H-013

LPF に耐性を持つ Adversarial Example の生成法
Adversarial Example Generation Resistant to LPF

小和田雄太† 菅間幸司† 和田俊和†1
Yuta Kowada Koji Kamma Tosikazu Wada

1. はじめに

近年,画像認識分野では Deep Neural Network(DNN)による認識技術が目覚ましい進歩をとげている.その一方で,入力データを加工することで,DNN の認識を意図的に誤らせる Adversarial Example(AE)という攻撃が生まれている.AE は人間の目には殆ど見えない微小な Adversarial Noise(AN)を入力画像に加えることで,学習済みの DNN に誤った予測をさせる(図 1 参照).画像の見た目を変えることなく,標的の画像識別システムを誤識別させることができれば,そのシステムの信頼性は大きく損なわれる.また,AE による攻撃が可能であることは,人が判断の根拠としているのとは異なる特徴に着目して画像識別システムが動作している証拠であり,システムに対して人間と同等の信頼をおけないことになってしまう.これらの結果,高い識別性能を持つシステムであっても,社会実装が行えなくなる.

このような問題に対して,誤識別を防ぐ Adversarial Defense(AD) [8][9]や AE を適用したデータを用いて学習を行う Adversarial Training(AT)[3][5]などの研究が行われている.

AD の一つに LPF を適用した画像を識別するという方法がある.これは,AE を作成する際に加える AN は高周波数成分を多く含む[2]ため,これを除去・抑制する方法である.しかし,LPF の影響を受けない AE を作成することができれば,新たな AD や AT が必要になる.

本論文では,画像の特徴的な周波数成分を変化させることで,固定的な周波数フィルタでの防御が不可能な AE の生成を提案する.

2. 関連研究

2.1 勾配を利用した AE

Fast Gradient Sign Method(FGSM)[1]は,Goodfellow らによって提案された損失関数の勾配を利用して AE を作成する手法である.FGSM では,識別器のパラメータをθ,入力をx,出力をy, ニューラルネットワークの学習に使用する損失関数を$J(\theta,x,y)$とすると加える AN であるηは式(1)のように表される.

$$\eta = \epsilon \mathrm{sign}(\nabla_x J(\theta,x,y)) \quad (1)$$

このようにして作成されたηを入力xに加えることで,識別器に認識を誤らせる AE を作成することができる.

Kurakin らは,FGSM を改良し,Basic Iterative Method(BIM)を提案した.BIM は, FGSM を反復的に行う手法である.

上記の BIM は元の入力画像で初期化するが,初期画像としてランダムノイズを用い,BIM と同じ方法で AE を作成する PGD[5]という手法も提案されている.

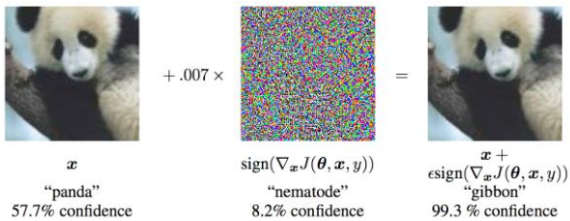

図 1 AE の例([1]より抜粋) ：識別器に 57.7%でパンダと認識される画像左に,微小な AN(中央)を加えた,右の画像は 99.3%でテナガザルと誤認識される.

2.2 画像周波数に着目した AE の防御に関する研究

図 2 に示すように入力画像では低周波成分が多くの割合を占めているのに対し,AE を作成する際に加えられる AN は高周波成分も多く含まれている.この二つの空間周波数での分布の違いに着目して,識別器を画像空間ではなく,空間周波数領域で学習させることにより,AE に対して頑強な識別モデルの作成に成功している.

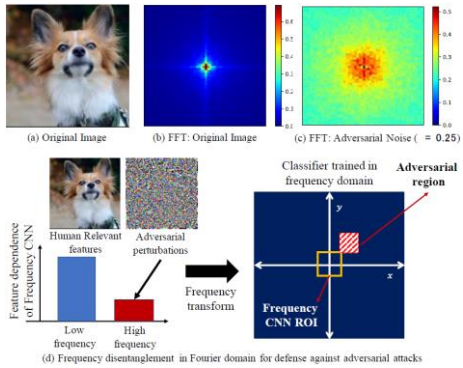

図 2：入力画像とノイズの周波数分布([2]より抜粋)

2.3 空間周波数領域で生成する AE

2.3.1 情報を削除して生成する AE

通常,元画像にノイズを付加させることにより AE を生成するが Duan らは[6],空間周波数領域において,元画像の情報を削除することにより AE を生成し,AT や JPEG 圧縮などの防御をした際に,PGD や FGSM などの他の攻撃手法よりも高い攻撃の成功率を成功している.

2.3.2 低周波領域で生成する AE

ブラックボックスモデルを用いた攻撃において Guo ら [7]は,AN をランダムにサンプリングする際,低周波領域に限定してサンプリングして AE を作成することを提案しており,

† 和歌山大学大学院システム工学研究科

限定しない場合と比べて,AE を生成する際のクエリコストを 2〜4 倍削減し,画像変換による防御を回避することに成功している.

3. 提案手法

3.1 アイデア

上記の関連研究[2][6][7]から,周波数領域で AN を生成することは有効であると言える.但し，常に同じ周波数領域で AN を生成する場合には，その領域を抑制・遮断してしまえば容易に防御することができてしまう．そこで本研究では次のように個々の画像ごとに異なる周波数特性を持った AN を生成し，単一の周波数フィルタでは防御することができない AE を生成する手法を提案する.

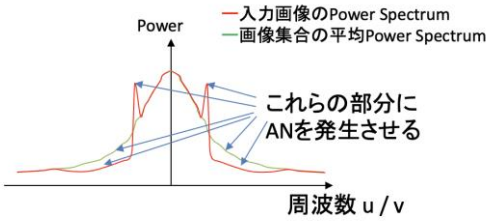

図 3:提案手法の概念図：画像集合の平均 PS から乖離した特徴的な部分に AN を発生させる.

図 3 は，入力として与える画像集合の平均 Power Spectrum(PS)と，ある入力画像の PS を周波数領域で表した図である．この図において，入力画像と平均画像の PS に乖離がある部分は，「入力画像に固有の特徴的な周波数領域」と見なすことができ，その部分に重点的に AN を生成すれば,誤識別が起きやすくなると考えられる．すなわち，画像集合の平均 PS とある入力画像の PS の差に比例して AN を発生させる．これにより，個々の画像に応じて異なる周波数特性を持った AN を発生させることができ，より効率的に誤識別を起こさせることができるはずである.

3.2 画像に対するフーリエ変換

画像を 2 変数関数$f(x,y)$として表すとき,フーリエ変換$F(u,v)$は以下のように定義される.

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j(ux+vy)}dxdy \qquad (2)$$

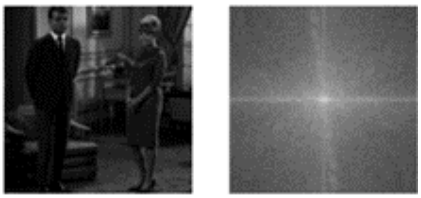

図 4：画像とそのパワースペクトル

図 4 左の画像の PS すなわち，$F^*(u,v)F(u,v)$を同図右に示す．右図では,中心に近いほど低い周波数,中心から遠いほど高い周波数を表しており，白い部分ほど成分が強いことを示している

3.3 LPF

今回の実験では,想定する防御手法として Low Pass Filter (LPF)を使用するが，LPF とは画像の低周波数成分だけ残し,

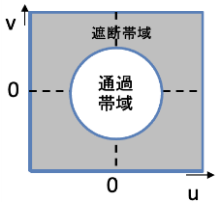

図 5:本研究で用いる LPF

高周波成分を除去するフィルタである．3.2 で述べた通り，周波数領域の表現では中心に近いほど低い周波数,中心から遠いほど高い周波数を含んでいる.したがって，図 5 のように中心部分の円形領域を取り出すフィルタを作成すれば LPF になる.

3.4 適応的フィルタを用いた AN の生成

ここでは，LPF のように，画像に関係なく固定的な周波数帯域の成分を通過・遮断するフィルタではなく，入力に応じて特性が変化する適応的フィルタの具体的な作成方法について説明する.まず識別器が正しく識別したi番目の画像に対して,フーリエ変換を適用して求めた PS を$F_i(u,v)$と表す.各(u,v)について,入力として想定される画像集合$i = 1,\dots,n$の平均$\frac{1}{n}\sum_{i=1}^{n} F_i(u,v)$を$\hat{F}(u,v)$と表す．次に AE 生成に使用する画像の PS を$F_j(u,v)$とした場合，式(3)のように,$\hat{F}(u,v)$と$F_j(u,v)$の差の絶対値を$D_j(u,v)$とする.

$$D_j(u,v) = \left|\hat{F}(u,v) - F_j(u,v)\right| \qquad (3)$$

式(3)を用いて,各(u,v)における差の大きさに比例し,一定の大きさの AN を通過させるフィルタ H_jを式(4)のように定義する.

$$H_j(u,v) \equiv \frac{D_j(u,v)}{\hat{D}_j}, \qquad \hat{D}_j = \iint D_j(u,v)dudv \qquad (4)$$

式(4)の分母は，フィルタで通過させる AN の大きさを正規化するためのスケーリングを表している．さらにH_jの一部の要素の値が大きくなりすぎる場合があるので,一定以上の値にならないように,式(5)を適用する.

$$H_j'(u,v) = \beta\left(1 - \exp\left(-\frac{1}{\gamma}H_j(u,v)^2\right)\right) \qquad (5)$$

ここでβは振幅を調整するハイパーパラメータである.また,γはH_jが持つ要素の値の分布を調整するハイパーパラメータである.実験では，$\beta = 1.3, \gamma = 0.1$の値を使用している.

3.5 適応的フィルタを利用した AE 作成の流れ

本節では,まず BIM による AE の作成手順を説明し,続いて提案手法について説明する.

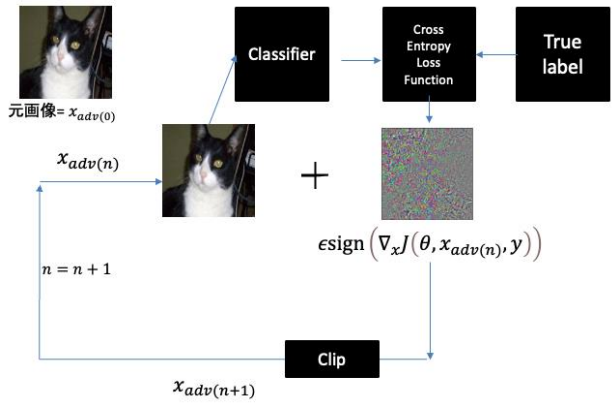

図 6：BIM のブロックダイアグラム

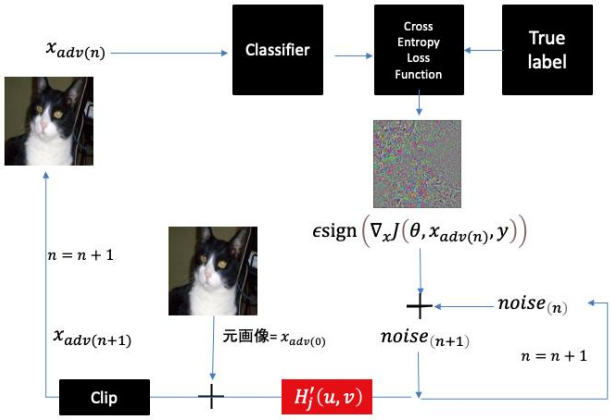

図 7：提案手法のブロックダイアグラム

BIM では，各サイクルにおいて，式(6)のようにして反復的に画像にノイズを加える．

$$x_{adv(n+1)} = Clip\left(x_{adv(n)} + \epsilon \text{sign}\left(\nabla_x J(\theta, x_{adv(n)}, y)\right)\right), \quad (6)$$

但し，x'_{adv}は AE，yは正解ラベル，ϵはノイズの大きさを決定するスケーリング，θは DNN モデルのパラメータ，Jは損失関数，sign は符号関数である．Clip は AE の各画素値が 0 以上 1 以下の範囲を逸脱しないように，クリップする関数である．実験で使用したパラメータは，$\epsilon = \frac{1}{255}$,反復回数は 10 回である．図 6 に BIM のブロックダイアグラムを示す．

一方，提案手法では，適応的フィルタ H'_jを損失の勾配から生成した AN に適用し，AN を蓄積するため，ブロックダイアグラムは図 7 のようになる．またアルゴリズムを Algorithm1 に示す．

4. 評価実験

4. 1 実験設定

データセット：本実験においてはデータセットとして，200 クラスの画像データセットである tiny-imagenet を用いた．そのうち，100，000 枚の学習用データから無作為に選択した 96，000 枚の画像を識別器の学習に用いた．学習率の初期値を 0．1 とし，30 エポックごとに学習率を 0．1 倍にして，90 エポックで学習を打ち切った．学習済みの識別器を用いて 10，000 枚のテストデータに対する推論を行ったところ，正解率は 56．01%であった．なお識別器には，入力層と出力層を tiny-imagenet 用に調整した Resnet18 を利用した．

Algorithm1：提案手法

Input：パラメータθ，損失関数J，入力x，
正解ラベルy，ステップサイズϵ，反復回数 t，

$x_{adv(0)} = x$
$noise_{(0)} = 0$
for *t* **do**
$\eta = \epsilon \text{sign}\left(\nabla_x J(\theta, x_{adv(n)}, y)\right)$
$noise_{(n+1)} = \eta + noise_{(n)}$
$noise_{(n+1)} = H' * noise_{(n+1)}$ //*は畳み込み積分を表す
$x_{adv(n+1)} = Clip(x + noise_{(n+1)})$
Output：x_{adv}

正しく識別できたテストデータ 5601 枚の画像に対して，BIM と提案手法で AE を作成する実験を行った結果を表 1 に示す．両手法において，一度に乗せるノイズの大きさは，$\epsilon = \frac{1}{255}$とし，元画像と AE の各画素における差の大きさについては，最大$\frac{8}{255}$までを許容し，誤識別率の比較を行った．

表 1 BIM と提案手法の誤識別率の比較

	フィルタなし	LPF18	LPF20	LPF22
元画像	0%	20．50%	9．34%	4．27%
BIM	77．22%	67．20%	70．52%	73．06%
提案手法	80．54%	72．23%	75．04%	77．43%

4. 2 性能評価

生成した AE を識別器に入力した際の誤識別率を求めた結果を表 1 に，図 8，9 に$\hat{F}(u, v)$と$F_j(u, v)$のパワースペクトラム（Power Spectrum，PS），BIM と提案手法で生成した AE とノイズの PS を可視化したものとノイズを加える前の原画像を示す．

表 1 において，各列は識別器に入力する際に適用した LPF とそのサイズ（周波数領域での半径）を示している．各行は，DNN に入力として与えたデータの違いを表しており，原画像，もしくは生成した AE の手法を示している．

表 1 より，BIM よりも提案手法のほうが誤識別率が増えていることが確認できる．また，図 8，9 より，BIM と比較して，提案手法の方が幅広い領域に AN の PS が分布していることがわかる．また，原画像の PS に応じて AN の PS が変化していることも確認できる．この特性から一定サイズの LPF で防御しても誤識別率が低下しにくい原因が説明できる．

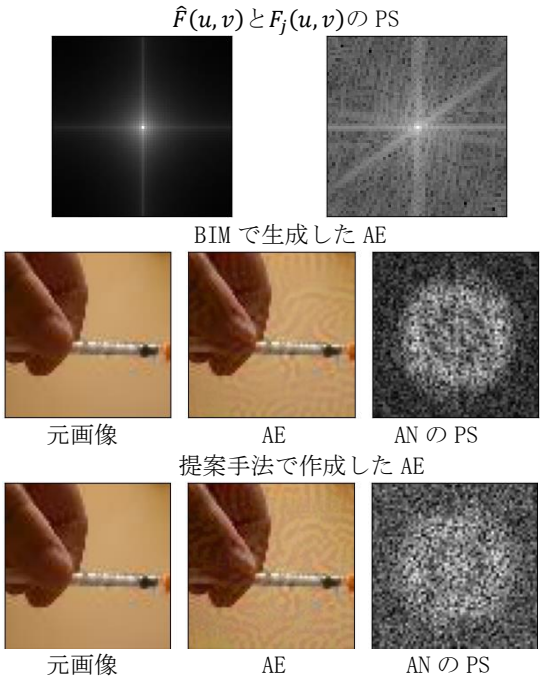

図 8:BIM と提案手法で生成した AE

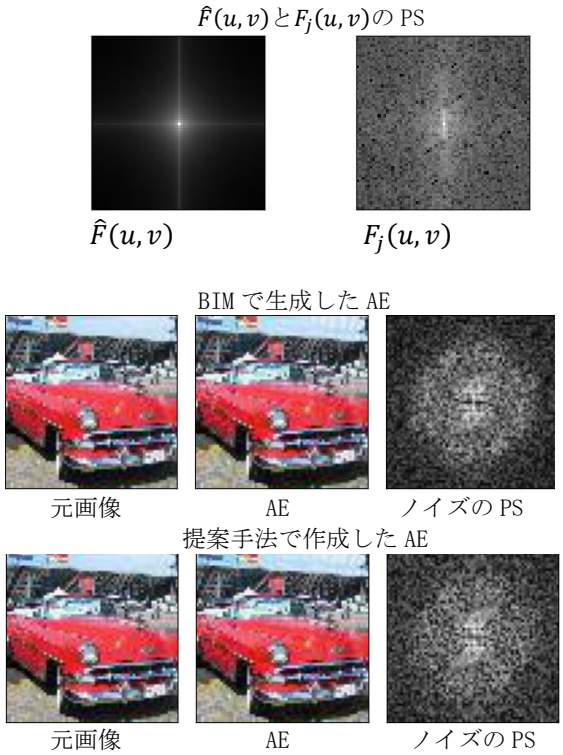

図 9:BIM と提案手法で生成した AE

5. まとめ

本研究では, 周波数領域において, 元画像とテストデータ全体の画像の平均の差を用いてフィルタを作成し, 作成したノイズにそのフィルタを適用することで AE を作成する方法を提案した. 実験結果から, 従来手法として用いた BIM より提案手法のほうが LPF で防御した場合, 誤識別率が上がることを示した. 今後は, 他の防御方法に対して AE の性能がどう変わるか, 他の攻撃手法と組み合わせた場合はどのような結果になるか調査する.

参考文献

[1] I. J.Goodfellow, ,J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” In Proc. of ICLR, 2015
[2] S. Chaudhury and T. Yamasaki, “Towards Adversarial Robustness of Learning in the Frequency Domain,” IEICE Technical Report.2021(3),vol.120,No_PRMU_409,pp.176-180
[3] Q. Z. Cai, M. Du, C. Liu, and D. Song, “Curriculum adversarial training”. arXiv preprint arXiv:1805.04807.2018
[4] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world”. In Artificial intelligence safety and security (pp. 99-112). Chapman and Hall/CRC.2018
[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks”, arXiv preprint arXiv:1706.06083.2017
[6] R. Duan, Y. Chen, D. Niu, Y. Yang, A. K. Qin, and Y. He, ” AdvDrop : Adversarial Attack to DNNs by Dropping Information. In Proceedings of ICCV (pp. 7506-7515,2021).
[7] C. Guo, J. S. Frank, and K. Q. Weinberger, “Low frequency adversarial perturbation”, arXiv preprint arXiv:1809.08758,2018
[8] D. G. Karolina, Z. Ghahramani, and D. M. Roy, “A study of the effect of jpg compression on adversarial images,” arXiv preprint arXiv:1608.00853 2016.
[9] K. Roth, Y. Kilcher, and T. Hofmann. “The odds are odd: A statistical test for detecting adversarial examples,” In Proceedings of ICML PMLR, 2019.