H-031

ノイズ付加学習手法を適用した CNN の Adversarial Examples 耐性の評価
Evaluation of the Resistance to Adversarial Examples of CNN Using Noise Addition Learning Method

野田 遼太郎 † 今井 信太郎 † 武田 敦志 ‡
Ryotaro Noda Shintaro Imai Atsushi Takeda

1 はじめに

近年，畳み込みニューラルネットワーク（Convolutional Neural Network, 以下 CNN）を用いた画像認識に関する研究及び実用化が広く進められている．しかし，CNN を用いた画像認識の実用化にあたり，Adversarial Examples[1] と呼ばれる，計算で求めた摂動を入力に加えることにより高確率で誤認識を発生させる手法が存在し，障害となることが危惧されている．我々の研究グループでは，学習時に複数回ノイズを付加することにより認識精度を向上させる CNN 学習手法 [2] を提案しているが，ノイズを付加することで CNN 中の信号バリエーションが増えるため，Adversarial Examples への耐性も期待できると考えられる．本稿では，ノイズを付加する学習手法の Adversarial Examples 耐性を評価・検証する．

2 関連研究

2.1 Adversarial Examples

Adversarial Exmaples は，画像に人の目では判別できない程度のノイズ（摂動）を加えることで，作為的に分類器の判断を誤らせる手法である．図 1 のように，人の目では判別が困難な摂動であっても分類器を騙すことができる．例えば，図 1 の一番上の元画像を Residual Network で学習した分類器に入力すると Apple と分類されるが，摂動を加えた画像を入力すると Trout と分類される．

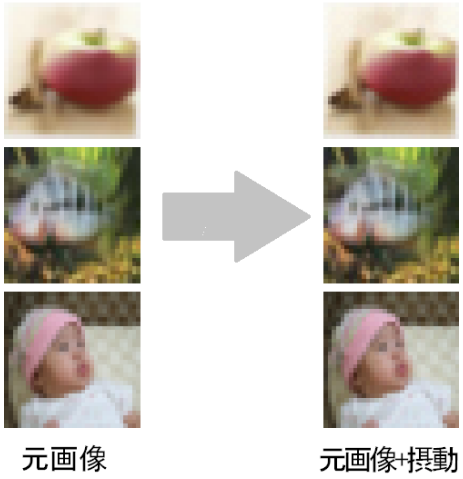

図 1: Adversarial image の例

Adversarial Exmaples の生成方法は様々あるが，一般的に学習済み分類器から計算によって求める．Fast method は，文献 [3] で紹介されている Adversarial Examples 生成手法の一つで，正解方向の勾配と逆方向の摂動を与えることで，分類器に間違った答えを出力させる．この他にも，Fast method を複数回適用する手法である Basic iterative method や，正解方向の勾配と逆方向ではなく分類器が最も低い確率としたクラスに近づける手法である Iterative least-likely class method が紹介されている．

本研究では，Fast method を使用して画像を作成し，Adversarial example 耐性を評価・検証する．

2.2 NoiseNet

NoiseNet[2] とは，我々の研究グループにおいて提案した，CNN の学習時にランダムにノイズを付加することにより学習の停滞を抑えテスト時の認識精度向上を目指す手法である．NoiseNet のネットワーク構成を図 2 および図 3 に示す．

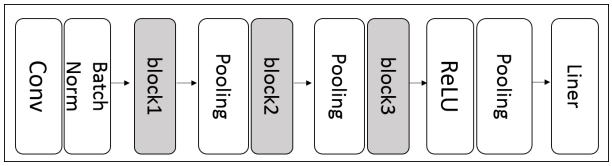

図 2: NoiseNet の構成

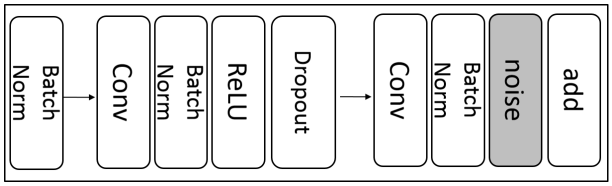

図 3: block の構成

この手法は，現在 CNN で主流となっている ResNet[4] を元にした構成にしており，ResNet における残差ブロックの最後にノイズを付加する層を配置している．ノイズ付加による画像認識精度の向上が評価実験から確認できたが，ノイズを付加することで CNN 中の信号バリエーションが増えるため，Adversarial Examples への耐性も期待できる．そのため，今回の実験により NoiseNet の Adversarial Example 耐性を評価・検証する．

† 岩手県立大学，Iwate Prefectural University
‡ 東北学院大学，Tohoku Gakuin University

3 実験

3.1 実験環境

本実験では，画像認識分野において有名なデータセットの一つである CIFAR-100 を用いる．実験環境を以下に示す．
OS: Ubuntu 18.04.2 LTS
GPU: GeForce GTX 1080 Ti
メモリ:16GB
言語: Python3
ライブラリ: chainer

3.2 SNR

SNR は文献 [5] で用いられている信号とノイズの比率のことで，元画像に対しどれだけの摂動が加えられているかの指標となる．本実験でも文献 [5] と同様に以下の式を用いて SNR を算出している．

$$\mathrm{SNR}(x, \delta_x) = 20 \log 10 \frac{\| x \|_2}{\| \delta_x \|_2}$$

x は元画像，δ_x は摂動にあたり，摂動が大きくなればなるほど SNR の値は小さくなる．本実験では SNR を統一した Adversarial image を用意し，ResNet で学習した分類器と NoiseNet で学習した分類器の精度を比較する．

3.3 比較実験

CIFAR-100 にある 100 ラベルからそれぞれ画像を 1 枚ずつ用意し，Fast method のパラメータを調整することにより，それぞれに SNR20~60 の 10 刻みの摂動を加えたものを生成した．ラベルが Apple の画像について，SNR20~60 の画像を左から並べたものを図 4 に示す．

図 4: SNR ごとの Adversarial image

生成した Adversarial image を ResNet と NoiseNet で学習した分類器にそれぞれ与えた場合の正解数を比較した．双方ともにデータセットは CIFAR-100 で学習したものを使用し，Residual block 数は 20，学習回数は 800 回とした．また，NoiseNet のノイズを含むブロック数は 3 とし，noise の標準偏差は 0.3 とした．実験結果を図 5 に示す．なお，実験で使用した元画像に対する accracy は ResNet で 89，NoiseNet で 96 であった．

3.4 考察

3.3 節の結果から，全体を通して ResNet に比べ NoiseNet で学習した分類器の方が精度が高いことが読み取れる．特に，SNR30,40 での正解率の差が顕著であり，微小な摂動であれば NoiseNet の学習により誤認識のリスクが減少することがわかる．SNR20 のような，大きな摂動が加えられた画像においても ResNet に比べ精度が高い．以上により NoiseNet の Adversarial Examples 耐性は ResNet に比べて高いと言える．

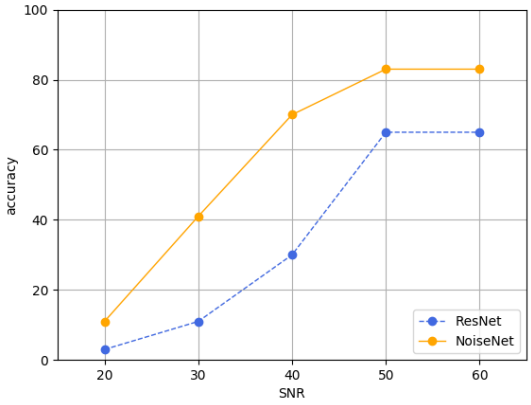

図 5: ResNet 及び NoiseNet で学習した分類器の正解率

4 おわりに

本稿では，100 種の画像を識別することで NoiseNet と ResNet の Adversarial Examples 耐性を比較した．その結果，NoiseNet で学習した分類器のほうが高い精度を示し，ノイズ付加学習手法を適用した CNN の Adversarial Examples 耐性が確認された．

今後の課題として，より多くのサンプルでの実験が挙げられる．本実験では 100 種から 1 枚ずつ取り出して実験を行ったが，データの偏りを考慮すると，より大規模な比較実験が必要と考えられる．また，今回は単純な Adversarial Image の生成手法である Fast method を用いて実験を行ったが，別の手法で生成した Adversarial Image に対しても同様に高い正解率を維持できるかを実験する必要がある．

参考文献

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, “Intriguing properties of neural networks”, arXiv:1312.6199 [cs.CV] (2013).
[2] 野田遼太郎，今井信太郎，武田敦志，“CNN による画像認識精度向上のための付加ノイズの検討”, FIT2018(第 17 回情報科学技術フォーラム) 講演論文集, pp. 3-139 – 3-140 (2018).
[3] A. Kurakin, I. Goodfellow, S. Bengio, “Adversarial examples in the physical world”, arXiv:1607.02533 [cs.CV] (2016).
[4] K. He, X. Zhang, S. Ren, J. Sun, “Deep Residual Learning for Image Recognition”, arXiv:1512.03385 (2015).
[5] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, N. Usunier, “Parseval Networks: Improving Robustness to Adversarial Examples”, arXiv:1704.08847 [stat.ML] (2017).