6S-09

複数モデルの出力を用いたAdversarial Examplesの検出

林 瑛晟[1,a)]　服部 隆志[1]　萩野 達也[1]

概要：Convolutional Neural Network をはじめとするニューラルネットワークは幅広く成果を挙げてきた一方で，Adversarial Examples と呼ばれる攻撃ノイズの付与に対して脆弱であることが報告されている．攻撃に対する様々な対応策が提案されたが，多くの場合対策を破る方法が考案されるイタチごっこの状態となっている．以上を踏まえた上で，本研究では動的な Adversarial Examples 検出器の生成手法を提案する．生成される検出器は最大で 90%を超える精度と未知のモデルへの未知の攻撃の検出を達成した．

1. はじめに

Szegedy ら [1] によって，正しく分類される入力に僅かな差分を加えると，ニューラルネットワークを含むいくつかの機械学習モデルが誤分類を起こす現象が報告された．これらは Adversarial Examples と名付けられた．攻撃手法には誤差関数の勾配を利用する Gradient-based Attacks の他に，出力のみを利用する Black-Box Attack も存在する．Black-box Attacks には Score-based Attacks と Decision-Based Attacks が含まれる [2], [3].

Adversarial Examples への対策は，防御と検出の 2 種類がある．実社会で悪意ある攻撃が仕掛けられた場合，対処が必要であるから，防御と同様に検出も重要となる．そのため，本研究では検出を対象とする．

2. 関連研究

Adversarial Examples の対策は画像を直接用いる手法や CNN の内部を利用する方法，GAN を利用した方法などが提案されたが，多くは突破可能なことが示されている [5]. 自動運転への攻撃に関しては画像に角度などが加わるため問題ないとの主張 [6] があったが，後にそれらへ robust な手法が提案された [7]．アンサンブルで攻撃に強いモデルを作成する手法 [8] も存在する．本研究もアンサンブルを用いていると言える．特に，Monteiro らの研究 [9] は本研究に近いアプローチを取っている．

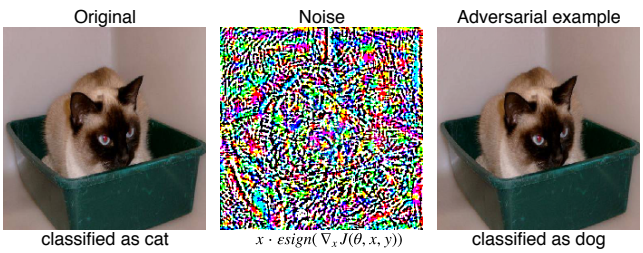

図 1 Fast Gradient Sigh Method(FGSM)[4] で作成した adversarial example．ノイズはスケーリングしてある．

表 1 各モデルのクロスエントロピー増加量．VGG16 に対して FGSM で生成した Adversarial Examples を用いて計測した．精度が 80%を下回ったモデルは学習に失敗したとして除外．

	Fine-tuning	転移学習	通常の学習
DenseNet121	0.00203334	0.183558	0.0163774
InceptionResNet	0.0199543	0.0584078	-0.0313711
InceptionV3	0.039255	-0.00419998	-0.026609
MobileNet	0.347232	0.140722	0.0167903
NASNetMobile	0.188956	0.0630474	NaN
ResNet50	0.0130675	NaN	0.0674035
VGG16	1.4866	0.636695	NaN
VGG19	0.119011	0.130622	NaN
Xception	0.00997292	0.317626	0.00542898

3. 提案手法

3.1 仮説

Szegedy らの研究では，データセットの異なるサブセットで学習した異なるアーキテクチャーのモデルが同一の adversarial example を誤分類することとが報告された [1]. いくつかのアーキテクチャーを用いて検証を行った結果，その程度は一様ではないことが分かった (**表 1**)．この結果をもとに，誤分類の起こる度合いを特徴量として用いることで Adversarial Examples を検出できると仮説を立て，実験を行った．

3.2 データセット

実験には Dogs vs. Cats データセットを用いた．このデータセットは CAPTCHA 用に作成された犬と猫の画像データセット [10] から，画像分類コンペティション用に 25000 枚を抜き出したもの[*1]である．

[1] 慶應義塾大学 環境情報学部
Faculty of Environment and Information Studies, Keio Uniersity
[a)] happyteru@keio.jp

[*1] https://www.kaggle.com/c/dogs-vs-cats/

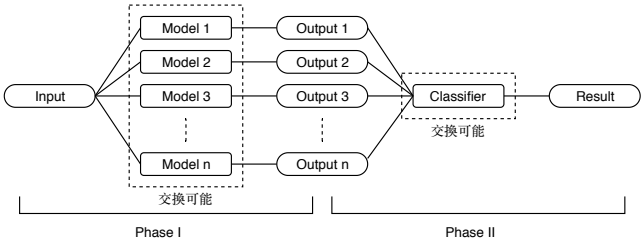

図 2　提案手法の概略図

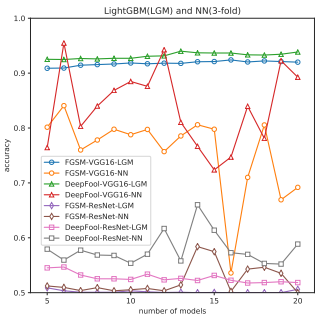

図 3　攻撃とモデルごとの精度

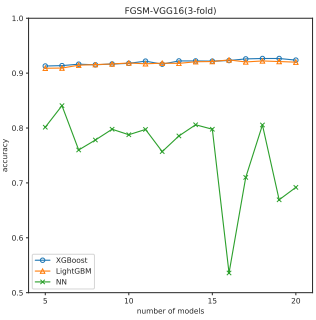

図 4　分類器ごとの精度

3.3　モデル

図 2 に手法の概要を示す．Phase I では ImageNet で事前学習した重み[*2]を用いて学習し，表 1 のように 27 種類のモデルを作成した．精度が 80%を下回った 4 モデルは除外した．学習には計 20000 枚の画像を用いた．

Phase II の前準備として FGSM を用いて VGG16 の Adversarial Examples を犬と猫のそれぞれ 1500 枚ずつ用意した．それらを Phase I の入力とし，出力を XGBoost, LightGBM, 全結合 NN の分類器 3 種類で学習した．

VGG16 の Adversarial Examples で学習した分類機を用いて，さらに別の攻撃手法として DeepFool Attack[11] や別のモデルとして ResNet50 に対しても評価を行った．

3.4　結果

最大で 90%超える正解率を達成した．また，攻撃手法やモデルの情報が学習データに含まれているかによって分類の難易度が大きく変化する (**図 3**)．特に攻撃対象のモデルの有無が結果に与える影響が大きい．全結合 NN がやや不安定であるものの，分類器の違いは精度に大きな影響を及ぼさない (**図 4**)．これは説明変数が高々 $num\,of\,classes \times num\,of\,models$ のタスクに帰着できているからだと考えられる．また，Phase II のみを行い検出器を生成する場合は Phase I も行う場合と比較して高速である．

3.5　既存手法との比較・結論

既知のモデルと攻撃に比べ難易度が上がるため精度が落ちるものの，画像の分類結果を元に未知のモデルへの未知の攻撃を検出できたことは重要かつ興味深い結果である．この成果を元に Black-box 検出器が実現できる．現実のタスクではモデルは既知であっても攻撃手法が既知であるとは限らない．そのため，攻撃手法と同様に検知手法もどれだけの情報を利用するかに応じて**表 2** のように分類することを提案する．また，実験のように 23 個のモデルから 10 個を選び 3 種類の分類機を使う場合 ${}_{23}C_{10} \times 3$ 通りの検出器が得られる．多数の検出器を同時にバイパスすることは困難であるから，提案手法は強固である．

表 2　利用する情報に基づいた分類 (今回の実験における対応)

	攻撃が既知	攻撃が未知/
モデルが既知	level1(FGSM-VGG16)	level2(DeepFool-VGG16)
モデルが未知	level3(FGSM-ResNet)	level4(DeepFool-ResNet)

[*2] https://keras.io/ja/applications/

4.　今後の展望

よりクラス数の大きなデータセットの利用や学習方法の工夫で，精度と安定性を改善することが課題である．図 3 からは，未知の攻撃のほうが検出精度が高いという直感に反する結果が示された．Phase I で学習の精度に差が出ていることや Phase II で過学習を起こしている可能性を考慮すべきである。また，実験結果からアークテクチャーの近いモデル同士で Adversarial Examples に対しより脆弱である傾向が示唆された．この傾向を厳密に議論するために，モデル間の類似度や距離を示す定量的な指標を定義することができれば，研究が一段進むと考えられる．

参考文献

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R.: Intriguing properties of neural networks, *arXiv:1312.6199* (2013).
[2] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. and Mukhopadhyay, D.: Adversarial Attacks and Defences: A Survey, *arXiv:1810.00069* (2018).
[3] Brendel, W., Rauber, J. and Bethge, M.: Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, *arXiv:1712.04248* (2017).
[4] Goodfellow, I. J., Shlens, J. and Szegedy, C.: Explaining and harnessing adversarial examples. CoRR (2015).
[5] Carlini, N. and Wagner, D.: Adversarial examples are not easily detected: Bypassing ten detection methods, *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, ACM, pp. 3–14 (2017).
[6] Lu, J., Sibai, H., Fabry, E. and Forsyth, D.: No need to worry about adversarial examples in object detection in autonomous vehicles, *arXiv:1707.03501* (2017).
[7] Athalye, A. and Sutskever, I.: Synthesizing robust adversarial examples, *arXiv:1707.07397* (2017).
[8] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D. and McDaniel, P.: Ensemble adversarial training: Attacks and defenses, *arXiv:1705.07204* (2017).
[9] Monteiro, J., Akhtar, Z. and Falk, T. H.: Generalizable Adversarial Examples Detection Based on Bi-model Decision Mismatch, *arXiv:1802.07770* (2018).
[10] Elson, J., Douceur, J. J., Howell, J. and Saul, J.: Asirra: a CAPTCHA that exploits interest-aligned manual image categorization (2007).
[11] Moosavi-Dezfooli, S.-M., Fawzi, A. and Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582 (2016).