論文 / 著書情報
Article / Book Information

論題	敵対的学習とノイズ付加を用いた深層学習の頑健性の向上
著者	中村 歩, 宇都 有昭, 篠田 浩一
出典	第27回 画像センシングシンポジウム, , ,
発行日	2021, 6
Note	第27回 画像センシングシンポジウムにて発表

敵対的学習とノイズ付加を用いた深層学習の頑健性の向上

中村　歩 †　宇都　有昭 †　篠田　浩一 †
† 東京工業大学
E-mail: nakamura@ks.c.titech.ac.jp

1　背景・目的

深層学習を用いて訓練したモデルは画像認識などの領域で高い性能が得られる。しかし、そのようなモデルに対して攻撃者が意図的に誤認識を引き起こすことができる脆弱性が存在することが判明している [1]。その脆弱性を利用してモデルを誤認識させ、被害をもたらす攻撃を敵対的攻撃と呼ぶ。敵対的攻撃は深層学習モデルが実社会に応用される際に大きな脅威となり得るため、敵対的攻撃に対して耐性のある深層学習手法の開発は重要な研究課題である。敵対的攻撃は深層学習モデルの誤差逆伝搬を用いて損失が大きくなるように入力を摂動させることで実行する。

敵対的攻撃に対する耐性を向上させる手法はいくつか提案されているが、最も頑健なものは学習時にモデルに対して敵対的攻撃を行い正しい出力が得られるように矯正する手法である [3]。この手法を敵対的学習と呼ぶ。敵対的学習は大きく精度を落とす強力な敵対的攻撃手法を用いれば頑健なモデルを学習できるが、強力な敵対的攻撃手法は誤差逆伝搬を複数回計算するため大きい計算量が必要となる。そのため敵対的学習に必要な計算量も多くなるという欠点が存在する。

もう一つの頑健性を向上させる手法として学習時の入力にランダムノイズを加えることで敵対的攻撃に対してある程度頑健なモデルが学習できる [4]。ランダムノイズの計算は少ない計算量で済むためこの手法は敵対的学習と比べると必要な計算量が少ないという利点があるが、敵対的学習ほど高い頑健性が得られないという欠点が存在する。

このように敵対的攻撃に対して頑健なモデルを学習する手法には得られる頑健性と必要な計算量のトレードオフが存在する。

本研究では少ない計算量で敵対的攻撃に対して頑健なモデルを学習するため、一度のみ誤差逆伝搬してその勾配の方向が出やすくなるように重みをつけたランダムノイズを学習時の入力に加える手法を提案する。この手法を敵対的勾配付きノイズ付与と呼ぶ。この手法は誤差逆伝搬の回数を一度のみにすることで敵対的学習と比べて少ない計算量で頑健な深層学習モデルを学習できる。

第 2 章では敵対的攻撃手法や敵対的攻撃に対する防御手法の詳細について述べる。第 3 章では提案手法の概要とアルゴリズムについて述べる。第 4 章では実験の詳細を記述し、第 5 章ではその実験結果を示して考察する。第 6 章では本研究を総括する。

2　従来手法

2.1　敵対的攻撃

深層学習は高い性能を持つモデルを学習できるが、そのようなモデルに対して攻撃者が意図的に誤作動を誘導する入力を作成できる [1]。この意図的に誤作動するように作成された入力を敵対的サンプルと呼び、敵対的サンプルを用いて深層学習モデルを誤作動させる攻撃を敵対的攻撃と呼ぶ。敵対的攻撃は深層学習が実社会に応用される際に深刻な被害をもたらす恐れがある。本研究では画像分類モデルにおける敵対的サンプルについて議論する。

敵対的サンプルを作成する様々な手法があるが、どの手法も深層学習モデルの誤差逆伝搬を用いる [2] [3]。誤差逆伝搬によって損失を大きくする勾配を計算して、その勾配を入力に加えることで敵対的サンプルを作成する。高い確率でモデルを誤認識させる強力な敵対的サンプルを作るためには誤差逆伝搬を何度も計算する必要があるため、そのような手法は大きい計算量を要する。少ない計算量で敵対的サンプルを作成する手法は誤差逆伝搬の回数を少なくする必要があるためモデルを誤認識させることができる確率も小さくなる。この誤差逆伝搬を実行する回数を敵対的サンプルのステップ数と呼ぶ。

2.2　Fast Gradient Sign Method (FGSM)

Fast Gradient Sign Method (FGSM) [2] は高速に敵対的サンプルを作成する手法である。一度だけ誤差逆伝搬を行い、入力の各次元の要素に対して損失が大きくなる方向に等しい大きさの摂動を加える。FGSM は 1 ステップで実行できるため少ない計算量で敵対的サンプルを作成できるが後に述べる PGD と比べると誤認識を引き起こせる確率は低くなる。

2.3 Projected Gradient Descent (PGD)

Projected Gradient Descent (PGD) [3] は FGSM よりも高い確率で誤認識を引き起こす強力な敵対的サンプルを作成する手法である。逐次的に細かいステップで最も損失の大きいサンプルを作成できる。しかし FGSM よりも大きな計算量が必要となる。

2.4 敵対的学習

敵対的学習 [3] は敵対的攻撃に対して頑健な深層学習モデルを学習する手法である。学習時に入力データを用いて敵対的サンプルを作成し、それに対して正しいラベルを出力するように矯正する手法である。

敵対的学習は PGD のような高い確率で誤認識を引き起こす強力な敵対的攻撃手法を用いると頑健なモデルが学習できるが、PGD の実行には大きな計算量が必要となるため敵対的学習に必要な計算量も大きくなるという問題点がある。また、PGD の代わりに FGSM のような 1 ステップの弱い攻撃手法を用いるとその攻撃手法に過学習してしまいより強力な手法に対しては脆弱になってしまう [5]。

2.5 ノイズ付与

深層学習モデルの学習時に敵対的サンプルの代わりにランダムノイズを加えた入力を用いても敵対的攻撃に対してある程度頑健になる [4]。この手法は余分な誤差逆伝搬を計算しないため敵対的学習と比べて少ない計算量で学習できる。しかし、敵対的学習ほど高い頑健性は得られない。

3 敵対的勾配付きノイズ付与

以上のように敵対的攻撃に対して頑健なモデルを学習するためには大きな計算量が必要となり、少ない計算量の手法を用いると頑健性が下がるというジレンマが存在する。そこで本研究では一度のみ誤差逆伝搬を計算し、その勾配の方向が出やすいように重みをつけたランダムノイズを入力に加えて学習することで、1 ステップの敵対的攻撃を用いた敵対的学習で起こる過学習を防ぎ、頑健性を高めつつ計算量を少なく抑える手法を提案する。この手法は FGSM を用いた敵対的学習とほぼ同等の計算量で実行できるが、ノイズの汎化性能により攻撃手法に対する過学習が起きないためより強力な攻撃手法に対して頑健になる。この手法を敵対的勾配付きノイズ付与と呼ぶ。

次元数 n、標準偏差 σ の正規分布から生み出されるサンプルは半径 $\sigma\sqrt{n}$ の球の表面近くに多く存在している。よってある勾配の方向に重みをつけたガウシアンノイズは

1. 勾配の L_2 ノルムを $\sigma\sqrt{n}$ に合わせる。
2. 1 の結果に平均 0、標準偏差 σ の正規分布を加える。
3. 2 の結果のノルムを $\sigma\sqrt{n}$ に合わせる。

という手順で計算できる。この手順を図で表したものを 1 で示す。

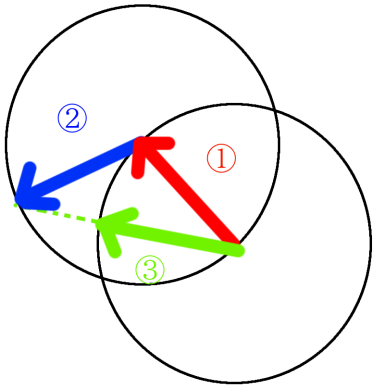

図 1　手順 1(赤) は誤差逆伝搬で求めた勾配、手順 2(青) はランダムノイズ、手順 3(緑) でそのノルムを $\sigma\sqrt{n}$ に合わせる。

このように損失関数の勾配を用いつつランダムノイズを加えることで 1 ステップの攻撃手法を用いた敵対的学習で発生する過学習を避けることができる。また、誤差逆伝搬を一度のみ行うため複数ステップの攻撃手法を用いた敵対的学習と比べて必要な計算量が少なくて済む。

4 実験

実験では画像分類のデータセットである CIFAR-10 [6] を用いる。深層学習モデルは Wide ResNet [7] を用いる。Wide ResNet は CIFAR-10 データセットで学習すると高い精度を得ることができる。

学習時のパラメータとしてエポック数は 200、バッチサイズは 128、重み減衰項の係数は 0.0005 とする。これはノイズ付与 [4] の論文と同じである。

計算機環境は Intel Xeon E5-2680 V4 Processor と GPU Tesla P100 を一枚搭載したコンピュータを用いる。

評価指標として頑健性を測るために敵対的サンプルを与えたときに認識率 (%)、計算量を測るためにモデルの学習に要した時間 (h)、モデルの精度を測るためにノイズの重畳されない画像に対する認識率 (%) の 3 つを用いる。頑健性を測るときの敵対的サンプルは 100 ステップの PGD で生成したものを用いる。

本研究の実験では 5 つの従来手法で訓練したモデルと提案手法で訓練したモデルを比較する。それぞれの

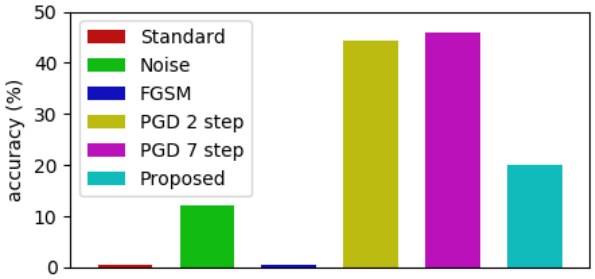

図 2　敵対的サンプルの認識率。

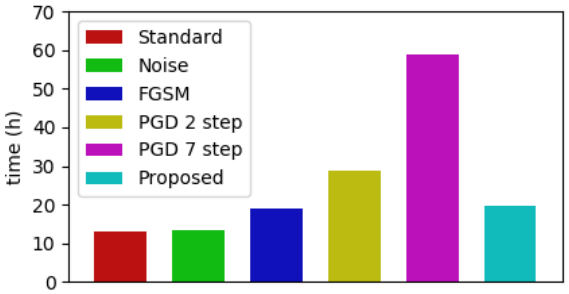

図 3　モデルの学習に要する時間。

モデルの訓練の手法を下に示す。

Standard 通常の学習
Noise ガウシアンノイズ付与
FGSM FGSM を用いた敵対的学習
PGD 2 step 2 ステップの PGD を用いた敵対的学習
PGD 7 step 7 ステップの PGD を用いた敵対的学習
Proposed 提案手法

以上の 6 つの手法で訓練したモデルを実験で比較する。

5　結果

敵対的サンプルに対する認識率を図 2 に示す。敵対的勾配付きノイズ付与を用いたときの認識率は 20.0%であり、通常の学習をしたモデルや FGSM を用いた敵対的学習をしたモデルと比べるとある程度の頑健性が得られた。ノイズのみを用いたモデルは 12.1%なのでそれと比較しても高い認識率である。2 ステップ以上の敵対的学習と比べると頑健性が低くなったが、それらの半分程度の頑健性は得られた。

モデルの学習に要する時間を図 3 に示す。敵対的勾配付きノイズ付与は 19.8 時間で学習を完了した。FGSM を用いた敵対的学習とほぼ同等の値である。2 ステップ以上の敵対的学習と比べると 2 ステップで 28.8 時間、7 ステップで 58.8 時間かかったためそれらより少ない計算量で学習できた。

元画像に対する認識率を図 4 に示す。敵対的勾配付きノイズ付与の認識率は 82.2%であり今回実験した手法の中で元画像に対する認識率が最も低くなった。

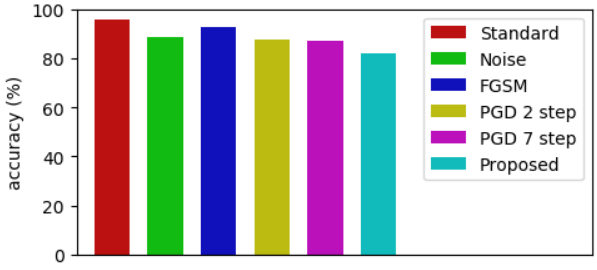

図 4　敵対的攻撃を加えていない元画像の認識率。

6　結論

本研究では少ない計算量で敵対的攻撃に対して頑健なモデルを学習するために学習時に一度誤差逆伝搬し、その勾配の方向に重みをつけたランダムノイズを入力に加える敵対的勾配付きノイズ付与という手法を提案した。

敵対的勾配付きノイズ付与を用いて学習したモデルは敵対的サンプルを 20.0%の精度で認識できた。また、敵対的勾配付きノイズ付与は GPU 一枚を搭載した計算機で 19.8 時間で実行することができた。

提案手法は敵対的攻撃に対してある程度の頑健性が得られたが、敵対的学習と比べると半分以下の頑健性になってしまうため計算量を維持しつつより頑健性を上げる手法の開発が今後の課題となる。

参考文献

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. Intriguing Properties of Neural Networks. *arXiv* 2013.

[2] Ian Goodfellow, Jonathon Shlens, Christian Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv* 2014.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* 2017.

[4] Nic Ford, Justin Gilmer, Nicolas Carlini, Dogus Cubuk. Adversarial Examples Are a Natural Consequence of Test Error in Noise. *arXiv* 2019.

[5] Alexey Kurakin, Ian Goodfellow, Samy Bengio. Adversarial Machine Learning at Scale. *arXiv* 2016.

[6] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. *https://www.cs.toronto.edu/ kriz/* 2009.

[7] Sergey Zagoruyko, Nikos Komodakis. Wide Residual Networks. *arXiv* 2016.