Towards universal limits on adversarial robustness

Hamza Fawzi

DAMTP, Cambridge

May 24, 2019
State-of-the-art classifiers are vulnerable to very small adversarial perturbations.

As of today, there is no provable and scalable way to defend against adversarial perturbations.

Is it possible to construct classifiers that are robust to adversarial perturbations?

Goal: derive upper bounds on the maximal achievable robustness to perturbations.
Prior works: tradeoff robustness-risk

Data-dependent upper bounds.

Theorem (Limit on the robustness of linear classifiers, [Fawzi et al., 2015])

For any linear classifier $f(x) = w^T x$, we have

$$\rho_{adv}(f) \leq \frac{1}{2} \|\mathbb{E}_{\mu_1}(x) - \mathbb{E}_{\mu_{-1}}(x)\|_2 + 2MR(f).$$

Similar tradeoff for quadratic classifiers.
Limits on the adversarial robustness of *any* classifier

Shift our point of view, and do not make any assumptions on the classifier.

Data \rightarrow \text{Classifier} \rightarrow \text{Labels}
Limits on the adversarial robustness of any classifier

Shift our point of view, and do not make any assumptions on the classifier.
Generative model

- **Data model**: The data distribution \( \mu = g(\mathcal{N}(0, I_p)) \), where \( g \) is a smooth generative model.

  Common scheme for generative models (e.g., BigGANS)

- **Classification function** \( f \) can be anything.
Setting illustration
Notions of robustness

- **In-distribution robustness:** Perturbations of latent vectors in $\mathcal{Z}$ measure the amount of change one needs to apply to meaningful latent features to cause data misclassification

$$r_\mathcal{Z}(z) = \min_{r \in \mathcal{Z}} \|r\|_2 \text{ s.t. } (f \circ g)(z) \neq (f \circ g)(z + r).$$

In the image space,

$$r_\mathcal{X}(z) = \min_{r \in \mathcal{Z}} \|g(z + r) - g(z)\| \text{ s.t. } (f \circ g)(z) \neq (f \circ g)(z + r),$$

where $\| \cdot \|$ denotes an arbitrary norm on $\mathcal{X}$. 
Notions of robustness

- **In-distribution robustness:** Perturbations of latent vectors in $\mathcal{Z}$ measure the amount of change one needs to apply to meaningful latent features to cause data misclassification

$$r_\mathcal{Z}(z) = \min_{r \in \mathcal{Z}} \| r \|_2 \text{ s.t. } (f \circ g)(z) \neq (f \circ g)(z + r).$$

In the image space,

$$r_\mathcal{X}(z) = \min_{r \in \mathcal{X}} \| g(z + r) - g(z)\| \text{ s.t. } (f \circ g)(z) \neq (f \circ g)(z + r),$$

where $\| \cdot \|$ denotes an arbitrary norm on $\mathcal{X}$.

- **Unconstrained robustness:**

$$r_{unc}(z) = \min_{r \in \mathcal{X}} \| r \| \text{ s.t. } f(g(z) + r) \neq f(g(z)).$$
Upper bound on the robustness

Theorem (High probability bounds in latent space)

For any classifier $f : \mathcal{X} \rightarrow \{1, \ldots, K\}$

$$
\mathbb{P}_z (r_Z(z) \leq \epsilon) \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2} .
$$

(1)

In the setting where all classes are equiprobable,

$$
\mathbb{P}_z (r_Z(z) \leq \epsilon) \geq 1 - \eta(K) e^{-\epsilon^2/2},
$$

(2)

where $\eta(K) \approx \sqrt{\frac{\pi}{2}} \exp(-\epsilon \sqrt{\log K})$. 

Consequence 1:

The normalized robustness $r_Z(z) \parallel z \parallel_2 = O(d^{-1/2})$, as $\parallel z \parallel_2 \approx \sqrt{d}$ for $z \sim N(0, I_d)$. 

Consequence 2:

Increasing probability of fooling with the number of classes $K$. 

8
Upper bound on the robustness

Theorem (High probability bounds in latent space)

For any classifier $f : \mathcal{X} \to \{1, \ldots, K\}$

$$\mathbb{P}_z (r_Z(z) \leq \epsilon) \geq 1 - \frac{1}{2} \pi \sqrt{\frac{\pi}{2}} e^{-\epsilon^2 / 2} .$$  \hfill (1)

In the setting where all classes are equiprobable,

$$\mathbb{P}_z (r_Z(z) \leq \epsilon) \geq 1 - \eta(K) e^{-\epsilon^2 / 2},$$  \hfill (2)

where $\eta(K) \approx \sqrt{\frac{\pi}{2}} \exp(-\epsilon \sqrt{\log K})$.

Consequence 1: The normalized robustness $\frac{r_Z(z)}{\|z\|_2} = O(d^{-1/2})$, as $\|z\|_2 \approx \sqrt{d}$ for $z \sim \mathcal{N}(0, I_d)$. 
Upper bound on the robustness

Theorem (High probability bounds in latent space)

For any classifier \( f : \mathcal{X} \to \{1, \ldots, K\} \)

\[
\mathbb{P}_z \left( r_Z(z) \leq \epsilon \right) \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2}. \tag{1}
\]

In the setting where all classes are equiprobable,

\[
\mathbb{P}_z \left( r_Z(z) \leq \epsilon \right) \geq 1 - \eta(K) e^{-\epsilon^2/2}, \tag{2}
\]

where \( \eta(K) \approx \sqrt{\frac{\pi}{2}} \exp(-\epsilon \sqrt{\log K}) \).

Consequence 1: The normalized robustness \( \frac{r_Z(z)}{\|z\|_2} = O(d^{-1/2}) \), as \( \|z\|_2 \approx \sqrt{d} \) for \( z \sim \mathcal{N}(0, I_d) \).

Consequence 2: Increasing probability of fooling with the number of classes \( K \).
Idea of proof: Isoperimetric inequality

Theorem (Gaussian isoperimetric inequality)

Let $\gamma_d$ be the Gaussian measure on $\mathbb{R}^d$. Let $A \subseteq \mathbb{R}^d$ and let $A_\eta = \{ z \in \mathbb{R}^d : \exists z' \in A \text{ s.t. } \| z - z' \| \leq \eta \}$. If $\gamma_d(A) = 1/2$ then $\gamma_d(A_\eta) \geq 1 - 1/2 \exp(-\eta^2/2)$.

A tiny widening of any set in the Gaussian space fills most of the space.
How tight is this bound?

Bound on probability becomes \textit{equality} for linear classifier (in the $\mathcal{Z}$ space) $\rightarrow$ \textbf{suggests that classifiers leading to linear separation in the latent space maximize robustness.}

Classifiers that create many disconnected regions in the latent space are much more vulnerable.
Transferability of perturbations

**Theorem (Transferability of perturbations)**

Let $f, h$ be two classifiers. Assume that $\mathbb{P}_z(f \circ g(z) \neq h \circ g(z)) \leq \delta$ (e.g., if $f$ and $h$ have a risk bounded by $\delta/2$ for the data set generated by $g$). Then,

$$\mathbb{P}_z \left( \exists v \in \mathcal{Z} : \|v\|_2 \leq \epsilon, f(g(z + v)) \neq f(g(z)), h(g(z + v)) \neq h(g(z)) \right) \geq 1 - \sqrt{\frac{\pi}{2}} e^{-\epsilon^2/2} - 2\delta .$$

**Interpretation.** When two classifiers have small risk, there exists adversarial perturbations that are *common* to both.
Robustness in the image space

Key assumption: **smoothness** of the generator.

**Assumption**

*We assume that $g$ admits an invertible modulus of continuity $\omega$; i.e.,

$$\forall z, z', \|g(z) - g(z')\| \leq \omega(\|z - z'\|_2).$$

(3)*

Under this assumption, the same results hold for the **robustness in the image space**.
Upper bound on expected robustness in image space

Theorem

Provided $\omega$ is a concave function, we have

$$\mathbb{E}_z r_{\mathcal{X}}(z) \leq \omega \left( \frac{\log(4\pi \log(K))}{\sqrt{2 \log(K)}} \right).$$

Assuming now that the generator $g$ only provides a $\delta$ approximation of the true distribution $\mu$ in the Wasserstein sense; that is, $W(g_d(\gamma_d), \mu) \leq \delta$, the following inequality holds provided $\omega$ is concave

$$\mathbb{E}_{x \sim \mu} (r_{\text{unc}}(x)) \leq \omega \left( \frac{\log(4\pi \log(K))}{\sqrt{2 \log(K)}} \right) + \delta,$$

where $r_{\text{unc}}(x)$ is the unconstrained robustness in the image space.
Relation between in-distribution and out-distribution robustness

- All our bounds are derived for the in-distribution.
- The bounds however trivially hold for out-distribution.
- Is it possible to get a better bound on unconstrained robustness directly?

No.

For any classification function $f$ consider the following modified classifier $\tilde{f}$:

$$\tilde{f}(x) = f(g(z^*)) \quad \text{with} \quad z^* = \arg\min_z \|g(z) - x\|.$$  \hspace{1cm} (4)

**Theorem**

For the classifier $\tilde{f}$, we have $\mathbb{E}_z r_{unc}(z) \geq \frac{1}{2} \mathbb{E}_z r_x(z)$. 


Experiments (SVHN)

For relatively simple tasks, powerful generative models that provide a good approximation of the data distribution exist.

<table>
<thead>
<tr>
<th>Error rate</th>
<th>Upper bound on robustness</th>
<th>2-Layer LeNet</th>
<th>ResNet-18</th>
<th>ResNet-101</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robustness in the $Z$-space</td>
<td>$16 \times 10^{-3}$</td>
<td>$6.1 \times 10^{-3}$</td>
<td>$6.1 \times 10^{-3}$</td>
<td>$6.6 \times 10^{-3}$</td>
</tr>
<tr>
<td>In-distribution robustness</td>
<td>$36 \times 10^{-2}$</td>
<td>$3.3 \times 10^{-2}$</td>
<td>$3.1 \times 10^{-2}$</td>
<td>$3.1 \times 10^{-2}$</td>
</tr>
<tr>
<td>Unconstrained robustness</td>
<td>$36 \times 10^{-2}$</td>
<td>$0.39 \times 10^{-2}$</td>
<td>$1.1 \times 10^{-2}$</td>
<td>$1.4 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

Upper bound is about one order of magnitude larger than the real robustness.
Experiments (CIFAR-10)

Bound predicts that any classifier defined on this task will have perturbations not exceeding 1/10 of the norm of the image.
Conclusions

- We prove that no classifier can be robust to adversarial perturbations when the latent space is sufficiently large and the generative model sufficiently smooth.
- Applicable to any classifier, including human visual system.
- Existence of transferable perturbations across different small-risk classifier.

[Fawzi, et. al., *Adversarial vulnerability for any classifier*, NeurIPS 2018]