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.

敵対的事例検出器を用いたロバストな画像分類システムの拡張

田中　美貴† 　長我部恭行† 　貴家　仁志†

† 東京都立大学システムデザイン学部，東京都
E-mail: †{tanaka-miki,osakabe-takayuki}@ed.tmu.ac.jp, ††kiya@tmu.ac.jp

あらまし　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に加えた敵対的事例攻撃によって，深層学習モデルの予測結果を操作される危険性があることが知られている．本稿では，敵対的事例にロバストな新しい画像分類システムを提案する．敵対的事例の防御手法として，敵対的攻撃に対してロバスト画像分類器を構築するする方法と，分類器への入力前に敵対的事例を検出する方法の二つが代表的である．従来これらは独立に研究されてきたが，先に両者の同時使用が，互いの短所を補うことが指摘された．本稿では，その先行研究をさらに拡張して，システム設計時に必要であった統計的仮定を緩和した画像分類システムを提案する．仮定外の攻撃に対して防御性能が大きく劣化するという従来法の課題を，複数の特徴を用いることによって，提案法は改善することが可能である．
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

Extention of robust image classification system with Adversarial Example Detectors

Miki TANAKA†, Takayuki OSAKABE†, and Hitoshi KIYA†

† Kiya's laboratory, Department of Computer Science, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo, 191-0065 Japan
E-mail: †{tanaka-miki,osakabe-takayuki}@ed.tmu.ac.jp, ††kiya@tmu.ac.jp

Abstract　In image classification with deep learning, there is a risk that an attacker can intentionally manipulate the prediction results of image classification by using images with a small designed noise, called adversarial examples. In this paper, we propose a robust image classification system against adversarial examples. In order to prevent the attack, there are two approaches: using a robust classifier, and using a detection method of adversarial examples. A robust image classification system with the combination of these two approaches was demonstrated to outperform conventional methods. In this paper, we extend the robust image classification system with the combination of the two approaches by using multiple features. The proposed method is more robust against various adversarial attacks and noise levels than the conventional one.
Key words　Adversarial example, Machine learning, Deep learning, Adversarial detection

1. ま え が き

深層学習モデルは，人為的に作成された微小なノイズを入力に付加した敵対的攻撃によって予測結果が操作される脆弱性を持っていることが知られている．ノイズは人間が判別不能な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，自動運転で使われる画像認識など，安全性や高いセキュリティが求められるアプリケーションにとって大きな脅威となる．従って，敵対的攻撃に対する防御策が急務の課題となっている．

敵対的攻撃に対する防御策は，主に 2 つのアプローチが検討されてきた．1 つはモデルを敵対的攻撃に対してロバストに設計する手法である．敵対的事例を含んだトレーニング画像によって学習を行う一種のデータ拡張によってロバスト性を持たせる手法 [1] や，秘密鍵を用いた画像変換によって，敵対的事例に対してロバストなモデルを学習する手法 [2] が提案されている．2 つ目のアプローチは，モデルへの入力前に敵対的事例を検出する手法である．学習分布から離れた入力を検出するために，クラス条件付きガウス分布のマハラノビス距離を求める手法 [3] や，クリーンな画像と敵対的事例の Feature Attribution の違いを用いた検出手法 [4] がある．

上記の 2 つのアプローチは，それぞれ独立に研究されてきたが，両者が攻撃手法やノイズレベルに対して異なる性質を持つことから，両者の同時使用がお互いの短所を補い合うことが指

摘された [5]．本稿では先行研究のシステムを，検出器を複数用いることによって拡張する．提案手法は学習に使用した攻撃手法やノイズレベルではない仮定外の敵対的攻撃に対して，従来法よりも高い防御性能を持つことを確認する．また検出手法の性質から，検出器を FGSM (fast gradient sign method) [1] のような弱いクラス誘導をするタイプの攻撃手法かつ，できるだけ弱いノイズレベルによって学習することで，仮定外の敵対的事例に対して頑健になることを示す．

2. 関 連 研 究

2.1 敵対的攻撃

敵対的攻撃は，ResNet などのニューラルネットワークモデルへの入力に，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加えることによって，意図的に予測結果を操作する攻撃である．付加されるノイズは，人間の目には識別ができないほど微小であっても攻撃が可能である．敵対的攻撃は，予測結果を本来のクラスから離れるようにする non-target attack と 攻撃者が予測結果を意図したクラスに誘導する target attack に分類できる．本稿では主に taget attack を対象とする．

敵対的事例の作成手法は，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は攻撃者が攻撃対象のモデルやパラメータ，防御手法にアクセスできると仮定される．一方で，ブラックボックス型は，モデルの出力を除いて，モデルに関する情報を持っていないとかていされる．グレーボックス型は，ホワイトボックス型とブラックボックス型の中間にあたり，モデルに関する部分的な情報を持っていると仮定される．

ホワイトボックス型の敵対的事例を作成する簡単で一般的な方法として, FGSM がある．PGD (Projected Gradient Descent) [6] は FGSM を反復して計算され, 攻撃者の意図したクラスに FGSM よりも強く誘導する．反復最適化に基づく攻撃には CW(Carlini and Wagner attack) [7] や，EAD(elastic-net attack) [8] などがある．

ブラックボックス型の攻撃手法は，勾配を推定する手法 [9]，入力を中心とした確率分布から抽出したサンプルを敵対的事例とする NATTACK と呼ばれる手法 [10]，数個のピクセルを変更することで敵対的事例を作成する OnePixel 攻撃 [11] などがある．

2.2 敵対的攻撃に対する防御手法

敵対的攻撃を防御する方法は主に，敵対的攻撃に対してロバストなモデルを設計する方法と，モデルへの入力前に敵対的事例を検出する方法に分けられる．

2.2.1 敵対的攻撃に対してロバストなモデル

ロバストなモデルとして，学習データに FGSM ベースの敵対的事例を混ぜることでモデルがロバスト性を獲得することが報告されている [1]．しかしこの手法は，学習した敵対的攻撃以外，例えば PGD のような反復的な敵対的事例が入力された場合には防御性能を持たないことが指摘されている [6], [12]．さらに，これらの防御手法はクリーンな画像の分類精度を低下させてしまう．攻撃手法を仮定しない方法として，MaungMaung の手法 [2] が提案されている．この手法は，入力画像に前処理として，秘密鍵を用いたブロック単位の画像変換：ピクセルシャッフリング，ビットフリッピング，FFX (Feistel-based encryption) 暗号化を行うことで，敵対的攻撃に対してロバストなモデルを学習する．

2.2.2 敵対的攻撃の検出

Lee らの手法 [3] では，学習分布から離れた入力を検出する OOD (out-of-distribution) 検出のひとつである．分類器モデルの特徴に関するクラス条件付きガウス分布を求め，そのマハラノビス距離に基づいた信頼度スコアから敵対的事例を検出する．また敵対的事例とクリーンな画像は，統計的に異なる Feature attribution を持つことが報告されている [4], [13]．先行研究 [5] では，ロバストな分類器に基づく検出器を提案している．[5] は敵対的事例が入力された場合に生じる一般的な分類器とロバストな分類器の出力の違いから，各分類器の Softmax 層に通す前の Logit を特徴量として用いる．本稿では，ロバストなモデルと敵対的事例検出器を組み合わせたロバストな画像分類システムを，検出器を複数用いることによって，ノイズレベルの変化や攻撃手法の変化に対してよりロバストに拡張する．

3. 敵対的事例検出器

3.1 二つの防御手法を組み合わせた画像分類システム

ロバストなモデルと敵対的事例の検出手法を組み合わせた画像分類システムが，先行研究 [5] で提案されている．ロバストな画像分類器 (図 1(a)) は強いノイズレベルに対しては防御性能が低下する一方，敵対的事例の検出器 (図 1(b)) は，より強いノイズレベルにおいて検出が容易になる．このように，二つの防御手法は敵対的攻撃に対して対照的な性質を持つため，これらを同時に使用する場合には，互いの性質を補い合った高い防御性能を維持できることが期待できる．具体的に，2 つの防御手法を組み合わせたロバストな画像分類システムは，ロバスト分類器が誤りやすい強いノイズレベルの敵対的事例を検出器によって取り除くことが可能である．また検出器が弱いノイズレベルの敵対的事例を誤ってクリーンな画像と識別した場合にも，ロバストな分類器は問題なく分類できると期待できる．ロバストな画像分類システムの概要を図 1(c) に示す．

3.2 拡張されたロバスト画像分類システム

敵対的事例検出器は，他の防御手法と同様に，攻撃手法に対して異なる性質を持つ．例えば，長我部らの手法は一般的な分類器とロバストな分類器の出力結果の差を特徴量とするので，特定のクラスに強く誘導するタイプの敵対的事例の方が検出が容易である．そのため，長我部の手法は PGD に対しては高い検出性能を持つが，FGSM は PGD の検出と比べて難しい．一方，Lee らの手法 [3] は，敵対的事例と OOD と呼ばれる学習分布から離れたサンプルの二つを検出することを目的としており，クラス条件付きガウス分布のマハラノビス距離を特徴量としている．従って，クラスの予測が曖昧になりやすい FGSM に対しては，長我部の手法と比べて高い検出性能を持つが，PGD においては長我部らの手法 [5] よりも検出性能が低い．

そこで本稿では，敵対的攻撃の検出器とロバストな分類器を組み合わせた画像分類システムに加えて，複数の異なる特徴を

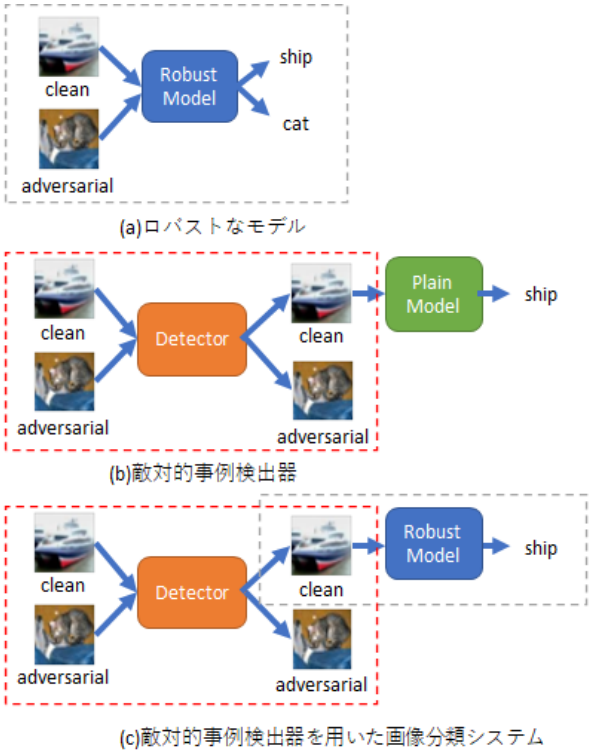

図 1　ロバストな画像分類システム

検出器に用いて画像分類システムをさらに拡張する．敵対的事例検出器は，一般に図 2(a) のように，特徴抽出器とロジスティック回帰などの線形モデルで構成される．特徴抽出器の構造は上述通り様々であるが，後列の線形モデルはほとんどの手法でロジスティック回帰が採用されている．線形モデルの出力は任意に決められる閾値 (th) に従って，出力が閾値より低い場合にはクリーンな画像，高い場合は敵対的事例と判断される．

まず，異なる特徴抽出器と対応する線形モデルを並列に使用する手法 (ensemble) の概要を図 2(b) に示す．検出の手順は以下のようである．

1. 異なる特徴抽出器から特徴量を抽出．
2. 各対応する線形モデルと適切な閾値によって敵対的事例を検出．
3. ２つの検出結果をアルゴリズムに従って統合．

特徴抽出器と線形モデルの組み合わせや，アルゴリズムの選択には自由度がある．本稿ではアルゴリズムに，どちらかの線形モデルが敵対的事例と判断した場合に敵対的事例とみなす (Ensemble (adv)) と，どちらかの線形モデルがクリーンな画像だと判断した場合にクリーンな画像とみなす (Ensemble (clean)) を適用する．

次に，異なる特徴抽出器で抽出した特徴を組み合わせて１つの特徴ベクトルとし，1 つの検出器 (ロジスティック回帰) を学習する手法 (Concat) の概要を図 2(c) に示す．検出の手順は以下のようである．

1. 異なる特徴抽出器で特徴量を抽出．
2. ２つの特徴量を 1 つの特徴量に組み合わせる．
3. 1 つの線形モデルと適切な閾値によって敵対的事例を検出．

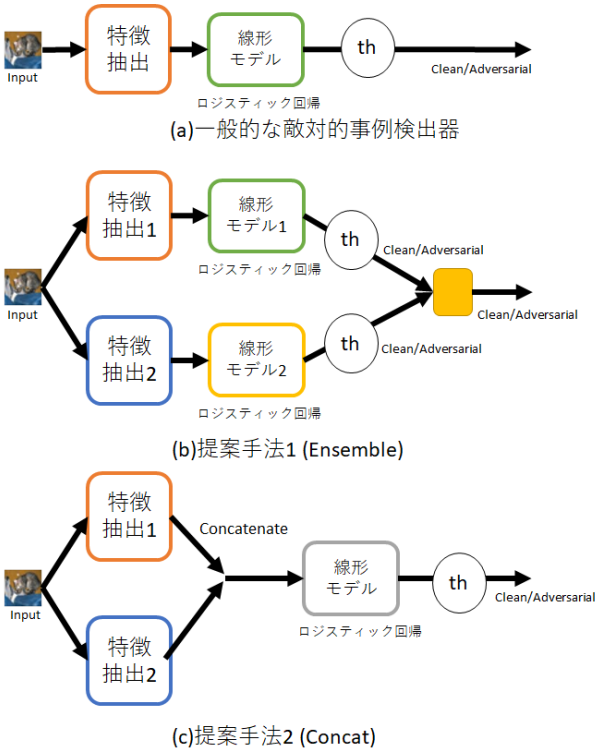

図 2　検出器の拡張

4. 実　　験

4.1 実験条件

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 データセットを使用する．本データセットは，動物や乗り物を含む 10 種類のクラスを持つ，32×32 ピクセルの画像 60000 枚 (トレーニング画像: 50000 枚，テスト画像: 10000 枚) で構成されている．トレーニング画像は，一般的な分類器とロバストな分類器の学習に用いた．

敵対的攻撃は，ホワイトボックス型攻撃である PGD [6], FGSM [1] によって行った．データセットのテスト画像の内 8000 枚と，敵対的攻撃が施された同 8000 枚の計 16000 枚を検出器の線形モデルの学習に使用した．テスト画像の残りの 2000 枚とその敵対的事例 2000 枚の計 4000 枚を，画像分類システムのテストに使用する．

検出器は，長我部らの手法 [5] と Lee らの手法 [3] について，単体で画像分類システムに適用した場合と，提案手法に従って 2 つの検出手法を組み合わせて適用した場合の防御性能を比較する．

一般的な画像分類器は，ResNet18 [14] を使用する．ロバストな分類器は，入力画像に秘密鍵を用いたブロック単位の画像変換 [2] 適用し，一般的な分類器と同様に ResNet18 を使用した．

4.2 評価指標

本稿では，以下のような識別器の Accuracy (Acc) によって提

案手法を評価する．

$$\mathrm{Acc} = \frac{N_{1yes}^{clean} + N_{2yes}^{clean} + N_{2}^{noise}}{N_1 + N_2} \quad (1)$$

ここで，クリーンな画像の枚数を N_1，敵対的事例の枚数を N_2 とする．検出器にクリーンと判別された枚数をそれぞれ N_1^{clean}，N_2^{clean}．敵対的事例と判別された枚数をそれぞれ N_1^{noise}，N_2^{noise} とする．このとき，分類器に入力される画像の中で，分類器に正しいラベルに分類された枚数をそれぞれ N_{1yes}^{clean}，N_{2yes}^{clean} とする．

4.3 実験結果

4.3.1 提案手法の基本防御性能

検出器の学習と同様の攻撃手法とノイズレベルで攻撃された場合の，提案手法の防御性能を確認する．検出器の学習にはテストと同様の攻撃手法とノイズレベルを使用している．PGD(ϵ = 3/255) を使用して攻撃した時の Acc を図 3 に，FGSM(ϵ = 3/255) を使用して攻撃した時の Acc を図 4 に示す．図 3，4 からわかることを以下にまとめる．

(a) 攻撃手法に関わらず Concat (提案手法 2) がほとんどの閾値で最も高い Acc となっている．

(b) 検出手法が一つである場合と比べて，Concat は閾値が変化してもほとんど Acc が変わらない．

(c) Ensemble (clean) (提案手法 1) は小さい閾値，すなわち少しでも疑わしい画像を敵対的事例と判別する場合に Acc が高い．

(d) Ensemble (adv) (提案手法 1) は大きい閾値，すなわちより疑わしい画像のみを敵対的事例と判別する場合に Acc が高い．

(e) Ensemble (clean) は PGD，FGSM ともに Ensemble (adv) よりも Acc が高い，

以上の結果から，この条件下では 2 つの提案手法のうち，Concat が最も優れた防御性能を示した．

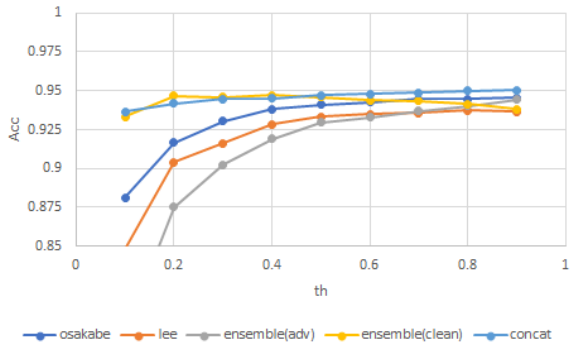

図 3　画像分類システムの Accuracy (PGD 攻撃)

4.3.2 ノイズレベルに対するロバスト性

検出器が学習していないノイズレベルに対するロバスト性について確認する．図 5 に，PGD(ϵ = 3/255) で検出器を学習した画像分類システムを，PGD(ϵ = 1,3,5,8,13/255) で攻撃した時の Acc を示す．図 5 から，Concat は学習したノイズレベルより高いノイズレベルの攻撃を容易に検出が可能であるため，弱いノイズレベルで学習を行うことが推奨される．

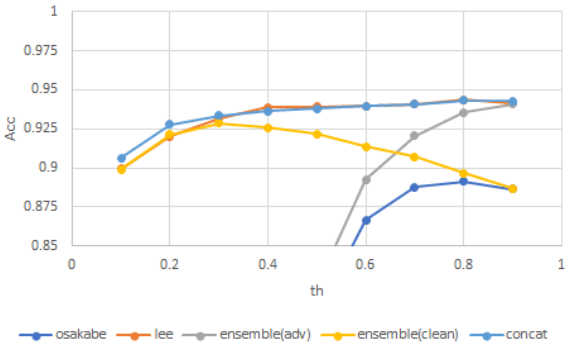

図 4　画像分類システムの Accuracy (FGSM 攻撃)

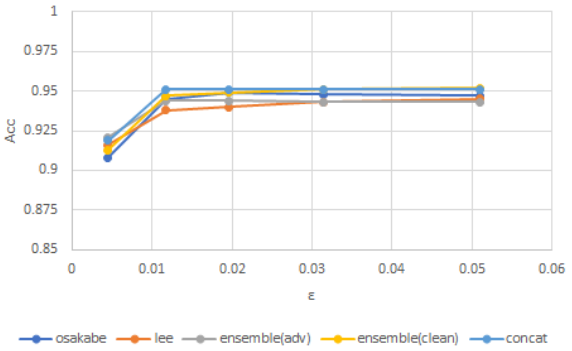

図 5　ノイズレベルに対するロバスト性

4.3.3 攻撃手法に対するロバスト性

検出器が学習していない攻撃手法に対するロバスト性について確認する．PGD(ϵ = 3/255) で検出器を学習した画像分類システムを，FGSM(ϵ = 1,3,5,8,13/255) で攻撃した Acc を図 6 に示す．また，FGSM(ϵ = 3/255) で検出器を学習した画像分類システムを，PGD(ϵ = 1,3,5,8,13/255) で攻撃した時の Acc を図 7 に示す，図 6，7 からわかることを以下にまとめる．

(a) 図 6 の条件 (PGD で学習) では，Concat と Ensemble (clean) は Lee の手法よりも低い Acc となった．

(b) 図 7 の条件 (FGSM で学習) で，どの手法も PGD で学習した場合と比べても Acc の低下が少なく，Concat と Ensemble (clean) は他手法よりも高い Acc となった．

以上の結果から，Concat は FGSM かつ弱いノイズレベルで学習した場合，異なる攻撃手法と広いノイズレベルに対して優れたロバスト性を示した．このことは，単独で使用した検出器やロバストな分類器に加えて，両者を単純に組み合わせた場合よりも高い防御性能であった．

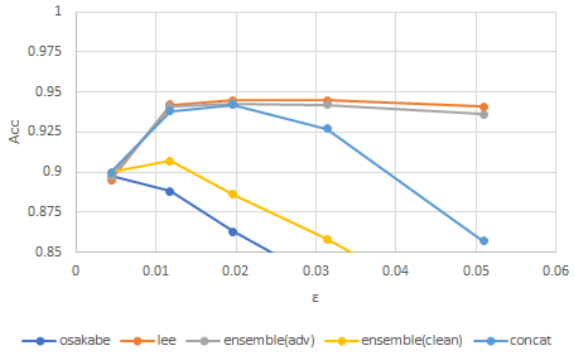

図 6　攻撃手法に対するロバスト性 (学習: PGD，テスト: FGSM)

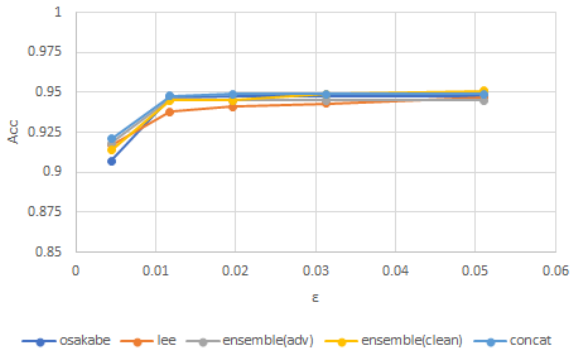

図 7　攻撃手法に対するロバスト性 (学習: FGSM，テスト: PGD)

5. む　す　び

本稿では，敵対的事例検出器を複数用いることによって，ロバスト画像分類器と敵対的事例検出器を同時使用するロバスト画像分類システムを，学習時と異なる攻撃手法やノイズレベルに対してよりロバストにできることを述べた．テストと同様の攻撃手法で検出器を学習した場合には，提案手法 2(Concat) が他手法の中で最も高い防御性能を示した．特に，FGSM で生成された弱いノイズレベルの敵対的事例を用いて Concat を学習した場合，単独で用いた検出器やロバスト分類器に加えて，両者を組み合わせた従来法よりも攻撃手法やノイズレベルに対して優れたロバスト性を示した．

文　　献

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 長我部恭行，エイプリルピョンマウンマウン，貴家仁志，“敵対的事例の検出器を用いた画像分類システムの防御性能の向上，” マルチメディア情報ハイディング・エンリッチメント研究会，第 IEICE-121 巻，pp.1–6，2022.

[6] 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.

[7] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” IEEE Symposium on Security and Privacy, pp.39–57, 2017.

[8] 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, pp.••–••, Apr. 2018.

[9] J. Uesato, B. O’Donoghue, P. Kohli, and A. van denOord, “Adversarial risk and the dangers of evaluating against weak attacks,” Proc. of the 35th International Conference on Machine Learning, vol.80, pp.5025–5034, 10–15 Jul 2018.

[10] Y. Li, L. Li, L. Wang, T. Zhang, and B. Gong, “NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks,” Proc. of International Conference on Machine Learning, vol.97, pp.3866–3876, 09–15 Jun 2019.

[11] J. Su, D.V. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” IEEE Transactions on Evolutionary Computation, vol.23, no.5, pp.828–841, 2019.

[12] 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.

[13] 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.

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, June 2016.