Analyzing the Robustness of Deep Learning Against Adversarial Examples

Jun Zhao
Nanyang Technological University

Abstract—Recent studies have shown the vulnerability of many deep learning algorithms to adversarial examples, which an attacker obtains by adding subtle perturbation to benign inputs in order to cause misbehavior of deep learning. For instance, an attacker can add carefully selected noise to a panda image so that the resulting image is still a panda to a human being but is predicted as a gibbon by the deep learning algorithm. As a first step to propose effective defense mechanisms against such adversarial examples, we analyze the robustness of deep learning against adversarial examples. Specifically, we prove a strict lower bound for the minimum $\ell_p$ distortion of a data point to obtain an adversarial example.

Keywords—Deep learning, adversarial examples, robustness, neural networks.

I. INTRODUCTION

Recently, a surge of studies [1]–[8] have presented the vulnerability of many deep learning systems under adversarial examples, which an attacker crafts by adding subtle perturbation to benign inputs in order to cause misbehavior of deep learning.

Among adversarial examples, those on image or object recognition systems have received much attention. Goodfellow et al. [1] show that adding carefully selected noise to a panda image can induce an image which is still considered by a human being as a panda, but is recognized by the neural network as a gibbon. Evtimov et al. [2] demonstrate attacks on neural network-based AI systems in self-driving cars. An attacker can put stickers such as “love” and “hate” on a stop sign and trick AI to believe that it is a speed limit sign. This can cause accidents if the car does not stop before the sign.

Adversarial examples also exist in the audio domain. Carlini and Wagner [3] present an audio adversarial example by adding subtle perturbations to the benign audio clip. For example, for a the benign audio clip reads “Without the dataset, the article is useless”, adding small noise to the clip gives the same sentence to a human, but the Google assistant understands the result as “Okay Google, browse to evil dot com” and hence will visit evil.com.

Kurakin et al. [4] introduce adversarial examples in the physical world. For example, adding small perturbation to a library photo makes the neural network predict a prison, and adding small perturbation to a washer image tricks the neural network to output doormat.

We discuss below the intuition for the existence of adversarial examples. As illustrated in Figure 1 on the next page, we have several data points, where a dot represents a cat image, and a triangle represents a dog image. Suppose the blue line is given by the deep learning-based AI system so that AI considers data above this line as a cat and considers data below this line as a dog. This AI system does a prefect job to predict existing data points. However, the true distribution for cat may look like the purple cloud shape, while the true distribution for dog may look like the yellow cloud shape. Hence, data in these two dashed parts will be predicted incorrectly by AI. For instance, the attacker may add noise to the green data point and obtain the red point as an adversarial example. This adversarial example will trigger AI to have misbehavior. Specifically, this image is a cat, but the AI system considers it as a dog. Similarly, the blue triangle represents an adversarial example, which is a dog, but is considered as a cat by AI.

Two challenges in the research of attack-resilient AI systems are as follows: 1) establishing resilience benchmarks against adversarial examples, and 2) proposing effective defenses against adversarial examples. Our work provides an initial step to address the two challenges, by analyzing the robustness of deep learning against adversarial examples.

The rest of the paper is organized as follows. In Section II, our Theorem 1 presents the result on the minimum distortion of a data point to obtain an adversarial example. We provide a detailed proof in Section III. Section IV surveys related work, and Section V concludes the paper.
II. THE RESULT ON THE MINIMUM DISTORTION OF A DATA POINT TO OBTAIN AN ADVERSARIAL EXAMPLE

Theorem 1. Let $D$ be the domain of data points, and $f : D \to \mathbb{R}^k$ be a multi-class classifier with $f = [f_1, f_2, \ldots, f_k]$ for continuously differentiable $f_j$, where $j \in \{1, 2, \ldots, k\}$. Define $c$ as the class predicted by classifier $f$ for data point $x$; i.e., $c = \arg\max_{j \in \{1, 2, \ldots, k\}} f_j(x)$. Let $\|\delta\|_{p_{\text{MinAdv}}}$ denote the minimum $\ell_p$ distortion of data point $x$ to obtain an adversarial example. Then we can prove that $\|\delta\|_{p_{\text{MinAdv}}}$ is greater than

$$\min_{j \in \{1, 2, \ldots, k\} \setminus \{c\}} \frac{f_c(x) - f_j(x)}{\max \left\{ \left\| \frac{\partial}{\partial z} [f_c(z) - f_j(z)] \right\|_q : z \in D \right\} }.$$ (1)

where $\frac{1}{p} + \frac{1}{q} = 1$; i.e., the term in (1) is a strict lower bound for $\|\delta\|_{p_{\text{MinAdv}}}$.

Concurrent studies by Hein and Andriushchenko [9] and Weng et al. [10] obtain similar results, but their results are more complex than ours and rely on either more involved or different proofs. In particular, with $D$ being $\mathbb{R}^d$ for some $d$, after defining $B_p(x, R) = \{ z : \|x - z\|_p \leq R \}$, Hein and Andriushchenko [9] replace (1) by

$$\max_{R > 0} h(R)$$

for $h(R)$ denoting $\min \left\{ \min_{j \in \{1, 2, \ldots, k\} \setminus \{c\}} f_c(x) - f_j(x) \right\}$

$$\max \left\{ \left\| \frac{\partial}{\partial z} [f_c(z) - f_j(z)] \right\|_q : z \in B_p(x, R) \right\} .$$

(2)

(2) is more complex yet tighter than (1). In fact, $h(R)$ of (2) in the special case of $R$ being $\infty$ reduces to (1). Weng et al. [10] assume Lipschitz continuity of $f_c(z) - f_j(z)$ in the $\ell_q$ norm; i.e., there exists $L_q^p$ such that

$$\|f_c(z_1) - f_j(z_1) - [f_c(z_2) - f_j(z_2)]\|_q \leq L_q^p \|z_1 - z_2\|_q$$

for any $z_1$ and $z_2$. Then [10] replaces (1) by

$$\min_{j \in \{1, 2, \ldots, k\} \setminus \{c\}} \frac{f_c(x) - f_j(x)}{L_q^p} .$$

(3)

The proofs for (1) and (2) are similar, while establishing the latter in [9] is more involved. In contrast, the proof in [10] for (3) uses a different technique.

III. PROOF OF THEOREM 1

Suppose that with a distortion $\delta$, the data point $x$ is changed to an adversarial example $x + \delta$. Since $c$ is the class predicted by classifier $f$ for data point $x$, the classifier $f$ predicts a class $j \in \{1, 2, \ldots, k\} \setminus \{c\}$ for the adversarial example $x + \delta$. This implies that

there exists $j \in \{1, 2, \ldots, k\} \setminus \{c\}$

such that $f_j(x + \delta) > f_c(x + \delta)$. (4)

For notation simplicity, we define

$$g_j(x) := f_c(x) - f_j(x)$$

(5)

so that

$$g_j(x + \delta) := f_c(x + \delta) - f_j(x + \delta).$$

(6)

Then (4) and (6) mean

there exists $j \in \{1, 2, \ldots, k\} \setminus \{c\}$

such that $g_j(x + \delta) < 0$. (7)

To analyze (7), we will bound $g_j(x + \delta)$. To this end, below we analyze $g_j(x) - g_j(x + \delta)$. For notation simplicity, we define

$$y := x + \delta,$$ (8)
and write

\[ x = [x_1, x_2, \ldots, x_d], \quad (9) \]
\[ y = [y_1, y_2, \ldots, y_d]. \quad (10) \]

Then we have

\[ |g_j(x) - g_j(y)| \]
\[ = |g_j(x_1, x_2, \ldots, x_d) - g_j(y_1, y_2, \ldots, y_d)| \]
\[ + \ldots + |g_j(y_1, x_2, \ldots, x_d) - g_j(y_1, y_2, x_3, \ldots, x_d)| \]
\[ = \sum_{i=1}^{d} \left| \frac{g_j(y_1, x_2, \ldots, y_{i-1}, x_i, \ldots, x_d) - g_j(y_1, y_2, \ldots, y_{i-1}, y_i, \ldots, y_d)}{x_i - y_i} \right| |x_i - y_i|. \quad (11) \]

Using the inequality of triangle in (11), we obtain

\[ |g_j(x) - g_j(y)| \]
\[ \leq \sum_{i=1}^{d} \left| \frac{g_j(y_1, x_2, \ldots, y_{i-1}, x_i, \ldots, x_d) - g_j(y_1, y_2, \ldots, y_{i-1}, y_i, \ldots, y_d)}{x_i - y_i} \right| |x_i - y_i|. \quad (12) \]

From the mean value theorem, there exists \( z_i \in [\min\{x_i, y_i\}, \max\{x_i, y_i\}] \) such that after defining

\[ \partial g_j^{(i)}(y_1, y_2, \ldots, y_{i-1}, x_i, x_{i+1}, \ldots, x_d) \]
\[ = \left( \left. \frac{\partial g_j(y_1, y_2, \ldots, y_{i-1}, x_i, x_{i+1}, \ldots, x_d)}{\partial \alpha} \right|_{\alpha = z_i} \right), \quad (13) \]

we have

\[ g_j(y_1, y_2, \ldots, y_{i-1}, z_i, x_{i+1}, \ldots, x_d) - g_j(y_1, y_2, \ldots, y_{i-1}, y_i, x_{i+1}, \ldots, x_d) \]
\[ = \partial g_j^{(i)}(y_1, y_2, \ldots, y_{i-1}, z_i, x_{i+1}, \ldots, x_d) \cdot |x_i - y_i|. \quad (14) \]

Substituting (14) into (12), we have

\[ |g_j(x) - g_j(y)| \]
\[ = \sum_{i=1}^{d} \left[ \partial g_j^{(i)}(y_1, y_2, \ldots, y_{i-1}, z_i, x_{i+1}, \ldots, x_d) \cdot |x_i - y_i| \right]. \quad (15) \]

From \( \frac{1}{p} + \frac{1}{q} = 1 \), the application of Hölder’s inequality to (12) further induces

\[ |g_j(x) - g_j(y)| \]
\[ \leq \left\{ \sum_{i=1}^{d} \left| \partial g_j^{(i)}(y_1, y_2, \ldots, y_{i-1}, z_i, x_{i+1}, \ldots, x_d) \right|^q \right\}^{1/q} \]
\[ \times \left\{ \sum_{i=1}^{d} (|x_i - y_i|)^p \right\}^{1/p}. \quad (16) \]

We now evaluate the terms in (16).

First, for \( z = [z_1, z_2, \ldots, z_d] \in D \), after defining

\[ \partial g_j^{(i)}(z_1, z_2, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_d) : i \in \{1, 2, \ldots, d\}, \quad (17) \]

where

\[ \partial g_j^{(i)}(z_1, z_2, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_d) := \left( \left. \frac{\partial g_j(z_1, z_2, \ldots, z_{i-1}, \alpha, z_{i+1}, \ldots, z_d)}{\partial \alpha} \right|_{\alpha = z_i} \right), \quad (18) \]

we note

\[ \left\{ \sum_{i=1}^{d} \left[ \partial g_j^{(i)}(y_1, y_2, \ldots, y_{i-1}, z_i, x_{i+1}, \ldots, x_d) \right]^q \right\}^{1/q} \]
\[ \leq \max \left\{ \left\| \partial g_j(z) \right\|_q : z \in D \right\}. \quad (19) \]

Also, it is clear that

\[ \left\{ \sum_{i=1}^{d} (|x_i - y_i|)^p \right\}^{1/p} = \|x - y\|_p. \quad (20) \]

Substituting (19) and (20) into (16), we obtain

\[ |g_j(x) - g_j(y)| \]
\[ \leq \max \left\{ \left\| \partial g_j(z) \right\|_q : z \in D \right\} \times \|x - y\|_p. \quad (21) \]

Recalling \( y := x + \delta \) as in (8), we write (21) as

\[ |g_j(x + \delta) - g_j(x)| \]
\[ \leq \max \left\{ \left\| \partial g_j(z) \right\|_q : z \in D \right\} \times \|\delta\|_p, \]

which means

\[ g_j(x) - \max \left\{ \left\| \partial g_j(z) \right\|_q : z \in D \right\} \times \|\delta\|_p \]
\[ \leq g_j(x + \delta) \]
\[ \leq g_j(x) + \max \left\{ \left\| \partial g_j(z) \right\|_q : z \in D \right\} \times \|\delta\|_p. \quad (22) \]

As explained in (7), for an adversarial example \( x + \delta \), there exists \( j \in \{1, 2, \ldots, k \} \setminus \{c\} \) such that \( g_j(x + \delta) < 0 \). Using
there exists $j \in \{1, 2, \ldots, k\} \setminus \{c\}$ such that
$g_j(x) - \max \{\|\partial g_j(z)\|_q : z \in D\} \times \|\delta\|_p < 0$.  
(23)

Then we obtain from (23) that

$$
\|\delta\|_p > \min_{j \in \{1, 2, \ldots, k\} \setminus \{c\}} \frac{g_j(x)}{\max \{\|\partial g_j(z)\|_q : z \in D\}}.  
(24)
$$

Recalling the definition $g_j(x) := f_c(x) - f_j(x)$ in (5), we write (24) as

$$
\|\delta\|_p > \min_{j \in \{1, 2, \ldots, k\} \setminus \{c\}} \frac{f_c(x) - f_j(x)}{\max \{\|\partial f_c(z) - \partial f_j(z)\|_q : z \in D\}}.  
(25)
$$

Let $\|\delta\|_{\text{MinAdv}}^p$ denote the minimum $\ell_p$ distortion of data
point $x$ to obtain an adversarial example. Then (25) implies

$$
\|\delta\|_{\text{MinAdv}}^p > \min_{j \in \{1, 2, \ldots, k\} \setminus \{c\}} \frac{f_c(x) - f_j(x)}{\max \{\|\partial f_c(z) - \partial f_j(z)\|_q : z \in D\}}.  
$$


\section{IV. RELATED WORK}

\subsection{A. Different Kinds of Adversarial Examples}

Athalye and Sutskever [6] show the existence of 3D adversarial examples. A novel algorithm to synthesize adver-
sarial examples for a selected distribution of transformations
is also proposed. Sharif et al. [11] present adversarial
examples against facial biometric systems. Cubuk et al. [7]
argue the inherent uncertainty in neural networks’ pre-
dictions as the reason why adversarial examples exist.
Balda et al. [12] leverage convex programming to gener-
ate adversarial examples under various constraints and for
different network types. Nguyen et al. [13] produce images
that are unrecognizable to human beings, but recognizable
by DNNs. Moosavi-Dezfooli et al. [14] introduce the Deep-
Fool algorithm to construct adversarial examples efficiently.
Zhao et al. [15] define a framework to generate legible and
natural adversarial examples. Brendel et al. [16] discuss the
significance of attacks which rely on only the final model
decision.

\subsection{B. Robustness against Adversarial Examples}

Kolter and Wong [17] introduce a method to train
deep ReLU-based classifiers such that they are provably
robust against adversarial examples with bounded norms.
Wong et al. [18] scale adversarial defense mechanisms
to larger models. Alvarez-Melis and Jaakkola [19] show
that the robustness of explanations is crucial for inter-
pretability. Under mild assumptions, Gal and Smith [20]
prove that idealized models are not vulnerable to adversarial
examples. Wang et al. [21] use bias-variance theory to un-
derstand adversarial examples. Raghunathan et al. [22]
leverage semidefinite relaxation to analyze adversarial
examples. Wang et al. [23] apply notions from topology to
quantify adversarial examples. Concurrent studies by
Hein and Andriushchenko [9] and Wang et al. [21] obtain results
similar to this paper, but their results are more complex than
ours and rely on either more involved or different proofs.
Fawzi et al. [24] discuss a diverse set of perturbations.
Carlini and Wagner et al. [25] argue that defensive distillation
of [26] does not work well in several cases.

\subsection{C. Defenses against Adversarial Examples}

Papernot et al. [26] propose defensive distillation to
counter adversarial examples. Huang et al. [27] present the
approach of learning with a strong adversary, where robust
classifiers are learnt from supervised data. Kurakin et al. [28]
scale adversarial training to ImageNet. Goodfellow et al. [1]
emphasize the linear nature of neural networks as a cause
of adversarial examples. Gu and Rigazio [29] investigate the
structure of adversarial examples and propose defenses ac-
cordingly. Meng and Chen [30] present the MagNet defense
framework, which can approximate normal examples’ mani-
fold. Song et al. [31] provide an empirical evaluation of ad-
versarial examples. He et al. [32] combine multiple defenses
to obtain a strong defense mechanism. Tramèr et al. [33]
show that adversarial training cannot defend against black-
box attacks. Cullina et al. [34] extend the Probably Ap-
proximately Correct (PAC)-learning framework to tackle an
adversary’s presence.

\section{V. CONCLUSION}

Many recent studies have demonstrated the vulnerability
of deep learning algorithms to adversarial examples, which
an attacker generates by adding subtle noise to benign inputs
in order to cause incorrect prediction by deep learning. For
example, an adversary can put stickers such as “love” and
“hate” on a stop sign and trick the deep learning system to
recognize it as a speed limit sign. In this paper, we analyze
the robustness of deep learning against adversarial examples.
Specifically, we prove a strict lower bound for the minimum
\( \ell_p \) distortion of a data point to obtain an adversarial example. Future directions include the development of effective defense mechanisms for various neural network models to counter adversarial examples.

REFERENCES


