ピクセル偏向法對抗敵対攻撃の評価

覃永智 [1,a)]　張海波 [1]　櫻井幸一 [1]

概要：Prakash はピクセル偏向という方法を提案した．敵対攻撃の摂動は自然な画像のような目立つオブジェクトがあることではなく，画像の隅々まで存在することが多い．活性化関数(activation map)により，CNN 分類にとって目立つところを見つける．その後，画像の目立たない部分にピクセル偏向をすると，攻撃の摂動を破壊できる同時に，きれいな画像の重要な部分も守れる．本稿では，Prakash が提案した方法の有効性を評価した．

キーワード：機械学習，分類学習，ニューラルネット

Evaluating the Pixel Deflection Defending against Adversarial Attacks

YONGZHI QIN[1,a)]　HAIBO ZHANH [1]
KOUICHI SAKURAI[1]

Abstract: Prakash proposed a method called pixel deflection. Perturbations of hostile attacks often exist in every corner of the image, rather than having a prominent object like a natural image. The activation map finds prominence in the CNN classification. Pixel biasing to the inconspicuous parts of the image can then destroy the perturbations of the attack while at the same time protecting the important parts of the clean image. In this paper, we evaluated the effectiveness of the method proposed in the paper.

Keywords: Machine learning, Classification learning, Neural network

1. 研究背景

最近畳み込みニューラルネットワーク（Convolutional neural network，略称: CNN)技術の成熟によって，人々と生活でよく使われている．しかし，CNN の普及とともに，「敵対攻撃」の脅威も酷くなる．

「敵対攻撃」の対策としていろいろな方法も提案されている．Prakash はピクセル偏向という方法を提案した．その方法の中で画像のピクセルと近所のピクセルを交換して，敵対攻撃の摂動を破壊することができる．しかし，実際に Prakash の方法を使う時予想内の結果をもらわなかった．Athalye[5]は Prakash の研究をホワイトボックスで検証をし，効果が良くないことを証明した．Li[6]の論文では，は 13 の防御をテストしたが，Prakash の方法は一番弱いことがわかった．論文[7,8]で Prakash の方法は画像の分類を誤りにするの可能性があると提唱した．論文[9,10]で Prakash のような方法は只敵対攻撃の摂動が小さいときが効くという判断がある．Li の論文[11]では、Prakash の方法でのピクセル偏向部分が作用が弱いと判断したが、本稿では，Prakash の方法のピクセル偏向は全然効果がないのではないが、ピクセル偏向とウェーブレットノイズ除去両方を同時に使う時の効果がただのウェーブレットノイズ除去よりいいではないという結果を発現した。

1 九州大学
Motooka 774, Nishi-ku, Fukuoka 819-0395, Japan
a) qin.yongzhi.485@s.kyushu-u.ac.jp

2. 紹介

2.1 敵対攻撃

敵対攻撃とは，画像に人の目にはあまり区別が見えない摂動を加え，CNN クラシファイアは攻撃された画像に騙され，間違い結果を出力することである．普通の対応方法はノイズを画像に加えるや画像を圧縮するなどの方法である．他には，GAN を使い，元の画像を生成という考えもあるが，どちらの方法も攻撃されていない画像に使うと分類確信度が減る問題がある．

2.1.1 Fast Gradient Sign Method(FGSM)

FGSM[2]は敵対攻撃の中で一番見やすいものである．FGSM の方法は以下の式の通りである．

$$\hat{\boldsymbol{x}} = \boldsymbol{x} + \epsilon sign(\nabla_x J(\boldsymbol{\theta}, \boldsymbol{x}, y))$$

$\boldsymbol{x}$は入力画像，$\hat{\boldsymbol{x}}$は敵対攻撃の結果，$\boldsymbol{\theta}$はモデルのパラメータ，yはターゲット，$J(\boldsymbol{\theta}, \boldsymbol{x}, y)$モデルのコスト，$\epsilon$はハイパーパラメータ．

2.2 ピクセル偏向(略称:PD)

敵対摂動にノイズを加え，敵対摂動を破壊するのは可能である．筆者はピクセル偏向という方向を提案した．画像からピクセルをランダムにサンプリングし，小さな正方形の近傍内からランダムに選択された別のピクセルに置き換える．

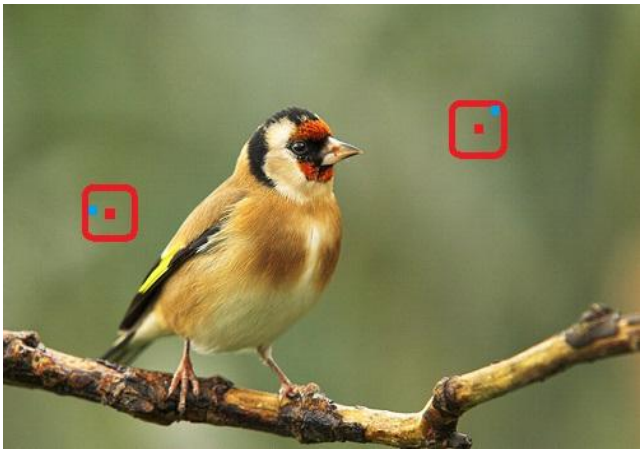

図 1　ピクセル偏向

2.3 活性化マップ

活性化マップ[3]とは，CNN クラシファイアは分類を判断する時，画像の各部分の重要性を描く図である．

CNN クラシファイアの FC 層からある分類に各特徴の w を重さにする．FC 層の各特徴は元の画像に投影できる．各投影と重さの掛け算の足し算は活性化マップである．

自然な画像の活性化マップは主な目標に集まり，敵対摂動はその特徴がなく，画像全体に分散している．

Prakash の論文の中で，筆者はきれいな画像の分類の正しさを保つめに，活性化マップを利用して，元の画像の情報をできるだけ守る上，ピクセル偏向の方法で敵対摂動を破壊する．

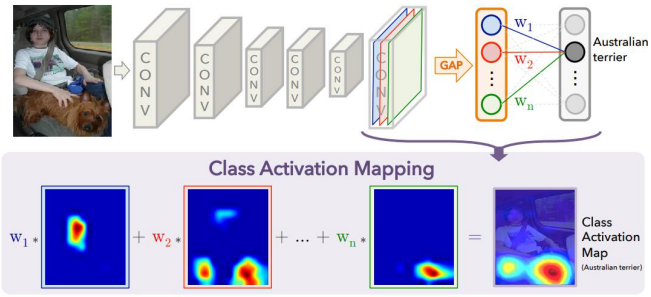

図 2　活性化マップ[3]

2.4 ウェーブレットノイズ除去(略称:WD)

画像のウェーブレット変換は画像圧縮やノイズ除去などでよく使われる方法．

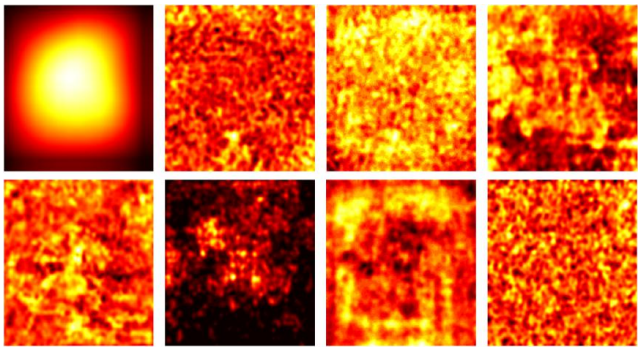

図 3　左上 1 は自然な画像の活性化マップ，他のは敵対摂動の分布[1]

3. 研究内容

Tensorflow の訓練完了のモデル Resnet50 を利用した．Imagenet の訓練データから各種類で 5 枚の画像をランダムに選んでデーターセットにした．敵対攻撃の画像を生成するために cleverhans を利用して，FGSM の画像を生成した．

その中でハイパーパラメータは以下の通り：ϵはウェーブレットノイズ除去のパラメータ：$\epsilon = 0.04$，Window はピクセル偏向の小さな正方形の大きさ：window $= 10$，L_{inf}は敵対攻撃のノルム：$L_{inf} = 12.8$.

そして，元の画像，FGSM 画像，PD(ここからの PD は活性化マップを利用した PD)後の FGSM 画像, WD 後の FGSM 画像，この方法を全部使う元画像と FGSM 画像の分類正しい確率を測定した．その結果は以下表 1 の通りである．

表 1　ピクセル偏向のパフォーマンス

Table 1　Pixel deflection performance.

deflection	N	T	PD	WD	defense	clean
100	3906	1117	1118	1220	1220	3696
250	3906	1114	1112	1223	1223	3683
500	3904	1099	1100	1216	1216	3667
1000	3906	1074	1102	1193	1193	3648
2000	3906	1032	1061	1143	1143	3599
4000	3906	996	1033	1117	1117	3511
8000	3906	862	927	972	972	3348

表 1 の中で Deflection は PD の中でピクセルを偏向する数，

N は元の画像が正しく分類できる数，T は FGSM の画像が正しく分類できる数，PD は FGSM の画像を PD した後正しく分類できる数，WD はは FGSM の画像を WD した後正しく分類できる数，Defense は FGSM の画像を PD して WD をした後正しく分類できる数, clean は元の画像を PD して WD をした後正しく分類できる数である．

4. 結論

データーから見るとピクセル偏向は FGSM 攻撃の防御において，効果がないとは言えないが，いい効果があるでもない．Deflection の数の増加によって，単純に PD を使う方法は何の処理もないよりいい結果があるが，PD と WD 両方を使う方法は WD よりいい効果がない．だから、FGSM 攻撃に対応する時にはPrakashの方法よりウェーブレットノイズ除去だけの方がいい。

参考文献

[1] Prakash, A., Moran, N., Garber, S., DiLillo, A., & Storer, J. (2018). Deflecting adversarial attacks with pixel deflection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8571-8580)
[2] I.J.Goodfellow,J.Shlens,andC.Szegedy,“Explaining and harnessing adversarial examples,”arXiv preprint arXiv:1412.6572,2014.
[3] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2921-2929).
[4] Chang, S. G., Yu, B., & Vetterli, M. (2000). Adaptive wavelet thresholding for image denoising and compression. IEEE transactions on image processing, 9(9), 1532-1546.
[5] Athalye, A., & Carlini, N. (2018). On the robustness of the cvpr 2018 white-box adversarial example defenses. arXiv preprint arXiv:1804.03286.
[6] Li, Y., Li, L., Wang, L., Zhang, T., & Gong, B. (2019, May). Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In International Conference on Machine Learning (pp. 3866-3876). PMLR.
[7] Mustafa, A., Khan, S. H., Hayat, M., Shen, J., & Shao, L. (2019). Image super-resolution as a defense against adversarial attacks. IEEE Transactions on Image Processing, 29, 1711-1724.
[8] Agarwal, A., Singh, R., Vatsa, M., & Ratha, N. (2020). Image Transformation-Based Defense Against Adversarial Perturbation on Deep Learning Models. IEEE Transactions on Dependable and Secure Computing, 18(5), 2106-2121.
[9] Borkar, T. S., & Karam, L. J. (2019). DeepCorrect: Correcting DNN models against image distortions. IEEE Transactions on Image Processing, 28(12), 6022-6034.
[10] Sun, B., Tsai, N. H., Liu, F., Yu, R., & Su, H. (2019). Adversarial defense by stratified convolutional sparse coding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11447-11456).
[11]Li,R., Feng,Y., Sakurai,K.,(2020).Evaluating the Effectiveness of Pixel Deflection Defending against Adversarial Attacks.IPSJ SIG Technical Report from https://www.ipsj-kyushu.jp/page/ronbun/hinokuni/1009/Papers/A4-1.pdf