社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

秘密鍵を用いた画像分類器の AutoAttack に対する頑健性評価

田中　美貴† 　エイプリルピョンマウンマウン† 　越前　功†† 　貴家　仁志†

† 東京都立大学システムデザイン学部，東京都
†† 国立情報学研究所, 東京都
E-mail: †tanaka-miki@ed.tmu.ac.jp, ††fugokidi@gmail.com, †††iechizen@nii.ac.jp, ††††kiya@tmu.ac.jp

あらまし　深層学習モデルの予測結果を不正に操作する敵対的事例攻撃に対する対策が，急務の課題となっている．本稿では，先に提案した秘密鍵を用いた画像分類器の頑健性を，敵対的事例攻撃のベンチマーク法である AutoAttack を用いて評価する．さらに秘密鍵を用いた画像分類器のための敵対的事例攻撃検出器を提案する。実験において，秘密鍵を用いた画像分類器はブラックボックス攻撃に対して脆弱性があること，提案された検出器を組み合わせて使用することによって，その脆弱性が改善され，最新のベンチマーク結果を上回ることを確認する．
キーワード　敵対的事例, 機械学習, 深層学習, 敵対的事例の検出

Adversarial Robustness of Secret Key-Based Defense against AutoAttack

Miki TANAKA†, Maungmaung APRILPYONE†, Isao ECHIZEN††, and Hitoshi KIYA†

† Kiya's laboratory, Department of Computer Science, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo, 191-0065 Japan
†† National Institute of Informatics (NII), 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, JAPAN
E-mail: †tanaka-miki@ed.tmu.ac.jp, ††fugokidi@gmail.com, †††iechizen@nii.ac.jp, ††††kiya@tmu.ac.jp

Abstract　Deep neural network (DNN) models are well-known to easily misclassify prediction results by using input images with small perturbations, called adversarial examples, so investigating countermeasures for adversarial examples is an urgent issue. In this paper, the secret key-based defense that we proposed is evaluated in terms of robustness against adversarial examples in accordance with a benchmark attack method, called AutoAttack. In addition, we propose a detection method of adversarial examples to be combined with the secret key-based defense. In an experiment, the secret key-based classification model is confirmed that it is not robust enough against a black box attack, and the combined use of the key-based defense and the proposed detector outperforms the latest benchmark.
Key words　Adversarial example, Machine learning, Deep learning, Adversarial detection

1. ま え が き

深層学習 (DNN) モデルは，急速に発展を続けており，その応用分野を広げている．しかし，DNN モデルに対する様々な攻撃が指摘され，その対策が急務の課題となっている [1]．その課題の一つが，敵対的事例と呼ばれる人為的に作成された微小なノイズを入力データに加えることによって，予測結果が操作される脆弱性を持つことである．ノイズは人間が知覚できないほど僅かであっても攻撃が可能であり，特に顔認証やマルウェアの検出，自動運転などの高い信頼性を要求される応用分野にとって大きな脅威となっている．

敵対的事例に対する防御策は，主に 2 つのアプローチが検討されてきた．第一のアプローチは，敵対的事例に対して頑健なモデルを作成することである．敵対的事例によってデータを拡張することで，モデルが頑健性を獲得することが指摘されている [2] が，一方で，攻撃を仮定する必要があることや，クリーンな画像の分類結果が低下する課題がある．攻撃を仮定しない方法として，秘密鍵を用いた画像変換によるロバストなモデル [3] が提案されている．しかし，秘密鍵を用いたモデルは，鍵を知られてしまった場合やブラックボックス攻撃に対してはロバスト性が低下する．この課題に対処するために，アンサンブルが提案されたが，異なる鍵やブロックサイズなどの暗号化条件の異なる条件で暗号化されたデータを用いて学習された多くのモデルをアンサンブルのために用意する必要がある [4]．第二のアプローチは，敵対的事例を事前に検出する方法である．クラス分布の距離 [5] や，ノイズ除去フィルタ処理前後の出力結果の差や Auto Encoder の組合せを用いて敵対的事例を検出できることが報告された [6], [7]．さらに，これらの二つのアプローチ

はそれぞれ独立に研究されてきたが，両者の敵対的事例に対する異なる性質に着目して，両者を同時に使用する画像分類システムが提案された [8].

このように多くの敵対的事例に対する防御手法が提案されているが，各防御手法の頑健性を客観的に比較することは容易ではない．本稿では，敵対的事例の防御手法を検証するために提案された AutoAttack と呼ばれる敵対的攻撃のアンサンブルによって，秘密鍵を用いた分類器の防御性能を検証する．さらに，秘密鍵を用いた分類器を敵対的事例検出器と組合わせる際に，秘密鍵を用いた分類器の弱点を補うことを可能とする敵対的事例検出器について考察する．実験では，秘密鍵を用いた画像分類器を，CIFAR10 データセットと ResNet18 を用いて学習し，AutoAttack に対して評価を行った．その結果，秘密鍵を用いた分類器の長所と短所が明らかになる．さらに，敵対的事例検出器を組合わせ，拡張された分類器は，分類器を単体で使用した場合の短所を補い，より頑健なシステムであることが示された．その AutoAttack に対する頑健性は，最先端のロバスト分類器を上回る結果である．

2. 関 連 研 究

2.1 敵対的事例攻撃

入力信号に敵対的事例と呼ばれる人為的に作成された微小なノイズを，ニューラルネットワークモデルに加えることによって，意図的に予測結果を操作する攻撃を敵対的事例攻撃がある．敵対的攻撃の作成手法はホワイトボックス型，ブラックボックス型，グレーボックス型に分けられる．ホワイトボックス型では，攻撃対象となるモデルの構造やパラメータ，防御手法全てにアクセスできると仮定される．一方，ブラックボックス型とは，入出力のみにアクセスでき，攻撃対象のモデルにアクセスできない攻撃である．グレーボックス型とは，モデルに関する一部の情報のみにアクセス可能な攻撃である．

ホワイトボックス型の攻撃として，最初に FGSM (fast gradient sign method) が [2] 提案された．PGD (Projected Gradient Descent) [9] では，FGSM を反復して計算することによって，より高い攻撃成功率を達成している．反復最適化に基づく攻撃には，C&W (Carlini and Wagner attack) [10] や EAD (elastic-net attack) [11] などがある．ブラックボックス型の攻撃手法は，主にクエリベースの攻撃と転送ベースの攻撃がある．RayS 攻撃 [12] や Square 攻撃 [13] などのクエリベースの攻撃は，攻撃対象のモデルに入力を繰り返し，その出力に応じてノイズを調整することによって敵対的事例を生成する．転送ベースの攻撃である Adaptive Black-Box 攻撃 [14] は，訓練データのラベルを取得するために攻撃対象のモデルに入力を繰り返し，ラベル付けされたデータを使用して訓練した別の分類器に対して攻撃を行うことで敵対的事例を生成する．

敵対的攻撃に対する防御手法の評価基準として，AutoAttack と呼ばれる攻撃のアンサンブルが提案されている [15]．AutoAttack は，ホワイトボックス型とブラックボックス型を含む 4 つの攻撃手法で構成され，一つの防御モデルを異なる攻撃手法によって繰り返し攻撃する．本稿では，秘密鍵を用いたロバストな分類器を AutoAttack によって評価する．

2.2 敵対的攻撃に対する防御手法

敵対的攻撃に対する防御手法は，敵対的攻撃に対して頑健な分類器を設計する手法と，敵対的攻撃を分類器の入力前に検出する手法に分けられる．

2.2.1 敵対的攻撃に対してロバストな分類器

ロバストな分類器は，種々の敵対的攻撃に対して十分な耐性を有し，かつクリーンな入力に対しても耐性を有しない一般の分類器と同等の分類精度を持つことが要求される．

FGSM に代表されるように，敵対的事例を学習に使用する画像に追加することによって，敵対的事例に対して分類器が頑健になることがまずは報告された [2]．しかし敵対的事例を用いて学習されたモデルは，学習時に想定されていない攻撃に対して頑健性を持たないだけでなく，クリーンな画像に対する精度を低下させる [9], [16]．さらに，付加されたノイズを除去する手法や自己教師あり学習 [17], [18] などの方法も提案されているが，頑健性維持の汎用性やクリーンな画像使用時における精度低下が同様に課題となっている．

一方，特定の攻撃法を仮定しないロバストな分類器として，MaungMaung らによって秘密鍵を用いて画像を変換してモデルを学習する方法が提案された [3]．この手法は，入力画像に対して，秘密鍵を用いたブロック単位の画像変換を施すことで，攻撃者が作成した敵対的事例を無効化することを目指すものである．提案されたブロック単位の画像変換は，DNN で学習可能な変換画像を生成可能であり，クリーンな画像における精度低下も小さい．しかし，課題として，攻撃者が正しい鍵を知っている場合や，勾配情報を用いないブラックボックス攻撃に対して防御できないことがある．ブラックボックス攻撃に対する対策として，秘密鍵を用いた分類器のアンサンブルが検討された [**?**].

2.2.2 敵対的事例検出器

敵対的事例攻撃に対する他の対策として，敵対的事例を事前に検出する方法がある．例えば，分類器の条件付きガウス分布のマハラノビス距離 [5] や，敵対的事例とクリーンな画像は特徴マップが異なること [19], [20] を用いて，敵対的事例が検出可能であることが報告された．また，ノイズ除去フィルター処理の影響を観察することによって，敵対的事例を検出する方法 [6], [7] が報告されている．

検出器は敵対的攻撃に頑健性を持たない一般的な分類器と共に使用されてきた．そのため，検出ミスは分類精度に直接影響する．したがって，検出器は理想的には全ての攻撃を検出する必要がある．また，敵対的事例の検出は，強いノイズが負荷された場合ほど容易である．これは，ロバスト分類器の持つ特徴と逆の特徴となる．

2.2.3 検出器を組み合わせた画像分類システム

敵対的事例の検出は，弱いノイズが付加された場合ほど難しい．一方，ロバストな分類器の実現では，強いノイズが付加された場合ほど正しく分類することが難しい．このような着眼点から，敵対的事例検出と画像分類器を組み合わせることが提案された [8]．この組み合わせは，互いの防御法の弱点を補い，分

類器の防御性能を向上させることが期待できる．

本稿の目的の一つは，この組み合わせを洗練させることにある．使用するロバスト分類器（秘密鍵を用いた分類器）を想定して，その弱点を補う検出器はどうあるべきかを考察する．

3. AutoAttack

種々の敵対的事例攻撃と同時に，種々の防御法が提案されている．しかし，各種防御法の客観的比較や評価は容易ではない．一般に防御法は，特定の攻撃手法には高い頑健性を持っていても，他の攻撃手法には頑健性を持たないことが多い．

AutoAttack [15] は，敵対的攻撃に対する防御をできる限りの条件の下で客観的に評価・比較するために提案された攻撃のアンサンブルである．複数の攻撃手法を用いて繰り返し防御モデルを攻撃することによって，防御法の頑健性を検証する．多様な攻撃手法によって防御モデルを攻撃することによって，各防御法の弱点を知ることもできる．

画像分類に使用されるテスト画像の枚数を N 枚，テスト画像として準備されたクリーンな画像の集合を X，X から生成される敵対的事例の集合を X_{adv} とする．分類器モデルを C とした時，AutoAttack による攻撃は以下の手順で実行される．

手順 1. テストしていない攻撃手法によって X から X_{adv} を作成する．

手順 2. X_{adv} を用いて C をテストする．C によって正しく分類された X_{adv} に対応するクリーンな画像の集合を X_{true} とする．

手順 3. X_{true} の要素が 0 枚の場合，または全ての攻撃手法が終了した場合，テストを終了する．

手順 4. $X = X_{true}$ として，手順 1 に戻る．

X_{true} の枚数を $|X_{true}|$ としたとき，Accuracy (Acc) は以下のように計算される．

$$\text{Acc} = \frac{|X_{true}|}{N} \times 100 \tag{1}$$

標準的な AutoAttack には，Auto-PGD-cross entropy [15] (APGD-ce)，Auto-PGD-target (APGD-t), FAB-target (FAB-t) [21], Square 攻撃 [13] の 4 つの攻撃手法が含まれる．このうち，Square 攻撃のみがブラックボックス型の攻撃であり，他の手法はホワイトボックス型攻撃である．また APGD-t と FAB-t は target 型，APGD-ce と Square は non-target 型である．各攻撃手法について詳述する．

- APGD [15]：固定のステップサイズを使用する PGD [9] を改善した攻撃で，最適化の進捗に応じたステップサイズの選択と，ステップサイズを減らした時に発見されている一番良い地点から最大化を再開することが特徴である．
- FAB [21]：誤分類を達成するために必要な摂動のノルムを最小化する手法であり，勾配マスキングが行われたモデルに対しても有効であることが指摘されている．
- Square [13]：ランダム探索を用いたクエリベースのブラックボックス型攻撃で，勾配近似を一切利用しない．

4. 秘密鍵を用いた分類器の評価

4.1 分類器の概要

本稿では，AutoAttack を用いて，秘密鍵を用いた分類器の敵対的事例攻撃に対する防御性能を評価する．秘密鍵を用いた分類器 [3] は，モデル学習用の入力画像とテスト画像に秘密鍵を用いたブロック単位の画像変換を施すことで，敵対的事例の無毒化を目指す．画像の変換方法としては，ピクセルシャッフリング，ビットフリッピング，FFX (Feistel-based encryption) 暗号化の 3 つが提案された．本稿では，ブロック内のピクセルをランダムに入れ替える画像変換であるピクセルシャッフリングを，先行研究で推奨されたブロックサイズ 4×4 の条件で使用する．

秘密鍵を用いた分類器は，攻撃手法に対して特に仮定を必要としないため，適応的攻撃を含むホワイトボックス型攻撃に対して高い防御性能を有することが確認された．しかし，鍵を攻撃者に知られてしまった場合には，一切の防御力を持たない．さらに，ブラックボックス型攻撃に対しては，攻撃者が鍵にアクセスができない条件でも，防御力が低下するが指摘された．本稿では，AutoAttack に対する秘密鍵を用いた分類器の防御性能を評価して，攻撃者に鍵を知られた場合の対策と，ブラックボックス型攻撃に対して頑健性の向上を考察する．

4.2 分類器の拡張

ロバストな分類器と敵対的事例検出器を同時に使用することによって，分類器の弱点を補強できることが示された [8]．敵対的事例検出器を組合わせた画像分類システム (図 1(a)) は，ロバストモデル (図 1(b)) の防御を突破する可能性のある敵対的事例を事前に検出することが期待される．加えて，敵対的事例検出器 (図 1(c)) では検出が困難な敵対的事例がロバスト分類器に入力されても，正しく分類することが期待される．本稿では，ロバスト分類器として秘密鍵を用いたロバストな分類器の使用を想定する．そして，その弱点を補強できる検出器を考察する．具体的には，秘密鍵を用いた分類器と一般の分類器の出力の差を利用した検出器を提案して，画像分類システム全体の頑健性を AutoAttack によって評価する．

4.3 鍵を用いるロバストな分類器のための検出器

敵対的事例検出器の一つとして，特徴の異なる分類器の出力の違いを特徴量に用いる方法が提案された [8]．図 2 に示すように，その検出器は特徴抽出部と判別器からなる．検出器は分類器の出力を特徴量としており，一方の分類器が攻撃され分類を誤った時にも，他方の分類器は正しく分類結果を出力することが仮定される．一般の分類器と秘密鍵を用いる分類器の間では，敵対的事例の影響が遷移され難い．

提案する検出器は，以下の点が考慮されている．

(a) 特徴量に Logit を用いる．

(b) 判別器は多項式回帰に基づき学習される．

(c) 判別器の学習には，クリーンな画像に加えて，2 つの分類器に異なる分類結果を出力させた敵対的事例のみを使用

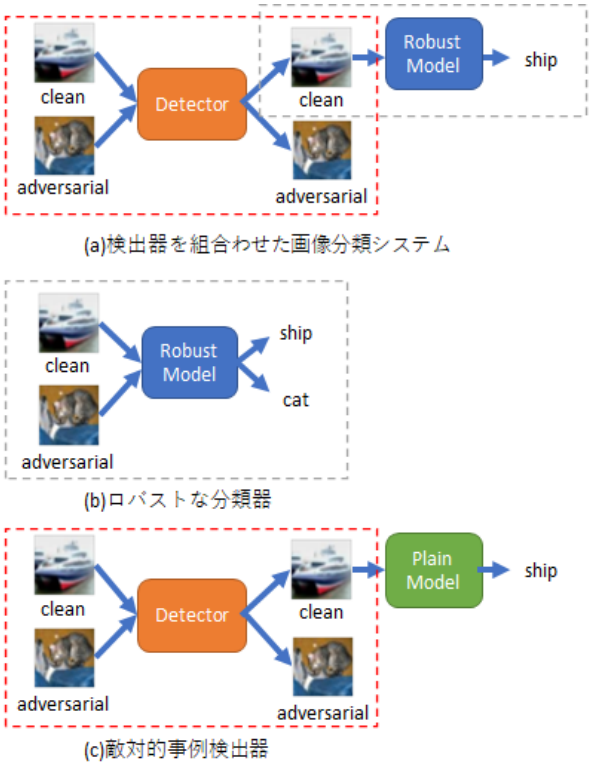


図 1　ロバストな画像分類システム

する．

まず (a) について，Softmax 層を通した分類器の最終的な出力は，各ラベルが $[0, 1]$ に正規化されるため敵対的事例の特徴が十分に特徴が表れにくいことから，Softmax 層を通す前の Logit を使用する．次に (b) について，ロジスティック回帰よりも表現力の高い多項式回帰を使用する．本稿では，攻撃に失敗した敵対的事例は判別部の学習には使用せず，多項式回帰を (c) のように学習する．テスト時には，判別器の出力が，予め決められた閾値 $(0 < \delta_{th} < 1)$ を超えた場合に，敵対的事例と判断する．

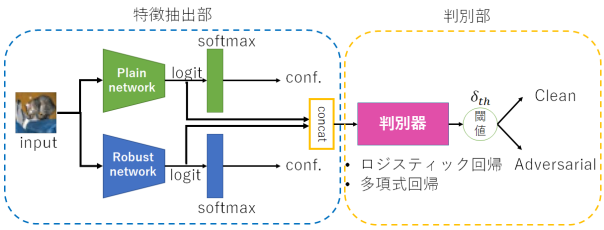


図 2　敵対的事例検出器

5. 実　　験

5.1 実験条件

5.1.1 敵対的事例の攻撃条件

APGD-ce，APGD-t，FAB-t，Square 攻撃の 4 つの攻撃を実行する標準的な AutoAttack [15] によって，モデルの頑健性を評価した．秘密鍵を用いた分類器への攻撃では，攻撃者が正しい鍵を知っている (鍵既知) 条件と，鍵を知らない (鍵未知) 条件の下でモデルの頑健性を評価した．鍵が未知の時，攻撃者は分類器のパラメータと画像変換方法を入手可能である．一方、鍵が既知の時には，それらに加え，攻撃者は秘密鍵を入手できる．鍵既知の場合，攻撃者はモデルの勾配を正しく学習することができる．従って，頑健性を持たない一般の分類器と同等の攻撃が可能となる..

5.1.2 防御手法の実験条件

使用したデータセットは，画像分類タスクに広く使用されている CIFAR10 データセットである．CIFAR10 は動物や乗り物を含む 10 種類のクラスを持ち，32×32 ピクセルの画像 60000 枚 (学習用：50000 枚，テスト用：10000 枚) で構成されている．

検出器の特徴抽出部の学習において (図 2 参照)，2 つの分類器 (一般的な分類器 (ResNet18) と秘密鍵を用いた分類器 (ResNet18)) を CIFAR10 の 50000 枚で学習した．一方，判別部は，2000 枚のクリーン画像と，そこから生成された敵対的事例 (FGSM) 2000 枚のうち攻撃が成功した (特徴抽出部の 2 つの分類器が異なる結果を出力した) 敵対的事例を用いて学習された．

テスト時には，拡張された分類器に対して，特徴抽出部に使用した秘密鍵を用いる分類器と同一の分類器を使用した．拡張された分類器全体のテストのために，テスト画像 10000 枚のうち，判別部の学習に用いた 2000 枚を除く，8000 枚から敵対的事例を生成し使用した．

一方，秘密鍵を用いた分類器のテストには，50000 枚のクリーンな画像を鍵を用いて変換した画像によってモデルを学習し，10000 枚の敵対的事例を用いてそのモデルの頑健性を評価した．なお，秘密鍵を用いた分類器の学習に使用された画像は，先に述べたように，4 × 4 のブロックサイズに分割され，ピクセルシャッフリングが適用され生成された．

5.2 評価指標

5.2.1 秘密鍵を用いた分類器

敵対的事例検出器を用いずに，秘密鍵を用いた分類器単独の性能を，式 (1) によって定義される Acc に基づいて評価する．式 (1) におけるテスト画像の枚数 N は，クリーンな画像とそこから生成される敵対的事例の枚数がともに 10000 枚であるので，N =10000 枚である．

ベンチマークとして比較された 2 つの先端研究 [17], [18] は，秘密鍵を用いた分類器と同一の条件の下で評価される．

5.2.2 拡張された分類器

敵対的事例検出器を組合わせた拡張された分類器では，入力画像はまず検出器に入力され，検出器にクリーン画像と判断された画像のみが分類器に入力される．従って，式 (1) の Acc における N は，判別部の学習に CIFAR10 のテスト画像 10000 枚のうち 2000 枚を使用しているため，N =8000 である．X_{true} の枚数は，検出器に敵対的事例と判断されかつその判断が正しかった画像の枚数と，検出器にクリーン画像と判断されかつ分類器の分類結果が正しかった画像の枚数の総和である．

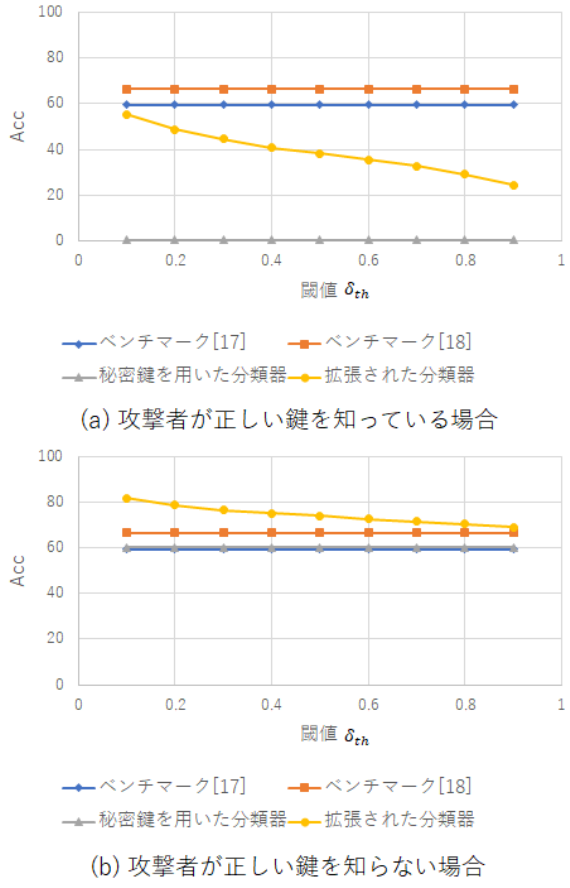
5.3 実験結果

5.3.1 秘密鍵を用いた分類器の評価

AutoAttack に対する各手法の Acc を図 3 に示す．図 3(a) には鍵既知の条件，図 3(b) には鍵未知の条件の結果がそれぞれ示されている．　各手法は，ロバスト分類器のランキングサイトであ

る Robust bench で最も高い頑健性を持つ手法 [18] と，文献 [15] において最も頑健性があった手法 [17] をベンチマークとして各手法を比較した．表 1 は，AutoAttack に含まれる 4 つの各攻撃法に対する Acc である．

図 3(a) から，鍵既知の場合，秘密鍵を用いた分類器は Acc がほぼ 0 であり，著しく Acc が低いことが分かる．また，図 3(b) に示すように鍵未知の場合であっても，秘密鍵を用いた分類器の Acc は，ベンチマーク [17] と同等であるが，ベンチマーク [18] の Acc を下回っている．表 1 から，耐性が Square（ブラック攻撃）に対してはベンチマークに比べ劣るが，他の攻撃に対しては優れていることが分かる．ブラックボックス攻撃である Square 攻撃の影響によって，AutoAttack 全体の Acc 低下につながったことがわかる．



(a) 攻撃者が正しい鍵を知っている場合

(b) 攻撃者が正しい鍵を知らない場合

図 3　AutoAttack に対する Acc

表 1　攻撃手法毎の Acc

		clean	APGD-ce	APGD-t	FAB-t	Square
ベンチマーク [17]		89.69	61.74	59.54	60.12	66.63
鍵既知	秘密鍵を用いた分類器	90.66	1.97	1.82	56.17	64.75
	拡張された分類器	85.35	87.16	85.71	73.09	82.51
鍵未知	秘密鍵を用いた分類器	90.66	90.39	90.45	90.66	64.75
	拡張された分類器	85.35	94.21	94.24	93.49	82.39

5.3.2　拡張された分類器の評価

図 3(a) より，敵対的事例検出器を組合わせることによって，鍵を用いた分類器のは鍵既知の場合にも向上することがわかる．さらに鍵未知の場合，鍵を用いた分類器の頑健性は，ベンチマークの Acc を上回った．表 1 から，拡張された分類器は鍵既知の場合にもベンチマークよりも高い Acc を維持した．鍵未知の場合も同様に，ブラックボックス攻撃も含め，すべての攻撃に対して Acc が向上した．しかし，clean の場合の Acc は，僅かに低下した．これは，クリーン画像に対する検出精度の低下が起因したものである．

6.　む　す　び

本稿では，まず敵対的事例攻撃のベンチマーク法である AutoAttack によって，秘密鍵を用いた画像分類器を評価した．CIFAR10 データセットと ResNet18 を用いた実験において，秘密鍵を用いた画像分類器は，ホワイトボックス攻撃に対しては高い耐性を有するが，攻撃者に鍵を知られた場合やブラックボックス攻撃に対して脆弱性があることを確認した．また秘密鍵を用いた画像分類器のための敵対的事例攻撃検出器を提案し，それと秘密鍵を用いた画像分類器を組み合わせた画像分類器を提案した．またその拡張された画像分類器の頑健性を AutoAttack によって評価した．その結果，攻撃者に鍵を知られた場合やブラックボックス攻撃に含め耐性が向上し，拡張された画像分類器の頑健性は，最先端のベンチマークを上回ることを確認した．

謝辞　この研究は 2022 年度国立情報学研究所公募型共同研究 (22S1401) および JST CREST JPMJCR20D3 の助成を受けています．

文　献

[1] H. Kiya, M. AprilPyone, Y. Kinoshita, S. Imaizumi, and S. Shiota, “An overview of compressible and learnable image transformation with secret key and its applications,” APSIPA Transactions on Signal and Information Processing, vol.11, no.1, e11, 2022.

[2] I.J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[3] M. AprilPyone and H. Kiya, “Block-wise image transformation with secret key for adversarially robust defense,” IEEE Transactions on Information Forensics and Security, vol.16, pp.2709–2723, 2021.

[4] M. AprilPyone and H. Kiya, “Ensemble of key-based models: Defense against black-box adversarial attacks,” Proc. of Global Conference on Consumer Electronics, pp.95–98, 2021.

[5] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” Proc. of Neural Information Processing Systems, p.7167–7177, 2018.

[6] Y. Yamasaki, M. Kuribayashi, N. Funabiki, H.H. Nguyen, and I. Echizen, “Feature extraction based on denoising auto encoder for classification of adversarial examples,” 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.1815–1820, 2021.

[7] 山崎裕真，栗林　稔，舩曳信生，グエン フイホン，越前　功，“Auto encoder に対する応答特性を用いた敵対的事例の検出法，” EMM)，pp.1815–1820，2021.

[8] T. Osakabe, M. AprilPyone, S. Shiota, and H. Kiya, “Adversarial detector with robust classifier,” Proc. of Global Conference on Life Sciences and Technologies, pp.179–182, 2022.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, 2018.

[10] N. Carlini and D. Wagner, “Towards evaluating the robustness of neu-

ral networks," IEEE Symposium on Security and Privacy, pp.39–57, 2017.
[11] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "Ead: Elastic-net attacks to deep neural networks via adversarial examples," Proc. of the AAAI Conference on Artificial Intelligence, vol.32, no.1, Apr. 2018.
[12] J. Chen and Q. Gu, "Rays: A ray searching method for hard-label adversarial attack," KDD '20, Association for Computing Machinery, New York, NY, USA, 2020.
[13] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," Computer Vision – ECCV 2020, eds. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, pp.484–501, Springer International Publishing, Cham, 2020.
[14] P.H. Nguyen, K. Mahmood, L.M. Nguyen, T. Nguyen, and M. vanDijk, "Buzz: Buffer zones for defending adversarial examples in image classification," CoRR, vol.abs/1910.02785, 2019. http://arxiv.org/abs/1910.02785
[15] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," Proceedings of the 37th International Conference on Machine Learning, ICML'20, 2020.
[16] A. Kurakin, I.J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017.
[17] Y. Carmon, A. Raghunathan, L. Schmidt, J.C. Duchi, and P.S. Liang, "Unlabeled data improves adversarial robustness," Advances in Neural Information Processing Systems, vol.32, 2019.
[18] S.-A. Rebuffi, S. Gowal, D.A. Calian, F. Stimberg, O. Wiles, and T.A. Mann, "Data augmentation can improve robustness," Advances in Neural Information Processing Systems, vol.34, pp.29935–29948, 2021.
[19] P. Yang, J. Chen, C.-J. Hsieh, J.-L. Wang, and M. Jordan, "Ml-loo: Detecting adversarial examples with feature attribution," Proceedings of the AAAI Conference on Artificial Intelligence, vol.34, no.04, pp.6639–6647, Apr. 2020.
[20] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks meet interpretability: Attribute-steered detection of adversarial samples," Proceedings of the 32nd International Conference on Neural Information Processing Systems, p.7728–7739, NIPS'18, 2018.
[21] F. Croce and M. Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," Proc. of the 37th International Conference on Machine Learning, eds. by H.D. III and A. Singh, vol.119, pp.2196–2205, Proceedings of Machine Learning Research, 13–18 Jul 2020.