CF-012

Target Neural Network のロバスト性が AdvGAN に及ぼす影響について
The Effect of Robustness of Target Neural Network on AdvGAN

西山 圭亮† 島田 英之†
Keisuke Nishiyama Hideyuki Shimada

1. 序論

機械学習技術はますます多くのドメインで重要な役割を担っている. 自律走行車や言語翻訳などの技術は, 機械学習をその中核に据えている. 畳み込みニューラルネットワーク (Convolutional Neural Networks) が ImageNet Large Scale Visual Recognition Competition (ILSVRC) で初期の成功以来, 深層学習は画像分類, セグメンテーション, 物体検出, 音声認識, 言語翻訳など, 多数の課題にうまく適用されている. しかし, 現代の機械学習技術は様々な複雑な課題の解決に成功しているにもかかわらず, 未解明の部分も多く, それゆえに機械学習特有のセキュリティ上の弱点が見つかっており, 現在盛んに研究されている. 近年の研究によって, 機械学習は入力データに微小な摂動を加えた敵対的サンプル (Adversarial Examples) に対して脆弱であることが示されている. 敵対的サンプルによる誤認識の例を図 1 に示す. 左は正常な画像であり, 機械学習を用いた画像認識モデルはこの画像を「パンダ」と判断する. しかし, 微小な摂動を意図的に加えた右の画像を画像認識モデルは「テナガザル」と誤認識を起こしてしまう.

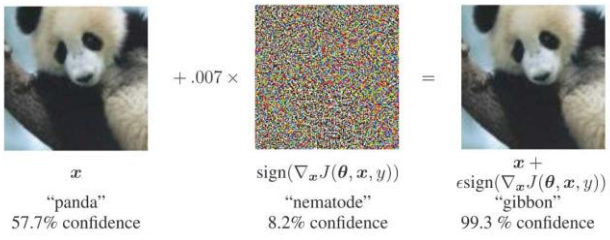

図 1　敵対的サンプルによる誤認識の例[1]

近年, ニューラルネットワーク (Neural Networks) など機械学習技術の著しい発展により, 画像認識などの応用において人間並みの判断能力を持つシステムが生まれている. 一方で, 意図的な攻撃に晒される環境では, 前述した敵対的サンプルのように思わぬ脆弱さを露呈することも分かってきた. 機械学習技術を安心して利用するためには, このような攻撃がどの程度可能なのか, なぜ可能なのか, 攻撃を防ぐにはどうしたらよいかを明らかにする必要がある. そのためには, 攻撃手法についての研究が有効解決策のひとつである. 敵対的サンプルによる攻撃手法は多数提案されているが, そのひとつとして, 敵対的生成ネットワーク (Generative Adversarial Network : GAN) を利用して敵対的サンプルを生成する AdvGAN [2]が提案されている. AdvGAN は一般的な Generator と Discriminator から構成される GAN とは異なり, Target Neural Network を導入した 3 つのネットワークから構成されている. Target Neural Network は一般的な画像認識モデルと同様なアーキテクチャである. Xiao ら [2]は Target Neural Network の学習データセットに Fast Gradient Sign Method (FGSM) [1]から生成した敵対的サンプルを加えている. 我々の先行研究[3]でも Target Neural Network の学習データセットに FGSM の他に, Jacobian-based Saliency Map Attack (JSMA) [4], Boundary Attack [5]より生成した敵対的サンプルを加えることを提案し, より優れた攻撃性能を獲得できることを示した. Xiao らの研究と我々の先行研究では同様に, Target Neural Network の学習データセットに敵対的サンプルを加えて, Target Neural Network のロバスト性を向上させている. しかし, ロバスト性が向上した Target Neural Network を用いた場合に, そのロバスト性が AdvGAN に及ぼす影響が不明瞭である. そこで, ロバスト性の違いによって, 敵対的サンプルの攻撃性能の向上を図ることができるか検証することが必要であろう.

本稿では, Target Neural Network のロバスト性が AdvGAN に及ぼす影響について実験的に確認し, その有効性について検証する. より具体的には, Target Neural Network の学習データセットとして, オリジナル画像データの他に, FGSM, JSMA, Carlini & Wagner Attack (C&W Attack) [6], Projected Gradient Descent Attack (PGDA) [7]から生成した敵対的サンプルを個別に用いて, 学習を行う.

†岡山理科大学　Okayama University of Science

2. 関連研究

2.1 敵対的サンプル

敵対的サンプルとは, 機械学習モデルへの入力の中でも特に, 人間にとって有効な入力に似ているにもかかわらず, モデルの予測を誤らせるように設計されたものである. 敵対的サンプルは, ターゲットモデルfを欺きながら, 距離メトリック$d(\cdot)$の下で$\tilde{x}$と非敵対的入力xとの差が最小となる入力$\tilde{x}$と定義される. 一般的に, 敵対的サンプルは次式,

$$d(\tilde{x}, x) < \epsilon \text{ such that } \hat{y}(\tilde{x}) \neq \hat{y}(x) \quad (1)$$

を満たすことを求める. ここで, ϵは摂動の大きさを制限する小さな定数であり, $\hat{y}(\cdot)$は分類モデルの予測ラベルを表す（すなわち, $\hat{y}(x) = \arg\max_c f(x)_{(c)}$）.

敵対的サンプルによる攻撃の種類は, ホワイトボックス攻撃とブラックボックス攻撃に分類される. ホワイトボックス攻撃は, 攻撃者はターゲットモデルのアーキテクチャ, パラメータの値, 学習手順, 学習データなどの情報が完全に把握できることを前提とした攻撃である. 一方, ブラックボックス攻撃はターゲットモデルの出力にのみアクセスし, その内部にはアクセスできないことを前提とした攻撃である. 攻撃者が被害者の内部を知ることはほとんどないため, ブラックボックス攻撃の方がより現実的な想定である. 敵対的サンプルによる攻撃は, さらに標的型攻撃と非標的型攻撃に分類することができる. 標的型攻撃では, 敵対的サンプルは特定の誤分類先のクラスに分類されるように設計される. 一方, 非標的型攻撃では, 敵対的サンプルのクラスに

関係なく，オリジナル以外のクラスに分類させるように設計される．敵対的サンプルは通常，ノルムを用いてノイズを評価する．敵対的サンプルのノイズを評価する指標として，利用されるのは，$\|\cdot\|_0$ ，$\|\cdot\|_2$ ，$\|\cdot\|_{\inf}$ の３つのノルムである．

2.2 AdvGAN

AdvGAN とは，GAN を用いて，敵対的サンプルを生成する生成手法である．AdvGAN は Neural Network を用いて，敵対的サンプルを導く摂動を生成することを目的として設計されている．AdvGAN では，Generator にオリジナルインスタンスを入力すると，摂動を出力する．また，Discriminator を追加することによって，GAN のフレームワークとして学習している．AdvGAN の概観を図２に示す．

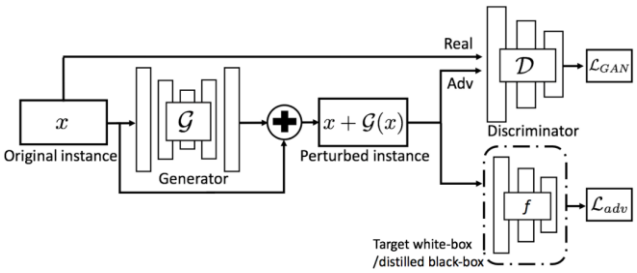

図２　AdvGAN の概観[2]

AdvGAN の目的関数は，classifier loss (L_{adv}) ，anti-discriminator GAN loss (L_{GAN}) , hinge loss (L_{hinge}) などの複数の項から成り立ち，Lは次のように表される．

$$L = L_{adv} + \alpha L_{GAN} + \beta L_{hinge} \tag{2}$$

また，L_{adv} , L_{GAN} , L_{hinge} は次のように表される．

$$L_{adv} = \mathbb{E}_x\left[L_f(x + G(x), t)\right] \tag{3}$$

$$L_{GAN} = \mathbb{E}_x[\log D(x)] + \mathbb{E}_x\left[\log\left(1 - D\left(x + G(x)\right)\right)\right] \tag{4}$$

$$L_{hinge} = \mathbb{E}_x\left[\max\left(0, \|G(x)\|_2 - c\right)\right] \tag{5}$$

ここで，αとβは損失項の重み付けを制御する定数，L_fは f が使用する分類損失関数 (例えば，cross entropy) , tはターゲットラベル，c は hinge loss の境界を表す定数である．

2.3 Adversarial Attacks

Adversarial Attacks とは，敵対的サンプルを生成する手法である．

2.3.1 Fast Gradient Sign Method

Fast Gradient Sign Method (FGSM) とは，与えられた入力に対して，ターゲットモデルの損失関数が増加するような摂動方向を素早く見つけ，分類の信頼度を下げ，クラス間混同の可能性を高めるように設計された生成手法である．損失を一定量増加させると誤分類になるという保証はないが，誤分類されたインスタンスの損失値は，そうでない場合よりも定義上大きくなるため，誤分類に導く効果が期待できる手法であるといえる．

FGSM は，入力に対する損失関数の勾配を計算し，選択した小さな定数と勾配の符号ベクトルを掛け合わせることによって小さな摂動を作り出すことで動作する．FGSM によって生成される敵対的サンプル$\tilde{x}$は，次のように表される．

$$\tilde{x} = x + \varepsilon\text{sign}\left(\nabla_x L(\theta, x, y)\right) \tag{6}$$

ここで，$\nabla_x L(\theta, x, y)$は，入力xに関する損失関数の導関数である．

2.3.2 Jacobian-based Saliency Map Attack

Jacobian-based Saliency Map Attack (JSMA) とは，誤分類を引き起こすために，一部の入力特徴をわずかに摂動させることで敵対的サンプルを生成する．これは，FGSM のように，入力特徴のすべてではないにせよ，ほとんどを変更する攻撃とは対照的である．

Neural Network による分類器から得られる予測されたソフトマックス確率ベクトル$f(x)$に基づいて，Saliency Map は次のように表される．

$$S\left(x_{(i)}, t\right) = \begin{cases} 0 \text{ if } \nabla_{x_{(i)}} f(x)_{(t)} < 0 \text{ or } \sum_{c \neq t} \nabla_{x_{(i)}} f(x)_{(c)} > 0 \\ -\nabla_{x_{(i)}} f(x)_{(t)} \sum_{c \neq t} \nabla_{x_{(i)}} f(x)_{(c)} \text{ otherwise} \end{cases} \tag{7}$$

ここで，$x_{(i)}$はxのi番目の要素を示し，tは攻撃の対象となる特定のラベルを表す．Saliency Map を元に，ターゲットクラスに対して重要度の高い入力要素を特定し，それらの要素を変更することで摂動を生成する．

2.3.3 Carlini & Wagner Attack

Carlini & Wagner Attack (C&W Attack) とは，多様な類似性メトリックを最小化する摂動を生成することを目的としている．C&W Attack の定式化は次のように表される．

$$\underset{w}{\text{minimize}}(\|\tilde{x}(w) - x\|_2^2 + c\, L_{CW}(\tilde{x}(w), t))$$
$$\text{where } \tilde{x}(w) = \frac{1}{2}(\tanh(w) + 1) \tag{8}$$

ここで，wは，$\tilde{x} = \frac{1}{2}(\tanh(w) + 1)$となるような変数の変更であり，$\tilde{x}$を$[0,\ 1]$内に制約するために導入される．また，$L_{CW}$は次のように表される．

$$L_{CW} = \max\left(\max_{i \neq t}\{Z(\tilde{x})_{(i)}\} - Z(\tilde{x})_{(t)}, -k\right) \tag{9}$$

ここで，$Z(\tilde{x})_{(i)}$は分類器の logits の i 番目の成分，t はターゲットラベル，k は敵対的サンプルに対する最小限の望ましい信頼マージンを反映するパラメータを表す．

2.3.4 Projected Gradient Descent Attack

Projected Gradient Descent Attack (PGDA) とは，Projected Gradient Descent という最適化アルゴリズムを利用した生成手法である．PGDA は FGSM を改良した手法であり，課せられた制約に違反しないように FGSM を複数回繰り返しながら生成する手法である．PGDA によって生成される敵対的サンプルは，次のように表される．

$$x^{t+1} = \prod_{x+S}(x^t + \alpha\text{sign}\left(\nabla_x L(\theta, x^t, y)\right) \tag{10}$$

ここで，$\nabla_x L(\theta, x, y)$は，入力xに関する損失関数の導関数である．

2.4 Adversarial defenses

Adversarial defenses とは，敵対的サンプルに対して防御する方法である．

2.4.1 Adversarial Training

Adversarial Training（敵対的学習）とは，学習イタレーションごとに敵対的サンプルを学習データセットに注入することによって，モデルを再学習させる防御法である．新たな学習損失関数$\tilde{L}$は次のように表される．

$$\tilde{L}(\theta, x, y) = \alpha L(\theta, x, y) + (1 - \alpha) L(\theta, \tilde{x}, y) \tag{11}$$

ここで，$L(\theta, x, y)$は分類損失関数，xは正常な入力，$\tilde{x}$は敵対的な入力を表す．αは正常な入力と敵対的な入力間の損失項の重み付けを制御する定数であり，αは通常 0.5 に設定される．

3. 検証内容と評価方法

Xiao らの研究と我々の先行研究では同様に, Target Neural Network を敵対的学習することによって, Target Neural Network のロバスト性を向上させ, 敵対的サンプルに対して耐性を上げた. ロバスト性を向上させるために, Xiao らの研究では, Target Neural Network の学習データセットとして, オリジナル画像データと, オリジナル画像データを元に FGSM によって生成した敵対的サンプルを扱っている. 我々の先行研究では, Target Neural Network の学習データセットとして, オリジナル画像データと, オリジナル画像データを元に FGSM, JSMA, Boundary Attack によって生成した敵対的サンプルを扱っている. Xiao らの研究と我々の先行研究の両者ともに, Target Neural Network にロバストなモデルを用いて, AdvGAN の学習をしているが, Target Neural Network のロバスト性が AdvGAN に及ぼす影響が不明瞭である.

よって, 本実験では Target Neural Network のロバスト性が AdvGAN に及ぼす影響について検証する. 評価方法として, 攻撃成功率について比較する.

4. 実験内容

本実験では学習対象のデータセットに手書き数字から構成された MNIST, 物体を中心としたカラー画像から構成された CIFAR-10 の 2 種類を利用した. また, 各画像の画素値を-1 から 1 の範囲に正規化した.

まず, 敵対的学習を行うため, 学習データセットとして必要となる敵対的サンプルを FGSM, JSMA, C&W Attack, PGDA を用いて生成する.

続いて, 6 種類の Target Neural Network を構築する. そのうちの 4 種類には, Target Neural Network の学習データセットに, オリジナル画像データと, 先程, 生成した敵対的サンプルを加える. つまり, Target Neural Network の学習データセットにオリジナル画像データと FGSM から生成した敵対的サンプルを用いた場合と, オリジナル画像データと JSMA から生成した敵対的サンプルを用いた場合等, 上記の 4 種類の生成手法を用いて, 4 種類の Target Neural Network の敵対的学習をして, ロバスト性を向上させる. 残る 2 種類には, Target Neural Network の学習データセットにオリジナル画像データのみと, オリジナル画像データと 4 種類の生成手法から生成した敵対的サンプルを用いる.

そして, 学習を終えた Target Neural Network を用いて, AdvGAN を学習する. このとき, AdvGAN の Generator や Discriminator のアーキテクチャなど, Target Neural Network の学習データセットに用いた学習データ以外は, すべて同様とする.

その後, 学習を終えた Generator を用いて, 異なるターゲットクラスに対する敵対的サンプルを生成し, それらに対する攻撃成功率の比較を行う.

5. 実験結果

ロバスト性が異なった 6 種類の Target Neural Network を用意した. この 6 種類の Target Neural Network を用いて, AdvGAN を学習した. そして, 学習を終えた Generator を用いて, オリジナル画像からオリジナル画像のクラスとは異なるターゲットクラスに対して, 敵対的サンプルを生成した.

まず, 学習を終えた Generator の評価として, アーキテクチャが異なる分類モデルに対して, 各種モデルに対する攻撃成功率を示す. 分類モデルは学習データセットにオリジナル画像のみを用いて, 学習したモデルである. 表 1 に MNIST における各種モデルの正解精度, および, 敵対的サンプルの攻撃成功率を示す. 表 2 に CIFAR-10 における各種モデルの正解精度, および, 敵対的サンプルの攻撃成功率を示す. None とは, 学習データセットがオリジナル画像データのみで学習した Target Neural Network を用いて, 学習をした AdvGAN である. All とは, 学習データセットにオリジナル画像データと 4 種類の生成手法で敵対的学習した Target Neural Network を用いて, 学習した AdvGAN である.

表 1　MNIST によるオリジナルデータに対する各種モデルの精度, および AdvGAN で各種モデルに対して生成した敵対的サンプルの攻撃成功率

Model		A	B	C
Accuracy[%]		98.96	99.06	99.24
Attack Success Rate[%]	**Adversarial Examples of AdvGAN's training data**			
	None	21.18	22.28	**22.32**
	FGSM	21.54	21.88	21.06
	JSMA	20.91	22.46	20.44
	C&W Attack	**22.16**	20.72	20.54
	PGDA	21.70	21.44	21.18
	All	21.52	**23.04**	21.12

表 2　CIFAR-10 によるオリジナルデータに対する各種モデルの精度, および AdvGAN で各種モデルに対して生成した敵対的サンプルの攻撃成功率

Model		D	VGG16
Accuracy[%]		74.62	76.78
Attack Success Rate[%]	**Adversarial Examples of AdvGAN's training data**		
	None	**83.04**	**86.28**
	FGSM	55.82	61.24
	JSMA	63.44	70.36
	C&W Attack	60.36	66.74
	PGDA	58.09	58.86
	All	63.88	67.62

表 1 より, 学習対象のデータセットに MNIST を利用した場合, Generator によって生成される敵対的サンプルの攻撃性能にあまり変化はないことが確認できた. しかし, 表 2 より, 学習対象のデータセットに CIFAR-10 を利用した場合, 学習データセットに敵対的サンプルを含むかどうかで, 攻撃性能に大きく変化があることが確認できた.

次に, Generator で生成した摂動を付与した際の, 全入力データ数に対する予測が変化した入力データの数の割合を計測した. 摂動の大きさ epsilon は 0 から 1.0 まで, 0.01 刻みで計測した. このとき, 生成する摂動は入力データのクラスとは異なるクラスをランダムに決定し, そのクラスをターゲットクラスとする. 図 3 に学習データセットに MNIST を用いた Generator によって, 計測した結果を示す. 図 4 に学習データセットに CIFAR-10 を用いた Generator によって, 計測した結果を示す.

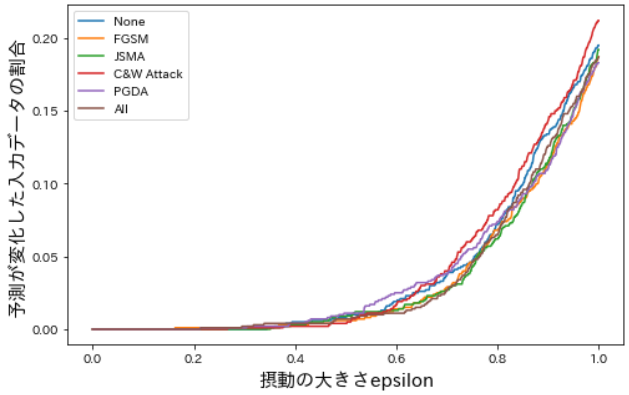

図３　MNIST による大きさ epsilon の摂動を加えた際に予測が変化した入力データ数の割合の推移

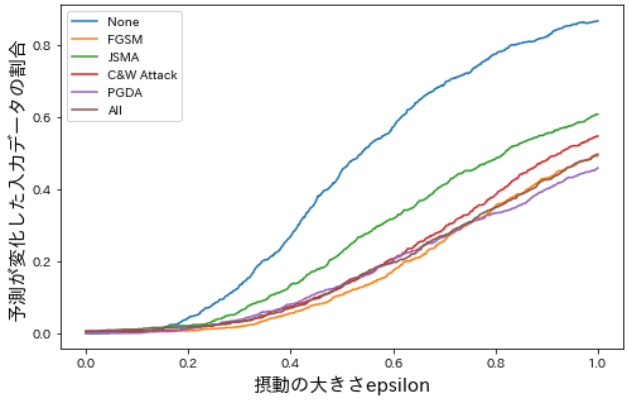

図４　CIFAR-10 による大きさ epsilon の摂動を加えた際に予測が変化した入力データ数の割合の推移

図３より, 同様に学習対象のデータセットが MNIST において, 摂動の大きさを変化させたとしても, 生成される敵対的サンプルの性能にあまり変化はないことが確認できた. しかし, 図４より, 学習対象のデータセットが CIFAR-10 において, Target Neural Network を敵対的学習することによって, ロバスト性を向上させた場合による Generator が生成する摂動は, 学習データセットに使用した敵対的サンプルによって, 予測が変化した入力データ数の割合が異なることが確認できた.

そして, FGSM, JSMA, C&W Attack, PGDA の 4 種類の生成手法を用いて敵対的学習を行い, ロバスト性を向上させたモデルに対する攻撃成功率を示す. 表３に MNIST における攻撃成功率を示す. 表４に CIFAR-10 における攻撃成功率を示す.

表３　MNIST による AdvGAN で各種モデルに対して生成した敵対的サンプルの攻撃成功率

Model		FGSM	JSMA	C&W Attack	PGDA
Accuracy[%]		99.26	99.24	99.16	99.28
Attack Success Rate[%] Adversarial Examples of AdvGAN's training data	None	18.64	18.58	18.68	17.34
	FGSM	18.52	17.11	18.32	18.09
	JSMA	18.02	**18.98**	**19.02**	18.68
	C&W Attack	**19.96**	18.36	18.56	18.82
	PGDA	18.08	18.64	17.80	19.06
	All	19.32	18.36	18.68	**21.09**

表４　CIFAR-10 による AdvGAN で各種モデルに対して生成した敵対的サンプルの攻撃成功率

Model		FGSM	JSMA	C&W Attack	PGDA
Accuracy[%]		64.08	75.22	74.74	65.86
Attack Success Rate[%] Adversarial Examples of AdvGAN's training data	None	63.85	**74.56**	73.76	**89.48**
	FGSM	71.14	58.04	74.42	69.14
	JSMA	**76.36**	66.36	**74.88**	75.14
	C&W Attack	72.38	62.01	64.48	70.32
	PGDA	67.92	57.26	57.54	67.02
	All	74.62	59.96	64.25	70.74

表３より, 同様に学習対象のデータセットが MNIST の場合, 攻撃対象のモデルが敵対的学習をしたモデルであっても, 生成される敵対的サンプルの性能に大きな変化は確認できなかった. しかし, 表４より, 学習対象のデータセットが CIFAR-10 の場合, 攻撃対象のモデルがロバストを向上させたモデルにおいて, Target Neural Network を敵対的学習することによって, 生成される敵対的サンプルの攻撃性能が大きく異なることが確認できた.

6. 結論

本稿では, Target Neural Network のロバスト性が AdvGAN に及ぼす影響について検証した. 具体的には, Target Neural Network の学習データセットにオリジナル画像データと共に, FGSM, JSMA, C&W Attack, PGDA から生成した敵対的サンプルを個別に用いて学習を行い, Target Neural Network のロバスト性を向上させた. その後, ロバスト性が向上した Target Neural Network を用いて, AdvGAN の学習を行った. その結果として, Target Neural Network を敵対的学習することによって, 高解像度のカラー画像を対象としたとき, Target Neural Network の学習データセットに敵対的サンプルを加えることによって, 生成される敵対的サンプルの攻撃性能が低下することが確認できた. しかし, 攻撃対象のモデルが敵対的学習を行い, 敵対的サンプルに対して耐性が上がったロバスト性が向上したモデルにおいて, Target Neural Network の学習データセットに敵対的サンプルを加えることで, 一部の条件下で, 優れた攻撃性能を発揮することが確認できた.

参考文献

[1] I. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples”, in International Conference on Learning Representations, 2015.
[2] C. Xiao, J.-Y. Zhu, W. He, M. Liu, and D. Song, “Generating adversarial examples with adversarial networks”, 2018.
[3] K. Nishiyama, and H. Shimada, “AdvGAN の classifier モデルに堅牢性を向上させたモデルを用いた場合の有効性について”, 情報処理学会第 85 回全国大会, 2023.
[4] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings”, in CoRR, 2015.
[5] W. Brendel, J. Rauber, and M.Bethge, “Decision-based adversarial attacks: Reliable attacks against black-box machine learning models”, in International Conference on Learning Representations, 2018.
[6] N. Carlini and D. A. Wagner, “Towards evaluating the robustness of neural networks”, in CoRR, 2016.
[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards Deep Learning Models Resistant to Adversarial Attacks”, in ICLR, 2018.