Adversarial Example

Cybersecurity Battleground in the ML Era

FFRI Security, Inc.
株式会社ＦＦＲＩセキュリティ
リサーチ・エンジニア 茂木裕貴

イントロダクション

機械学習は近年発展が著しく、自動運転 [1] や株価予測 [2]、リコメンデーション [3] など、幅広く応用がされている。

その一方で、機械学習技術に対する脅威も取り沙汰されるようになった [4,5]。 [6] において、微小なノイズのような摂動を画像に載せることで、機械学習モデルが誤分類を起こすことが発見された。この摂動の載せられたデータを Adversarial Example(s)[*1] と呼ぶ。微小な摂動に限らず、ある「パッチ」を置くことで誤分類させる方法も [7] で提案された。これは Adversarial Patch と呼ばれる。

こうした現象は現実的に様々な場面で脅威となる。例えば McAfee 社が、道路標識にテープを貼ることで Tesla 社の自動運転車を騙すことに成功した [8]。

本文書では、Adversarial Example（Adversarial Patch などの variant も含む）について、機械学習の各分野ごと・攻撃者の必要な情報ごとに説明する。また、防御手法については、複数の防御手法が破られてきた経緯に触れ、その上で有効とされている防御手法について述べる。

[*1] 複数形の Adversarial Examples とする場合もあるが、本文書では単数形で統一する。

Table of Contents

1. Adversarial Example とは何か

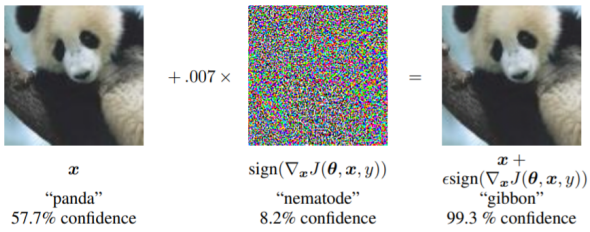


図 **1-1**: パンダの画像に微小な摂動を載せることで機械学習モデルにテナガザルと誤分類させる例

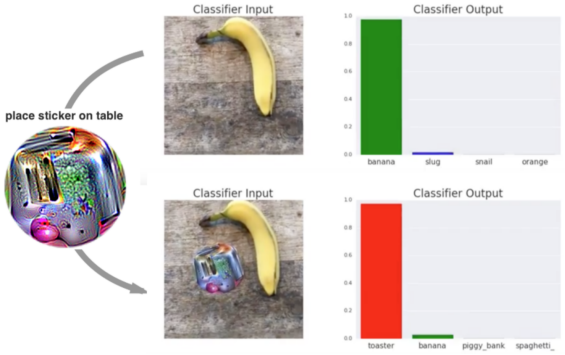


図 **1-2**: バナナの横にパッチを置くことで機械学習モデルにトースターだと誤分類させる例。画像は CleverHans [10] のリポジトリ [11] より

本章では Adversarial Example とその variant の定義について簡単に説明する。

1.1 Adversarial Example の定義

[6] において、微小なノイズのような摂動を画像に載せることで、機械学習モデルが誤分類を起こすことが指摘された*1。この摂動の載せられたデータを Adversarial Example と呼ぶ。 [6] の著者らが次に発表した [9] に載せられた図 1-1 は、Adversarial Example の説明の際に引用されることが多い*2。

微小な摂動に限らず、ある「パッチ」を置くことで誤分類させる方法も [7] で提案された（図 1-2）。これは Adversarial Patch と呼ばれる。

理論的には、Adversarial Example は元の画像に加えると機械学習モデルに分類されるクラスが変わるような摂動の大きさ*3の最小化問題として [6] で定式化されている*4。

このとき、攻撃者が指定したクラスに分類されるように作られたものを Targeted Adversarial Example と呼び、元のクラスと異なるクラスに分類されることのみを目的として作られたものを Untargeted Adversarial Example と呼ぶ [12]。

1.2 Adversarial Example の性質

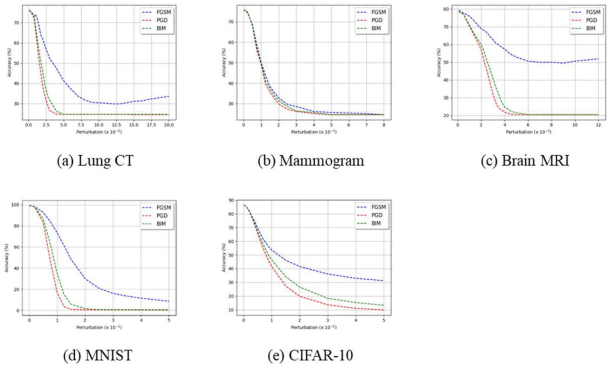
Adversarial Example は、機械学習モデルの推論時における攻撃である [13]。ただし実際に Adversarial Example を作成するにあたっては、対象の機械学習システム・モデルの調査といった事前の作業が必要となりうる [14]。

Adversarial Example は、ある攻撃対象の機械学習モデルに対し、ある入力データ（画像やテキスト）に摂動を加え、誤分類させるものである。しかし、[15] ではある一つのデータだけでなく、複数のデータに対して有効な摂動が作成された。これを Universal Adversarial Perturbation という*5。この摂動が載ったデータを Universal Adversarial Example と呼ぶ [16]。 [15] では、Universal Adversarial Perturbation は、異なる複数のデータに対し “Universal” なだけではなく、異なる複数のアーキテクチャの機械学習モデルにおいても有効であることが示されている。この意味で “Doubly-universal” であるとされている。

Adversarial Example は、図 1-1 のように摂動を加える前後で違いが人間の目には判別が困難なケースもあるが、判別できるものもある。一般には摂動を大きくすることで、人間にとっては摂動が知覚しやすくなる*6が、分類器への攻撃成功率は上がる（図 1-3）。

1.3 リアルワールドにおける Adversarial Example

Adversarial Example はリアルワールドの様々な場面で脅威となっている。例えば McAfee 社が、道路標識にテープを貼ることで Tesla 社の自動運転車を騙すことに成功し [18]、メディアでも大きく取り上げられた [8]。また、自然言語処理でも Adversarial Example は問題と



(a) Lung CT (b) Mammogram (c) Brain MRI

(d) MNIST (e) CIFAR-10

図 1-3: 様々な分類タスクと攻撃手法で、摂動を大きくすると機械学習モデルの分類精度が落ちる。画像は [17] より

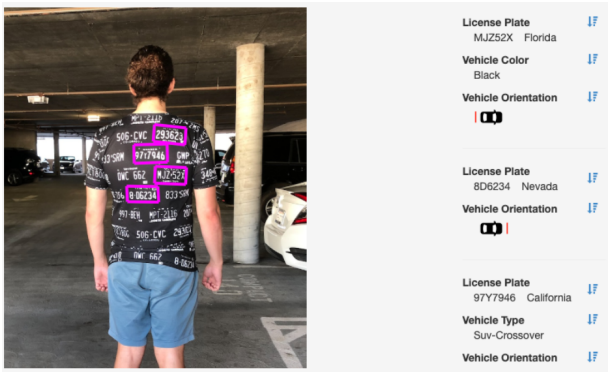


図 1-4: Adversarial Fashion により、人間ではなく車であるとカメラに認識されている。画像は [21] のスライドより

なっており、研究者がスパムフィルターのスパム検知のバイパスに成功し、CVE[*7] が発行された [19]。

こうした状況を踏まえ、CERT/CC[*8] は勾配降下法を用いて訓練された機械学習モデルについて、Adversarial Example を「アルゴリズムの脆弱性」として公表した [20]。

機械学習モデルへの攻撃である Adversarial Example を用いて、「防御」に利用した事例も存在する。Adversarial Fashion [22] は、監視カメラを騙し、監視から逃れられるようデザインされた服である（図 1-4）。また [23] では、顔画像から Adversarial Example を作成し、この画像を元に DeepFake を作成することが困難になる手法が提案された。 [24] で報じられた事例のように合意のない DeepFake は被害者の人権への脅威であり、こうした機械学習による脅威への対策として Adversarial Example が使われうる。

[*1] ただし、Adversarial Example という言葉は使われてはいないものの、 [25] において現在の Adversarial Example に相当する攻撃が提案されている。さらに 2005 年には Spam Filter の Evasion が論じられており [26]、本論文以前から機械学習モデルの誤分類を引き起こす攻撃は研究されてきた。

[*2] 例えば TensorFlow のチュートリアル [27] や PyTorch のチュートリアル [28] でも引用されている。

[*3] [6] では l_2 norm が考察されていたが、他の l_p norm や、SSIM などが使用されることもある [29]。

[*4] ただし、この最小化問題を直接解くことは困難であるため、実際の攻撃では近似した最小化問題を解くことを考察している。

[*5] Adversarial Example を作成する際にデータにのせる摂動を Adversarial Perturbation と呼ぶ。

[*6] 2 章で後述するようにいかに人間に知覚できないように大きな摂動を作るか、という研究も存在する [30]。

[*7] 脆弱性を識別するために広く使われている識別子。詳しくは情報処理推進機構の解説 [31] を参照のこと。

[*8] カーネギーメロン大学のソフトウェア工学研究所のもとに作られた、世界初の CSIRT（サイバーセキュリティインシデントに対応するチーム）である [32]。

2. Adversarial Example の種類

2.1 各分野における Adversarial Example

表 2-1: 各分野における Adversarial Example。サーベイ論文 [33–36] を参考に作成。後述するように Whitebox Attack であっても Transferability により Blackbox で攻撃が成功することがあり、それをもって Blackbox であるとしている論文には Required に “Whitebox, Blackbox (transferability)” と記載した。ただし特定のモデルへの Whitebox Attack について単に Transferability のチェックをするのではなく、Transferability を利用しつつ汎用的に有効な Blackbox Attack を提案しているものは “Blackbox (transferability)” と記載した

分野	タスク	Required	論文
Computer Vision	Image Classification	Whitebox	[9, 37]
		Blackbox (hard label)	[38]
	Face Recognition	Whitebox	[39]
		Whitebox, Blackbox (transferability)	[40]
	Object Detection	Whitebox, Blackbox (transferability)	[41]
		Blackbox (position, hard label)	[42]
	Video Classification	Whitebox	[43]
		Blackbox (score)	[44]
Natural Language Proceeding	Text Classification	Whitebox	[45]
		Blackbox (score)	[46]
	Machine Translation	Whitebox, Blackbox (translated sentence)	[47]
	Sentiment Analysis	Whitebox, Blackbox (score)	[48]
Audio	Speech-to-text	Whitebox	[49]
		Blackbox (generated text)	[50]
	Speaker Recognition	Whitebox	[51]
		Blackbox (score, hard label)	[52]
Reinforcement Learning	Game	Whitebox, Blackbox (transferability)	[53]
		Blackbox (actions)	[54]
	Autonomous Vehicles	Blackbox (position, velocity, actions)	[55]
	Energy Management System	Whitebox, Blackbox (transferability)	[56]
Graph	Node Classification	Whitebox	[57]
		Blackbox (score)	[58]
	Link Prediction	Whitebox, Blackbox (transferability)	[59]
		Blackbox (predicted graph)	[60]
	Community Detection	Blackbox (transferability)	[61]

前ページから続く			
分野	タスク	Required	論文
Time Series	Electrocardiogram Classification	Whitebox	[62]
		Blackbox (hard label)	[63]
	Regression	Whitebox, Blackbox (transferability)	[64]
Cyber Security	Malware Detection	Whitebox	[65]
		Blackbox (hard label)	[66]
	Network Intrusion Detection	Whitebox	[67]
		Blackbox (hard label)	[68]
Tabular Data	Classification	Whitebox	[69]
		Blackbox (score, hard label)	[70]

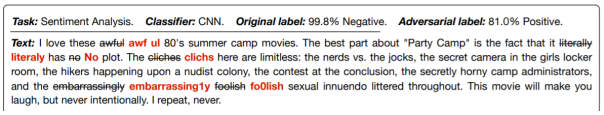
図 **2-1**: [71] の手法により生成された Adversarial Example

Computer Vision（以下、CV）の分野における Adversarial Example は、画像や映像といった扱う対象に対し、微小なノイズのような摂動が加えられることで、機械学習モデルに誤分類を起こすものである。この摂動は、図 1-1 のような画像全体にかけるもの以外にも存在する。 [71] では、画像全体ではなく、一部のピクセルのみを変動させるだけで、誤分類させることに成功している（図 2-1）。このような攻撃は Sparse Adversarial Example と呼ばれる [72]。その一方で、対象のピクセルを変動させる量は大きく、変動したピクセルは知覚しやすい。これを改善し、人間に知覚しづらいような Sparse Adversarial Example を作成する研究も存在する [73]。

この他にも、前述した「パッチ」を置く Adversarial Patch や、画像の外枠に摂動を加える Adversarial Frame [74]（図 2-2）などの variant が存在する。

Natural Language Processing（以下、NLP）の分野における Adversarial Example の生成では、CV におけるそれとは異なった困難さが生じる。これは機械学習モデルに対する入力が、画像の場合連続である[*9]のに対し、テキストの場合は離散になるためである [34]。例えば “FFRI” という単語の文字をそれぞれ 0.5 後ろにずらす、ということは不可能である。また、それぞれ 1 後ろにずらすと “GGSJ” となり、人間にとって大きく意味が異なってしまう。このため、この操作により機械学習モデルが誤分類をしても、有効な Adversarial Example とは言い難い。このことに留意し、人間に取って不自然にならないよう、“FFR1” とするといった文字や単語の挿入

図 **2-2**: [74] の手法により生成された Adversarial Frame。コリー犬がカヌーへと誤分類されている


Task: Sentiment Analysis. ***Classifier:*** CNN. ***Original label:*** 99.8% Negative. ***Adversarial label:*** 81.0% Positive.

Text: I love these ~~awful~~ awf ul 80's summer camp movies. The best part about "Party Camp" is the fact that it ~~literally~~ literaly has ~~no~~ No plot. The ~~cliches~~ clichs here are limitless: the nerds vs. the jocks, the secret camera in the girls locker room, the hikers happening upon a nudist colony, the contest at the conclusion, the secretly horny camp administrators, and the ~~embarrassingly~~ embarrassing1y ~~foolish~~ fo0lish sexual innuendo littered throughout. This movie will make you laugh, but never intentionally. I repeat, never.

図 **2-3**: [34] で生成された Adversarial Example の例。単語の一部を大文字にしたり、空白を入れることで、人間には元の文と同様に読める一方で機械学習モデルには誤分類させることに成功している

や削除・置き換えが使用されている（図 2-3 を参照）。

Audio の分野においても、Adversarial Example は存在する。Audio は時系列の波のデータであり、微小な波を足し合わせることで Adversarial Example を生成できる。 [49] では、人間にとって聞こえる文章と、機械学習モデルが transcribe する文章を全く異なるものにしたり、音楽のフレーズに機械学習モデルが transcribe してしまう文章を埋め込むことに成功した[*10]。また心電図などの時系列の波データについても Audio と同様に Adversarial Example が生成できる [62]。

Reinforcement Learning（以下、RL）における Adversarial Example は、2 つのアプローチに大別される。1 つは攻撃対象の Agent の Observation に直接改変を加える

ものである。 [53] では、ゲーム画面に摂動を加え、ゲームをプレイする Agent への攻撃を行っている。もう 1 つは、攻撃対象の Agent とは別の Agent を操作するといった手法により、間接的に Agent の Observation に改変を加えるものである。 [55] では、標的の自動運転車に対し、標的に事故の責任があるようなクラッシュを起こせるよう、別の攻撃用 Agent を Deep Q-Learning [75] で訓練している。

Graph データを扱う Neural Network として、Graph Neural Network (以下、GNN) が存在する*11。Neural Network を Graph データに適用する試みは 1997 年の [76] からあり、[77] で GNN が提唱された。GNN は、Node 間に Link が存在するかどうか（Edge が存在するかどうか）を予測する Link Prediction や、Graph から Community に属する Subgraph を抽出する Community Detection などに応用されている。こうした Graph データを扱う機械学習モデルに対し誤分類をさせるため、入力する Graph に対し Edge を追加・削除したり、Node を追加するといった手法を用いて Adversarial Example が作成されている [35]。

次に Cyber Security の分野における Adversarial Example について解説する。スパムフィルターやマルウェア検知器への Adversarial Example を作成することは、検知器のバイパスを試みることである。しかし、Cyber Security の分野では Adversarial Example が提唱される前から検知器のバイパスは活発に行われてきた。例えば、2005 年の [26] ではスパムでないメールに含まれる単語を挿入・追加することによるスパムフィルターのバイパスが論じられている。同様にマルウェア検知器のバイパスも昔から行われてきた*12。深層学習が応用されたマルウェア検知器への攻撃として、 [65] ではマルウェアのバイナリの一部を改変することで 60 個中 52 個をバイパスさせることに成功した。マルウェアを改変するにあたっては、改変後のマルウェアが動作することが必要であるという制限が存在する。これを実現するため、例えば [65] では、インポート関数の追加や、ファイルの最後にジャンクデータを追加するなど、動作に影響を及ぼさないような手法が採用されている。

最後に、Tabular Data における Adversarial Example について述べる。Tabular Data において特徴的な点は、Tabular の要素には数値やカテゴリカルデータといった複数のデータ型が存在することである。この対処として、 [69] では離散値を連続値として扱い、順序のないカテゴリカルデータは削除している。 [70] では、生成した Adversarial Example について、整数値のフィールドの値は丸める、ブール値のフィールドの値はしきい値との大小関係により 0 か 1 に振り分けるといった工夫を行っている。

このように、Adversarial Example は [6] で考察された CV の領域を超え、様々な分野で研究が進んでいる。

2.2 攻撃者の得られる情報の設定による Adversarial Example の分類

攻撃者が得られる情報の設定により、Adversarial Example の生成手法は Whitebox Attack と Blackbox Attack に大別することができる [78]。Whitebox Attack は、対象の機械学習モデルに対する完全な知識及び直接的なアクセスを必要とする。モデルのアーキテクチャや訓練済みモデルの各種パラメータ、訓練に使用したデータの分布など、全てを攻撃者が知っており、また攻撃対象のモデルにアクセスができる設定である。例えば Whitebox Attack の一つである Fast Gradient Sign Method [9]（以下、FGSM）において摂動を計算するには、訓練済みモデルのパラメータや、それを用いた損失関数の勾配の計算が必要であり、Whitebox の設定でなければこれらの値を得ることは困難である。実際に特定の機械学習モデルを攻撃する場合には、Whitebox であることは非現実的であるが、後述するように Model Extraction*13 と組み合わせることで Blackbox Attack が実現できたり、3 章で後述するように Adversarial Training に使うことでモデルの Adversarial Example への耐性を高めることができるなど、Whitebox Attack 自体は有用である。

Blackbox Attack は、攻撃者の得られる情報が制限された攻撃である。ただし、この定義は文献によって異なる。例えば [79] では「モデルのアーキテクチャに対し知識はなく、モデルの入出力のみを知っている」設定とされているが、 [80] においては提案する攻撃を「Transformer Architectures に対する Blackbox Attack」であると主張しているように、必ずしも攻撃対象のモデルに対し知識を全く仮定しない訳ではない。このように、Whitebox Attack ほどではないものの攻撃対象のモデルに対し何らかの知識を仮定する場合、Graybox (Greybox) Attack

と呼ばれることがある[*14]。

ここで機械学習モデルの「出力」として、(1 つ、もしくは複数の) ラベルに対する機械学習モデルのスコアを必要とするものを Score-based Attack と呼び、予測されるラベルのみを必要とするものを Decision-based Attack と呼ぶ [36]。機械学習モデルの出力するスコアを Soft-label と呼び、出力するラベルを Hard-label と呼ぶことがあるため [81]、Decision-based Attack は Hard-label Attack と呼ばれることもある [82]。Blackbox Attack の手法は、以下の 2 つに大別される。対象のモデルへクエリし、データを徐々に AE になるよう改善していく Query-based Attack と、攻撃対象のモデルの代替モデルに対し Whitebox Attack を行う Transfer-based Attack である [83]。Query-based Attack は、 [84] や [74] が該当する。 [85] では、ImageNet [86] 上で訓練されたモデルに対し、900 弱程度の数のクエリにより、100% の精度で Adversarial Attack を成功させた。

Transfer-based Attack は、Model Extraction などで対象の機械学習モデルを元に代替モデルを作成するものと、そうでないものに分かれる。Adversarial Example には、ある機械学習モデルに対し有効なものが、他の機械学習モデルに対しても有効であることがある、Transferability という性質が存在する。例えば [87] では、Support Vector Machine (以下、SVM) や決定木、Deep Neural Network など異なるアルゴリズムをまたいだ Adversarial Example の Transferability が検討された。ここでは、同じアルゴリズム同士での Transferability が最も高いものの、Logistic Regression のモデルに対し作成された Adversarial Example の 90% 以上が SVM に対し有効であるといったことが示された。一方、Transferability をより効率的に応用するため、攻撃対象の機械学習モデルを元に代替モデルを作成する手法としては、 [88] がある。 [88] では、Amazon、Google、Microsoft 及び Clarifai のクラウド画像分類サービスに対し代替モデルを作成し、PGD Attack [89] を変形した Whitebox Attack を代替モデルに適用することで、各サービスに対し有効な Adversarial Example を作成した。このような工夫を行うことなく、単に攻撃対象の機械学習モデルと全く異なるモデルに Whitebox Attack を行っても、前述の Transferability により対象の機械学習モデルに対しても攻撃が成功することはある。そのため、Whitebox Attack の手法の論文であっても、この Transferability によって、Blackbox の設定で実験を行うことがある[*15]。そのような場合と直接的な Blackbox Attack (Query-based Attack 及び Model Extraction を行う Transferability Attack) を行う場合を区別するため、表 2.1 では前者を “Blackbox (transferability)” と記載した。

また、[84] において、Nobox Attack という設定が提案された。Nobox Attack では、攻撃者は攻撃対象となる機械学習モデルについての知識がなく、さらに Adversarial Example を作成するために攻撃対象のモデルにデータをクエリすることもできない。 [84] では Nobox Attack の具体的な手法は提案されなかったが、後に [90] において実現された[*16]。

[*9] ピクセル値自体は離散であるが、前処理により連続値として扱われる。例えば TensorFlow のチュートリアル [27] では、“float32” 型にキャストされた後、標準化が行われる。

[*10] 論文の著者の Web サイト [91] で実際の例を聴くことができる。

[*11] Graph Neural Network の歴史やタスクについての説明は [92] による。

[*12] 例えば 2004 年の論文 [93]。

[*13] 対象の機械学習モデルと同等の機能・性能を持つモデルを複製する攻撃で、例えば [94] では API 経由のアクセスにより Amazon のようなクラウドサービスの機械学習モデルを複製することに成功した。

[*14] 例えば [95] や [96] はそれぞれの論文では Blackbox であるとしているが、サーベイ論文 [36] では Graybox であると位置づけられている。

[*15] 例えば [64] など。

[*16] ここでも、Whitebox Attack の Transferability を用いた場合、Nobox Attack が成立することはある。

3. Adversarial Example の防御

3.1 Adversarial Example の防御手法とそのバイパス

Adversarial Example の脅威に対処するため、防御手法も研究が進んだ。防御手法は主に二種類に大別される。Adversarial Example を検知する手法と、機械学習モデルそのものを Adversarial Example に対して Robust にする（耐性を高める）手法である。

Adversarial Example を検知する手法としては、機械学習モデルに新たに Adversarial Example 用の分類クラスを追加する [97] や、Adversarial Example であるか否かを分類する分類器を作成する [98] などが提案された。機械学習モデルを Robust にする手法としては、Distillation*17 を用いて Robust な機械学習モデルを作成する Defensive Distillation [99] などが提案された。

しかし、こうした防御手法は完全ではなかった。

Defensive Distillation [99] が C&W Attack [100] により破られたのを皮切りに、様々な Adversarial Example の防御手法が破られた。 [12] では前述した [97] や [98] を含む Adversarial Example の検知手法が 10 個バイパスされ、 [101] では ICLR 2018 にて提案された防御手法のうち 6 つが完全にバイパスされた。

ここでは [97] のバイパス手法について簡単に例を用いて説明する。猫の画像と犬の画像の 2 値分類を行う分類器を考える。 [97] の手法では、まず通常通りこの 2 値分類を行う分類器を作成する。次に、この分類器に対し Adversarial Example を複数作成する。そして、今度は猫の画像か、犬の画像か、Adversarial Example かの 3 値分類を行う分類器を作成する。これによって Adversarial Example を検知しよう、というものであった。これに対し、[12] では、次のようにバイパスを行った。まず、ある犬の画像を猫の画像に誤分類させたいとする。そこで、攻撃対象となる上記の 3 値分類器を、次のような 2 値分類器とみなす。「猫の画像」か、「犬の画像または Adversarial Example」かを判定する分類器である。この 2 値分類器に対し、猫の画像と判定される Adversarial Example を作成することができ、これにより検知をバイパスすることができる。 [98] では Adversarial Example かどうかを判定する分類器を作成して検知する手法が提案されたが、これも同様にバイパスできる*18。

その後も、NeurIPS 2018 で Spotlight Paper*19 となった AMI [102] が [103] で破られ、 [104] では ICLR や ICML、NeurIPS といった機械学習の国際トップ学会で採択された 13 の防御手法に対し攻撃を成功させた。

3.2 Two Promising Direction

多くの防御手法が破られる中で、機械学習モデルを Robust にできるとされている手法が 2 つある*20。Adversarial Training と Certified Defense*21 である。Adversarial Training は、Adversarial Example を訓練データに組み込む手法である [9]。Certified Defense は、理論的に Robustness を保証する手法の総称である。典型的には、ある大きさの摂動に対しどの程度の Robustness が保証できるかを考える [105]。またこの 2 つの手法を組み合わせた手法も提案されている [106]。

ただし、これらの防御手法にも弱点が存在する。まず、Adversarial Training は機械学習モデルの訓練時間の増大を引き起こす。FGSM のような勾配を一度だけ計算する “Single Step” で高速に生成できる Adversarial Example を訓練データとして使用するだけでは、十分に Robust にならなかったり、Adversarial Example に Overfit してしまい Adversarial Example でないデータの分類精度が大きく低下する “Catastrophic Overfitting” を引き起こしうることが知られている [107]。そこで PGD Attack のような勾配の計算を数イテレーションに渡って行い少しずつ摂動を足していく “Multi Step” の攻撃を使うことが推奨されている [89] が、これは訓練時間を増大させる [107]。そのため、 [107] や [108] など、“Single Step” の Adversarial Example を用いても Robust になるよう工夫を行い Adversarial Training を行う研究も盛んである。

次に摂動の大きさに制限を設ける Certified Defense は、摂動が大きければ Robust な保証ができない。もちろん摂動が大きければ Adversarial Example は人間の目で判別しやすくなるはずである。しかし、 [30] では、大きな摂動を加えつつも、人間の目にも自然に見える

Adversarial Example を作成することに成功した。

また、これらの手法により機械学習モデルが Robust になるといっても、他の提案手法より Robustness の評価の数値が高い訳ではない。例えば Adversarial Training のサーベイ論文 [109] の Table 1 からは、一部を除いて 40%～50% 台の Accuracy が多い事がわかる[*22]。また Certified Defense についても、[105] の Table IV から同様の状況であることがわかる。その一方で、前述の通り破られた Defensive Distillation [99] では、ほんの数パーセントしか Adversarial Example による攻撃が成功していなかった。つまり、バイパスしうるような防御手法であっても、実際にバイパスされるまでは、極めて有効に見える。逆に言えば、実際にその間は有効な防御であるとみなして、Adversarial Training や Certified Defense に限定せず、最新の手法を用いる方針も考えうる。

[*17] Distillation は [110] により提案された、機械学習モデルの軽量化手法である。ただし Defensive Distillation [99] においては軽量化は行わない。

[*18] 詳細は [12] を参照のこと。

[*19] Spotlight Paper は submit された論文のうち上位 3% [103] で、5 分間の発表を行う [111]。

[*20] 例えば [112] ではこの 2 つが “Promising Direction” とされ、[113] では防御手法を Certified なものとそうでない Heuristic なものに分け、Adversarial Training は Heuristic な防御手法の中で最も成功しているとしている。

[*21] Provable Defense と呼ばれることもある [114]。これは Certified Defense を提案する論文 [115] の第一版が出たすぐ後に独立に Provable Defense の論文 [114] が出たという時系列によると思われる。実際、[114] の第一版 [116] では、[115] はまだ ICLR 2018 のレビュー中となっており、リファレンスに載っているだけである。現在はそれぞれ相互に論文中で言及がなされている。

[*22] この数値はデータセットや攻撃の種類、さらに摂動の大きさに依存するため、ここに記載されている数値の大小のみで手法の比較はできないことに注意。

4. まとめ

Adversarial Example とその variant の定義を述べ、各分野における White Box/Black Box Attack について紹介した。Adversarial Example は CV だけでなく、NLP など、他の分野にも存在する。このように、どのような分野・用途であっても、機械学習を使う際は、Adversarial Example が存在しうることを想定する必要がある。また防御手法については、コンセンサスの得られた完全な防御が存在しないことを述べた。さらに現実的には、機械学習モデル単体だけではなく、機械学習システム全体として防御策を考える必要もある。防御しなければならない攻撃は Adversarial Example だけではなく、またどれほど攻撃に対し Robust なモデルができても、システムそのものが侵害されれば無力化されうる。そうしたシステム全体としての防御のベストプラクティスは固まりきっておらず*23、今後の研究や議論を注視していくことや、各機械学習システムに応じて脅威分析や脆弱性検査を行うことも必要ではないだろうか。

*23 将来的な標準・及びベストプラクティスの策定のため、NISTIR 8269 のドラフト版 [117] が公開されている。

参考文献

[1] Marin Toromanoff, Emilie Wirbel, and Fabien Moutarde. End-to-end model-free reinforcement learning for urban driving using implicit affordances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[2] Kentaro Imajo, Kentaro Minami, Katsuya Ito, and Kei Nakagawa. Deep portfolio optimization via distributional prediction of residual factors, 2020.

[3] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. Deep learning recommendation model for personalization and recommendation systems, 2019.

[4] 機械学習アルゴリズムに脆弱性（JVN）. `https://scan.netsecurity.ne.jp/article/2020/03/27/43869.html`. (Accessed on 02/17/2021).

[5] 韓国の人気チャットボット、憎悪表現でサービス停止. `https://www.afpbb.com/articles/-/3326565`. (Accessed on 02/17/2021).

[6] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[7] Tom Brown, Dandelion Mane, Aurko Roy, Martin Abadi, and Justin Gilmer. Adversarial patch. 2017.

[8] マカフィー、テスラ車をダマしてスピード違反させることに成功. `https://www.gizmodo.jp/2020/03/mcafee-tesla.html`. (Accessed on 02/17/2021).

[9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[10] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, Rujun Long, and Patrick McDaniel. Technical report on the cleverhans v2.1.0 adversarial examples library, 2018.

[11] Adversarial Patch. `https://github.com/cleverhans-lab/cleverhans/blob/master/examples/adversarial_patch/AdversarialPatch.ipynb`. (Accessed on 02/17/2021).

[12] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, p. 3–14, New York,

NY, USA, 2017. Association for Computing Machinery.

[13] X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, and A. V. Vasilakos. Privacy and security issues in deep learning: A survey. *IEEE Access*, Vol. 9, pp. 4566–4593, 2021.

[14] Adversarial ML Threat Matrix. https://github.com/mitre/advmlthreatmatrix. (Accessed on 02/25/2021).

[15] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 86–94, 2017.

[16] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 30, No. 9, pp. 2805–2824, 2019.

[17] Marina Z. Joel, Sachin Umrao, Enoch Chang, Rachel Choi, Daniel Yang, Antonio Omuro, Roy Herbst, Harlan Krumholz, and Sanjay Aneja. Adversarial attack vulnerability of deep learning models for oncologic images. *medRxiv*, 2021.

[18] モデルハッキング：自動運転車の安全な走行のための ADAS を調査. https://blogs.mcafee.jp/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles. (Accessed on 03/01/2021).

[19] CVE-2019-20634 Detail. https://nvd.nist.gov/vuln/detail/CVE-2019-20634. (Accessed on 03/01/2021).

[20] VU#425163 - Machine learning classifiers trained via gradient descent are vulnerable to arbitrary misclassification attack. https://www.kb.cert.org/vuls/id/425163/. (Accessed on 03/23/2021).

[21] DEFCON 27 Crypto & Privacy Presentation. https://adversarialfashion.com/pages/defcon-27-crypto-privacy-presentation. (Accessed on 03/01/2021).

[22] Adversarial Fashion. https://adversarialfashion.com/. (Accessed on 03/01/2021).

[23] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, Vol. 12538 of *Lecture Notes in Computer Science*, pp. 236–251. Springer, 2020.

[24] ディープフェイク・ポルノ、やっと動き出した規制への道. https://www.technologyreview.jp/s/234910/deepfake-porn-is-ruining-womens-lives-now-the-law-may-finally-ban-it/. (Accessed on 03/22/2021).

[25] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. *Lecture Notes in Computer Science*, p. 387–402, 2013.

[26] Daniel Lowd and Christopher Meek. Good word attacks on statistical spam filters. In *CEAS*, Vol. 2005, 2005.

[27] Adversarial example using FGSM : TensorFlow Core. https://www.tensorflow.org/tutorials/generative/adversarial_fgsm. (Accessed on 02/17/2021).

[28] Adversarial Example Generation - PyTorch Tutorials 1.7.1 documentation. https://pytorch.org/tutorials/beginner/fgsm_tutorial.html. (Accessed on 02/17/2021).

[29] M. Sharif, L. Baue, and M. K. Reite. On the suitability of lp-norms for creating and preventing adversarial examples. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1686–16868, 2018.

[30] Amin Ghiasi, Ali Shafahi, and Tom Goldstein. Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates. In *International Conference on Learning Representations*, 2020.

[31] 共通脆弱性識別子 CVE 概説. https://www.ipa.go.jp/security/vuln/CVE.html, July 2015. (Accessed on

05/11/2021).

[32] 米国 CERT Coordination Center (CERT/CC) と「CERT」について. https://www.ipa.go.jp/security/vuln/CVE.html, October 2004. (Accessed on 05/11/2021).

[33] Fatemeh Vakhshiteh, Ahmad Nickabadi, and Raghavendra Ramachandra. Adversarial attacks against face recognition: A comprehensive study, 2021.

[34] Wenqi Wang, Lina Wang, Run Wang, Zhibo Wang, and Aoshuang Ye. Towards a robust deep neural network in texts: A survey, 2020.

[35] Lichao Sun, Yingtong Dou, Carl Yang, Ji Wang, Philip S. Yu, Lifang He, and Bo Li. Adversarial attack and defense on graph data: A survey, 2020.

[36] Ihai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain, 2021.

[37] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2582, 2016.

[38] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[39] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face ID system. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pp. 819–826. IEEE, 2020.

[40] D. Deb, J. Zhang, and A. K. Jain. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10, 2020.

[41] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. *Lecture Notes in Computer Science*, p. 52–68, 2019.

[42] Yajie Wang, Yu an Tan, Wenjiao Zhang, Yuhang Zhao, and Xiaohui Kuang. An adversarial attack on dnn-based black-box object detectors. *Journal of Network and Computer Applications*, Vol. 161, p. 102634, 2020.

[43] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V. Krishnamurthy, Amit K. Roy Chowdhury, and Ananthram Swami. Stealthy adversarial perturbations against real-time video classification systems. *Proceedings 2019 Network and Distributed System Security Symposium*, 2019.

[44] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. Black-box adversarial attacks on video recognition models. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, p. 864–872, New York, NY, USA, 2019. Association for Computing Machinery.

[45] Nicolas Papernot, Patrick D. McDaniel, Ananthram Swami, and Richard E. Harang. Crafting adversarial input sequences for recurrent neural networks. In Jerry Brand, Matthew C. Valenti, Akinwale Akinpelu, Bharat T. Doshi, and Bonnie L. Gorsic, editors, *2016 IEEE Military Communications Conference, MILCOM 2016, Baltimore, MD, USA, November 1-3, 2016*, pp. 49–54. IEEE, 2016.

[46] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2890–2896, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[47] Javid Ebrahimi, Daniel Lowd, and Dejing Dou. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 653–663, Santa

Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[48] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. *Proceedings 2019 Network and Distributed System Security Symposium*, 2019.

[49] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7. IEEE, 2018.

[50] Shreya Khare, Rahul Aralikatte, and Senthil Mani. Adversarial Black-Box Attacks on Automatic Speech Recognition Systems Using Multi-Objective Evolutionary Optimization. In *Proc. Interspeech 2019*, pp. 3208–3212, 2019.

[51] Qing Wang, Pengcheng Guo, and Lei Xie. Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition. In *Proc. Interspeech 2020*, pp. 4228–4232, 2020.

[52] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. Who is real bob? adversarial attacks on speaker recognition systems. In *Proceedings of the 42nd IEEE Symposium on Security and Privacy (IEEE S&P, Oakland)*, 2021.

[53] Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[54] Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[55] S. Zhang, H. Peng, S. Nageshrao, and H. E. Tseng. Generating socially acceptable perturbations for efficient evaluation of autonomous vehicles. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1341–1347, 2020.

[56] Pengyue Wang, Yan Li, Shashi Shekhar, and William F. Northrop. Adversarial attacks on reinforcement learning based energy management systems of extended range electric delivery vehicles, 2020.

[57] Xiaoyun Wang, Minhao Cheng, Joe Eaton, Cho-Jui Hsieh, and Felix Wu. Attack graph convolutional networks by adding fake nodes, 2020.

[58] Jiaqi Ma, Shuangrui Ding, and Qiaozhu Mei. Towards more practical adversarial attacks on graph neural networks. In *Advances in Neural Information Processing Systems*, 2020.

[59] Wanyu Lin, Shengxiang Ji, and Baochun Li. Adversarial attacks on link prediction algorithms based on graph neural networks. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, ASIA CCS '20, p. 370–380, New York, NY, USA, 2020. Association for Computing Machinery.

[60] Houxiang Fan, Binghui Wang, Pan Zhou, Ang Li, Meng Pang, Zichuan Xu, Cai Fu, Hai Li, and Yiran Chen. Reinforcement learning-based black-box evasion attacks to link prediction in dynamic graphs, 2020.

[61] Jia Li, Honglei Zhang, Zhichao Han, Yu Rong, Hong Cheng, and Junzhou Huang. Adversarial attack on community detection by hiding individuals. *Proceedings of The Web Conference 2020*, Apr 2020.

[62] Xintian Han, Yuxuan Hu, Luca Foschini, Larry Chinitz, Lior Jankelson, and Rajesh Ranganath. Adversarial examples for electrocardiograms, 2019.

[63] Jonathan Lam, Pengrui Quan, Jiamin Xu, Jeya Vikranth Jeyakumar, and Mani Srivastava. Hard-label black-box adversarial attack on deep electrocardiogram classifier. In *Proceedings of the 1st ACM International Workshop on Security and Safety for Intelligent Cyber-Physical Systems*, SecICPS '20, p. 6–12, New York, NY, USA, 2020. Association for Computing Machinery.

[64] Gautam Raj Mode and Khaza Anuarul Hoque. Adversarial examples in deep learning for multivariate time series

regression, 2020.

[65] Luca Demetrio, Battista Biggio, Giovanni Lagorio, Fabio Roli, and Alessandro Armando. Explaining vulnerabilities of deep learning to adversarial malware binaries. In Pierpaolo Degano and Roberto Zunino, editors, *Proceedings of the Third Italian Conference on Cyber Security, Pisa, Italy, February 13-15, 2019*, Vol. 2315 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[66] Raphael Labaca-Castro, Corinna Schmitt, and Gabi Dreo. Aimed: Evolving malware with genetic programming to evade detection. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pp. 240–247. IEEE, 2019.

[67] Z. Wang. Deep learning-based intrusion detection with adversaries. *IEEE Access*, Vol. 6, pp. 38367–38384, 2018.

[68] Zilong Lin, Yong Shi, and Zhi Xue. Idsgan: Generative adversarial networks for attack generation against intrusion detection, 2021.

[69] Vincent Ballet, Xavier Renard, Jonathan Aigrain, Thibault Laugel, Pascal Frossard, and Marcin Detyniecki. Imperceptible adversarial attacks on tabular data, 2019.

[70] Francesco Cartella, Orlando Anunciação, Yuki Funabiki, Daisuke Yamaguchi, Toru Akishita, and Olivier Elshocht. Adversarial attacks for tabular data: Application to fraud detection and imbalanced data. In Huáscar Espinoza, John McDermid, Xiaowei Huang, Mauricio Castillo-Effen, Xin Cynthia Chen, José Hernández-Orallo, Seán Ó hÉigeartaigh, and Richard Mallah, editors, *Proceedings of the Workshop on Artificial Intelligence Safety 2021 (SafeAI 2021) co-located with the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021), Virtual, February 8, 2021*, Vol. 2808 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[71] A. Modas, S. Moosavi-Dezfooli, and P. Frossard. Sparsefool: A few pixels make a big difference. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9079–9088, 2019.

[72] Xiaoyi Dong, Dongdong Chen, Jianmin Bao, Chuan Qin, Lu Yuan, Weiming Zhang, Nenghai Yu, and Dong Chen. Greedyfool: Distortion-aware sparse adversarial attack. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[73] Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. In *ICCV*, 2019.

[74] Francesco Croce, Maksym Andriushchenko, Naman D. Singh, Nicolas Flammarion, and Matthias Hein. Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks, 2020.

[75] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.

[76] A. Sperduti and A. Starita. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, Vol. 8, No. 3, pp. 714–735, 1997.

[77] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Vol. 2, pp. 729–734 vol. 2, 2005.

[78] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey, 2018.

[79] Aminul Huq and Mst. Tasnim Pervin. Adversarial attacks and defense on texts: A survey, 2020.

[80] Adrian de Wynter. Mischief: A simple black-box attack against transformer architectures, 2020.

[81] Christian Thiel. Classification on soft labels is robust against label noise. In Ignac Lovrek, Robert J. Howlett, and Lakhmi C. Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, pp. 65–73, Berlin,

Heidelberg, 2008. Springer Berlin Heidelberg.

[82] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, JinFeng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *International Conference on Learning Representations*, 2019.

[83] Chilin Fu, X. Zhang, and Jun Zhou. Poster: Query-efficient adversarial attack framework by utilizing transfer gradient. 2020.

[84] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, p. 15–26, New York, NY, USA, 2017. Association for Computing Machinery.

[85] Yang Bai, Yuyuan Zeng, Yong Jiang, Yisen Wang, Shu-Tao Xia, and Weiwei Guo. Improving query efficiency of black-box adversarial attack. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pp. 101–116, Cham, 2020. Springer International Publishing.

[86] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, Vol. 115, No. 3, pp. 211–252, 2015.

[87] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, 2016.

[88] Dou Goodman. Transferability of adversarial examples to attack cloud-based image classifier service, 2020.

[89] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[90] Qizhang Li, Yiwen Guo, and Hao Chen. Practical no-box adversarial attacks against dnns. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[91] Audio Adversarial Examples. `https://nicholas.carlini.com/code/audio_adversarial_examples`. (Accessed on 03/22/2021).

[92] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 32, No. 1, p. 4–24, Jan 2021.

[93] Mihai Christodorescu and Somesh Jha. Testing malware detectors. *SIGSOFT Softw. Eng. Notes*, Vol. 29, No. 4, p. 34–44, July 2004.

[94] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *Proceedings of the 25th USENIX Conference on Security Symposium*, SEC'16, p. 601–618, USA, 2016. USENIX Association.

[95] Weiwei Hu and Ying Tan. Generating adversarial malware examples for black-box attacks based on gan, 2017.

[96] Ishai Rosenberg, Asaf Shabtai, Lior Rokach, and Yuval Elovici. Generic black-box end-to-end attack against state of the art api call based malware classifiers. In Michael Bailey, Thorsten Holz, Manolis Stamatogiannakis, and Sotiris Ioannidis, editors, *Research in Attacks, Intrusions, and Defenses*, pp. 490–510, Cham, 2018. Springer International Publishing.

[97] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical)

detection of adversarial examples, 2017.

[98] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and clean data are not twins, 2017.

[99] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597, 2016.

[100] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

[101] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, July 2018.

[102] Guanhong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. In *Advances in Neural Information Processing Systems 31*, pp. 7728–7739, 2018.

[103] Nicholas Carlini. Is ami (attacks meet interpretability) robust to adversarial examples?, 2019.

[104] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1633–1645. Curran Associates, Inc., 2020.

[105] Linyi Li, Xiangyu Qi, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks, 2020.

[106] Mislav Balunovic and Martin Vechev. Adversarial training and provable defenses: Bridging the gap. In *International Conference on Learning Representations*, 2020.

[107] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

[108] B. S. Vivek and R. Venkatesh Babu. Single-step adversarial training with dropout scheduling. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 947–956, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[109] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness, 2021.

[110] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[111] 福地一斗, 福馬智生. 小特集：「AI トレンド・トップカンファレンス NeurIPS 2018，AAAI 2019 報告会」NeurIPS 概要・今年の傾向. 人工知能, Vol. 34, No. 5, pp. 686–688, 2019.

[112] Nicholas Carlini. A (short) Primer on Adversarial Robustness. `https://www.facebook.com/mediaforensics2020/videos/669127567151553/`, June 2020. (Accessed on 04/12/2021).

[113] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, Vol. 6, No. 3, pp. 346–360, 2020.

[114] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 5286–5295. PMLR, 10–15 Jul 2018.

[115] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

[116] Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope, 2017.

[117] Elham Tabassi, Kevin Burns, Michael Hadjimichael, Andres Molina-Markham, and Julian Sexton. A taxonomy

and terminology of adversarial machine learning. *NIST IR*, 2019.

株式会社ＦＦＲＩセキュリティ

株式会社ＦＦＲＩセキュリティは、日本発のサイバーセキュリティをリードする専門家集団です。国際的なセキュリティカンファレンスでの研究発表実績もある世界トップレベルのサイバーセキュリティ専門家集団が、 先進的な調査研究により、今後予想される脅威を先読みし、一歩先行くコンセプトで製品・サービスを展開しています。

株式会社ＦＦＲＩセキュリティ
〒100-0005 東京都千代田区丸の内３丁目３番１号 新東京ビル２階
E-mail: research-feedback[at]ffri.jp URL: https://www.ffri.jp/