat{z}_{2j})$$

s.t. $\vec{z}_2 = W_1 \vec{z}_1 + \vec{b}_1$

$\hat{z}_3 = W_2 \vec{z}_2 + \vec{b}_2$

4a) $\int \Big\{ \quad c^T \hat{z}_3 + I_{B_\epsilon(\vec{x})}(\vec{z}_1) + \sum I_{Z_j}(z_{2j}, \hat{z}_{2j})$

$\text{s.t. } \vec{z}_2 = W_1 \vec{z}_1 + \vec{b}_1 \quad \Big\}$

$\hat{z}_3 = W_2 \vec{z}_2 + \vec{b}_2 \quad \Big\}$

$= c^T \hat{z}_3 + I_{B_\epsilon(\vec{x})} z_1 + \sum I_{Z_j}(z_{2j}, \hat{z}_{2j}) + v_3^T \hat{z}_3 - v_3^T W_2 z_2 -$

$v_3^T b_2 + v_2^T \hat{z}_2 - v_2^T W_1 z_1 + v_2^T b_1$

$= c^T \hat{z}_3 + v_3^T \hat{z}_3 + I_{B_\epsilon(x)}(z_1) - v_2^T W_1 z_1$

$+ \left(\sum_{j=1}^{n_2} \min_{z_{2j}, \hat{z}_{2j}} \left(I_Z(z_{2j}, z_{2j}) - v_3^T (W_2)_j z_{2j} + v_{2j} \hat{z}_{2j} \right) \right)$

$- \sum_{i=1}^{2} v_{i+1}^T b_i$

4b) Even part of the dual is nicely subdivided according to it's z or $\hat{z}$. Thus

$g(v_1, v_2) = \min_{\hat{z}_3} (c + v_3)^T \hat{z}_3 + \min_{z_1} I_{B_\epsilon(x)}(z_1) - v_2^T W_1 z_1$

$+ \left(\sum_{j=1}^{n_2} \min_{z_{2j}, \hat{z}_{2j}} \left(I_Z(z_{2j}, z_{2j}) - v_3^T (W_2)_j z_{2j} + v_{2j} \hat{z}_{2j} \right) \right)$

$- \sum_{i=1}^{2} v_{i+1}^T b_i$

4c) • $\min_{\tilde{z}_3} (c + v_3)^T \tilde{z}_3 = 0$ since $\nabla = 0$

$\rightarrow v_3 = -c$

• $\min_{z_1} \left(\mathcal{I}_{B_\epsilon(x)}(z_1) - \sigma_2^T W_1 z_1 \right) = - \max_{z_1} \sigma_2^T W_1 z_1 - \mathcal{I}_{B_\epsilon(x)}(z_1)$

$= - \mathcal{I}^*_{B_\epsilon(x)}(\sigma_2)$

• Similarly: $\min_{z_{2j}, \hat{z}_{2j}} \left(\mathcal{I}_{\mathcal{Z}}(z_{2j}, \hat{z}_{2j}) - v_3^T (W_2)_j z_{2j} + v_{2j} \hat{z}_{2j} \right)$

$= - \mathcal{I}^*_{\mathcal{Z}_j}(\hat{\sigma}_{2j,1} - v_{2j})$ if we define

$v_3^T (W_2)_j = (W_2^T)_j v_3 = \hat{v}_j$

Thus $\mathcal{L}$ dual is

$$\max_{v} \; - \mathcal{I}^*_{B_\epsilon(x)}(v_1) + \sum_{j=1}^{n_2} - \mathcal{I}_{\mathcal{Z}_j}(\hat{v}_{2j}, -v_{2j}) - \sum_{i=1}^{2} v_{i+1}^T b_i$$

$$\text{s.t.} \quad v_3 = -c$$

$$\hat{v}_2 = W_2^T v_3$$

$$\hat{v}_1 = W_1^T v_2$$

5. $I^*_{B_\epsilon(\vec{x})}(\vec{r}) = \sup_{s \text{ s.t. } \|s-x\|_\infty \le \epsilon} \{r^T s\}$

$s = x + \epsilon u$

$= \sup_{u \text{ s.t. } \|u\|_\infty \le 1} r^T x + \epsilon r^T u$

$= r^T x + \epsilon \|v\|_1$

$$\max_{V} \; -I^*_{B_\epsilon(x)}(V_1) + \sum_{j=1}^{n_2} -I_{Z_j}(\hat{v}_{2j}, -v_{2j}) - \sum_{i=1}^{2} v_{i+1}^T b_i$$

$$\text{s.t.} \quad v_3 = -c$$
$$\hat{v}_2 = W_2^T v_3$$
$$\hat{v}_1 = W_1^T v_2$$

$$\max_{V} \; -x^T v_1 - \epsilon \|v_1\|_1 - \sum_{j=1}^{n_2} v_{i+1}^T b_i + \sum_{j \in S} l_j \mathrm{ReLU}(v_{2j})$$

$$\text{s.t.} \quad v_3 = -c$$
$$\hat{v}_2 = W_2^T v_3$$
$$\hat{v}_1 = W_1^T v_2$$
$$v_{2j} = 0 \quad \forall j \in S^-$$
$$v_{2j} = \hat{v}_{2j} \quad \forall j \in S^+$$
$$v_{2j} = \frac{u_j}{u_j - l_j} \quad \forall j \in S$$

Since we have upper bounded each Fenchel conjugate I^*_Z and the bounds are determined by the variables, this is a valid substitution.

7. $\hat{z}_2 = W_1 z_1 + b_1$

We know that $z_1 = x + \varepsilon u$. Thus

$\hat{z}_2 = W_1 x + \varepsilon W_1 u + b_1$. Thus

$\max W_{ij}^T u = \|W_{ij}\|_1$

s.t. $\|u\|_\infty \leq 1$

To maximize/minimize $\hat{z}_2$, you can add/subtract respectably (since the quantity is always positive).

$\max \hat{z}_2 = W_1 x + b_1 + \varepsilon \|W_1\|_1$

$\min \hat{z}_2 = W_1 x + b_1 - \varepsilon \|W_1\|_1$

8. Monotonic: We assume $y_i \leq y_i'$ for i pertaining to the incorrect classes and $y_{i_{true}} \geq y'_{i_{true}}$. Then

$$L(y, y_{true}) = -\log\left(\frac{e^{y_{i_{true}}}}{\sum_{j=1}^{m} e^{y_j}}\right)$$

$$= -\left[\log\left(e^{y_{i_{true}}}\right) - \log\left(\sum_{j=1}^{m} e^{y_j}\right)\right]$$

$$= \log\left(\sum_{j=1}^{m} e^{y_j}\right) - \log\left(e^{y_{i_{true}}}\right)$$

$$\leq \log\left(\sum_{j=1}^{m} e^{y_j}\right) - \log\left(e^{y'_{i_{true}}}\right)$$

$$= -\sum_{i=1}^{m} (\vec{y}_{true})_i \log\left(\frac{e^{y_i'}}{\sum_{j=1}^{m} e^{y_j}}\right)$$

Translation-Invariant: We have that

$$= -\sum_{i=1}^{m} (\vec{y}_{true})_i \log\left(\frac{e^{y_i' - a}}{\sum_{j=1}^{m} e^{y_j - a}}\right)$$

$$= = -\sum_{i=1}^{m} (\vec{y}_{true})_i \log\left(\frac{e^{a}}{e^{a}} \frac{e^{y_i'}}{\sum_{j=1}^{m} e^{y_j}}\right)$$

$$= -\sum_{i=1}^{m} (\vec{y}_{true})_i \log\left(\frac{e^{y_i}}{\sum_{j=1}^{m} e^{y_j}}\right)$$

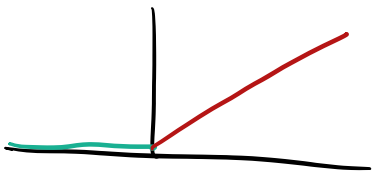
9. See attached

10.

$$f^*(y) = \sup_x \{ y^T x - f(x) \} \quad \forall y \in \mathbb{R}^n$$

What is $f(x)$?

If $\ell \le u \le 0$, then the convex



hull is

thus $f(x) = r \ge 0$

$$\sup_s \{ s_1 \hat{v} \} = 0$$

s.t. $s_1 = 0$

$(s_1, s_2) \in Z_j$

Similarly, if $0 \le \ell \le u$ we have

$v = \hat{v}$ and

$$\sup_s \{ (s_1 - s_2) \hat{v}_j \} = 0$$ by taking the gradient

s.t. $\ell \le s_2$

$s_2 \le u$

$(s_1, s_2) \in Z_j$

11. Problem reduces to

$$\sup_s \begin{cases} s_1 \hat{v} - s_2 v \\ \text{s.t.} \\ (s_1, s_2) \in Z_j \end{cases}$$

line through $(\ell_j, 0)$; (u_j, u_j)

$$\frac{u_j - 0}{u_j - \ell_j} = m$$

$$y = \frac{u_j}{u_j - \ell_j}(x - \ell_j) = \frac{u_j}{u_j - \ell_j} x - \frac{u_j \ell_j}{u_j - \ell_j}$$

$$\rightarrow y = \frac{u_j}{u_j - \ell_j} x - \frac{u_j \ell_j}{u_j - \ell_j}$$

$$s_1 = \frac{u_j s_2}{u_j - \ell_j} - \frac{u_j \ell_j}{u_j - \ell_j}$$

$$\sup_{s_1} \frac{u_j \hat{v} s_2}{u_j - \ell_j} - \frac{u_j \ell_j \hat{v}}{u_j - \ell_j} - s_2 v$$

s.t. $(s_1, s_2) \in Z_j$

Linear, take derivative, we get

$$\frac{u_j \hat{v}}{u_j - \ell_j} = v \quad \text{and} \quad \sup_{s_1} = -\frac{u_j \ell_j \hat{v}}{u_j - \ell_j}$$

$= -\ell_j v$. Thus, since the optimum lies at the border, either it is 0 or $-\ell_j v$. Thus

$$I^*_{Z_j} = \begin{cases} \mathrm{ReLU}(-\ell_j v) & \text{iff } \dfrac{u_j \hat{v}}{u_j - \ell_j} = v \\ +\infty & \text{otherwise} \end{cases}$$