A-014

遺伝的アルゴリズムを用いた敵対的サンプルの最適化

田中 大樹 †
東京理科大学 理工学研究科 情報科学専攻 修士 1 年 ‡

1. はじめに

近年、ニューラルネットワークを多層にしたものである深層学習が様々な分野で従来の結果を上回るような結果を出している。特に、畳み込みを利用した畳み込みニューラルネットワークは画像認識だけでなく自然言語や音声の分野でも優秀な性能を示している。しかし、悪意のあるノイズを掛けられた画像などにおいては、ニューラルネットワークは高確率で誤認識してしまうことが知られている。これらの画像のことを敵対的サンプルという。ニューラルネットワークを実社会で用いる際はこのような攻撃について熟知し、それに対策できなくてはならない。よって敵対的サンプルの界隈では、新たな攻撃手法を提案し、それに対する防御手法を提案することでより防御性能の高いニューラルネットワークを研究している。
本稿では、攻撃手法の 1 つである FGSM を用いて敵対的サンプルを生成し、より攻撃性能をあげるために遺伝的アルゴリズムを用いて最適化する実験を行い、その結果について考察する。

2. 基礎知識

2.1 ニューラルネットワーク

ニューラルネットワークとは、脳の構造を模して造られた機械学習の手法の 1 つである。脳細胞を模したニューロンを層状に組み合わせて計算モデルを構築する。そのモデルに対し、求める結果を出力するように与えられたデータを用いてパラメータを調整することを学習という。学習を繰り返すことで、未知のデータに対しても高確率で正しく分類することが可能となる。畳み込み層とプーリング層を用いた畳み込みニューラルネットワーク (Convolutional Neural Network) は、画像認識において非常に優秀な結果を収めており、画像認識だけでなく様々な分野でもよく用いられている。

2.2 敵対的サンプル

敵対的サンプル (Adversarial Examples) とは、学習器に対して意図的に出力を誤らせるような摂動を与えた入力サンプルのことである。画像だけでなく、映像や音声などの敵対的サンプルも存在するが、今回は画像のみを扱う。敵対的サンプルの特徴として、人間の目には元の画像とほぼ同じか、多少ノイズが載っている程度としか認識できないといった点が挙げられる。そのため、与える摂動の量には制限が与えられている。主に L ∞ノルムや L2 ノルムを用いて摂動の最大値を設定していることが多い。
また、攻撃対象となる学習器の情報がどれだけ得られるか、どのように出力を誤らせるかによって攻撃の難易度が大きく変わるため、異なる攻撃と分類する。学習器の情報が手に入るか否かでホワイトボックス攻撃とブラックボックス攻撃に分類され、出力を特定のクラスに誤らせるか否かで targeted 攻撃と untargeted 攻撃に分類される。

2.3 遺伝的アルゴリズム

遺伝的アルゴリズムとは、生物の進化を模して造られた最適化手法の 1 つである。解の候補を遺伝子として数値で表現し、選択、交叉、突然変異を繰り返し行うことで、よりよい遺伝子を探索する。様々な選択方法、交叉方法が存在し、問題の条件によって使い分けることが重要である。
また、遺伝子の表現方法として、実数を用いる方法と実数を二進数を用いて表す方法がある。前者はバイナリ型遺伝的アルゴリズム、後者は実数値型遺伝的アルゴリズム (以下 RCGA) などと呼ばれる。RCGA で主に使われている交叉方法では、親個体の分布よりも広い範囲に子個体が生成されるので、突然変異を実装する必要性が薄い。
本稿では RCGA で用いられている選択手法、交叉手法について予備実験を行い、敵対的サンプルの最適化に適した手法を調べ、それを採択した。以下で採択した手法について簡単に述べる。

2.3.1 世代交代モデル

1. Minimal Generation Gap
   佐藤らによって提案された世代交代モデルである。MGG では複製選択で 2 体の親を選択するため、交叉方法は二親交叉に限られる。MGG を多親交叉に拡張したモデルも考えられており、複製選択として n 体の親を選択するが交叉に用いる親の個体数は変わらず 2 体である。
   - 複製選択
     適応度を無視して、現世代から個体をランダムに 2 体選び非復元抽出する。
   - 生存選択
     非復元抽出した親個体 2 体と交叉によって生成された子個体群の中から、エリート選択によって選ばれた 1 個体とルーレット選択によって選ばれた 1 個体を現世代に加え、それを次世代とする。

2.3.2 交叉手法

1. BLX-α
   ブレンド交叉は Esheleman によって考案された交叉手法である。主に二親交叉で用いられ、多親交叉でも用いられることはあるが生成に用いる親個

Adjustment of adversarial examples with genetic algorhythm
†Taiki Tanaka
‡Tokyo University of Science

体の個数は 2 体である。BLX-α では、子個体を以下の手順に沿って生成する。

(a) 交叉に用いる親個体を 2 体選択し、親個体の遺伝子をそれぞれ $X_i^1, X_i^2 (1 \leq i \leq n)$ とする

(b) $d_i = |X_i^1 - X_i^2|$
$minX_i = min(X_i^1, X_i^2)$
$maxX_i = max(X_i^1, X_i^2)$ を求める

(c) 子個体の遺伝子 Y_i を
区間 $[minX_i - \alpha d_i, maxX_i + \alpha d_i]$ から一様乱数に従って決定する

BLX-α では、親個体のパラメータの距離が広ければ子個体も広い範囲に生成され、距離が狭ければ子個体も狭い範囲から生成される。

3. 関連研究

3.1 Fast Gradient Sign Method(FGSM)

FGSM とは Goodfellow ら [1] が提案した、敵対的サンプルを生成する際の手法の 1 つである。初期から存在する手法であるが、アルゴリズムがシンプルかつ高速に敵対的サンプルを生成できるのでよく用いられている。この手法では、元の画像を x、与える摂動を η、敵対的サンプルを $x + \eta$ で表している。このとき、摂動 η は以下の式を用いて計算される。

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

ここで J は対象となるモデルの損失関数、$\boldsymbol{\theta}$ はモデルのパラメータ、$\boldsymbol{x}, y$ はモデルへの入力と出力である。また、$sign$ は signature の略であり符号関数である。ϵ は摂動の改変量を決めるパラメータであり、この値によって FGSM を用いて作られた摂動 η の条件 $\|\eta\|_\infty \leq \epsilon$ が決定する。通常のニューラルネットワークの学習では、損失関数の勾配を用いて損失が減少するように重みの値を調整することで分類の精度を高める。しかしこの手法では、損失関数の勾配を用いて、正しいラベル y に対する損失を増加させるように摂動を計算することによって、学習器が誤認識を引き起こすようにしている。損失関数の勾配の符号のみを用いることによって摂動の各ピクセルの改変量は一定となってしまうが、高速に敵対的サンプルを生成することができる。
また、FGSM 以外にも多くの敵対的サンプルの生成方法が提案されており、それらは Parenot らがまとめた Cleverhans[2] というライブラリで公開されており、簡単に実験ができるようになっている。

3.2 ブラックボックス上での FGSM の再現

FGSM によって生成された摂動の各ピクセルの改変量が一定であることを利用して、ブラックボックス上で擬似的に FGSM を再現することができる [3]。各ピクセルの改変量は $+\epsilon, -\epsilon, 0$ のいずれかであるが、損失関数の勾配が 0 となることはほぼ存在しないので $+\epsilon, -\epsilon$ のどちらかであるとする。このとき、あるピクセルに $+\epsilon$ して誤認識率が高まればそのまま、逆であれば符号を逆にするという操作を各ピクセルごとに行うことにより、ブラックボックス上で FGSM と同程度の改変量の敵対的サンプルの生成を可能としている。

4. 提案手法

FGSM は L∞ノルムによって摂動の量を制限しており、各ピクセルに定数 ϵ を足して敵対的サンプルを高速に生成することを実現している。よって、求められた摂動の各ピクセルの値は $-\epsilon$ か $+\epsilon$ のどちらかである。このとき $(-\epsilon, +\epsilon)$ の範囲に、より誤認識率が高い解がある可能性を捨ててしまっている。しかし、各ピクセルが $(-\epsilon, +\epsilon)$ の範囲で局所解を持つか否かは明らかではないので、摂動の各ピクセルで最適解を求めるのは難しい。
よって本稿では、疑似 FGSM を用いて MNIST の画像から敵対的サンプルを生成し、生成した摂動の各ピクセルにおいて、より誤認識率が高い解を RCGA によって探索した。具体的な流れを以下に示す。

4.1 RCGA による探索の流れ

MNIST の画像 1 枚当たりの画素数は 28×28 であり、これを同時に探索するにはあまりにもパラメータが多いため、1 枚を 28 行に分けて 1 行ずつ探索を行う。疑似 FGSM によって得られた摂動を η、改変量を ϵ、現世代の個体数を N_{pop}、生成個体数を N_c として手順を示す。

1. η の r 行目を遺伝子の初期値 η_r とする
2. 現世代として η_r を遺伝子に持つ個体を 1 体、$\eta_r \pm \frac{\epsilon \cdot u(0,1)}{5}$ を遺伝子に持つ個体を $N_{pop} - 1$ 体生成する。$u(0,1)$ は区間 $[0,1]$ の一様乱数である。
3. 評価関数として元々の画像の正解ラベルの推論値を用いて、現世代の個体の適応度を計算する
4. 親となる個体を選択方法に基づいて選択し、親個体群を生成する
5. 交叉方法に基づいて親個体群から子個体を生成する
6. 子個体群の個体数が N_c になるまで子個体を生成する
7. 世代交代モデルに基づいて、親個体群、子個体群から次の世代を決定する
8. 世代数が 50 となるまで 3~7 を繰り返す

5. 実験

5.1 実験概要

手書き数字のデータセットである MNIST の一部のテストデータに対して敵対的サンプルを生成し、RCGA を用いてその敵対的サンプルを改善した結果を示す。

5.2 実験条件

敵対的サンプルの攻撃対象とした学習器について、構造とパラメータを表 1, 表 2 に示す。

表 1, 表 2 で述べた学習器を対象にして敵対的サンプルを生成する。また、敵対的サンプルの分類はブラックボックスかつ Non-Targeted 攻撃であるので、表 1,2 のパラメータは生成に利用していない。
用いる画像は MNIST のテストデータセットのうち、正しく認識されている 9922 枚中の 100 枚とする。疑似 FGSM を用いて敵対的サンプルを作る際の摂動の制限

表 1: 対象となる学習器の構造

層	構成
Convolition	$5 \times 5 \times 32$
MaxPooling	2×2
Convolition	$5 \times 5 \times 64$
MaxPooling	2×2
Dropout	0.25
Dense	512
Softmax	10

表 2: 対象となる学習器の学習パラメータ

パラメータ	値
最適化アルゴリズム	Adam(デフォルトパラメータ)
Batch Size	200
Epochs	10

を $\epsilon = 0.1$ としたところ、100 枚中 38 枚が攻撃成功となった。

この 38 枚の敵対的サンプルについて、実数値型 GA を用いて最適化を行う。実数値型 GA の選択方法、交叉方法などのパラメータは表 3 に示す。また、全てのピクセルについて一度に最適化を行うとあまりにもパラメータが多いため、得られた敵対的サンプルを一行ごとに分けて最適化を行い、一行あたりの最適化された敵対的サンプルを取得し、それを組み合わせて最終的な出力としている。

表 3: RCGA のパラメータ

パラメータの種類	値
遺伝子長 (dim)	28
世代数	50
現世代個体数	$15 \times dim$
生成個体数	$10 \times dim$
親個体群個体数	$dim + 1$
(BLX-α の α)	0.5

5.3 結果

最適化を行った 38 枚のうち、改善率の上位 3 枚、下位 3 枚についての表を示す (表 4, 表 5)。なお、改善率は

$$\frac{\text{正解ラベルの元の推論値} - \text{最適化後の正解ラベルの推論値}}{\text{正解ラベルの元の推論値}}$$

として計算しており、この値が大きいほど学習器はより誤った認識をしているので、敵対的サンプルの精度がより改善されているとできる。38 枚全ての結果についての表は付録に載せる。

5.4 評価

RCGA を適用した全ての敵対的サンプルが改善されている (付録、表 6)。上位では 50~60%、下位では 15%

表 4: 実験 2 の結果 (上位 3 枚)

番号	元の推論値	改善後の推論値	改善率
38	0.01179016	0.00419930	0.64383010
59	0.00008243	0.00003746	0.54555380
72	0.29179689	0.14393593	0.50672562

表 5: 実験 2 の結果 (下位 3 枚)

番号	元の推論値	改善後の推論値	改善率
53	0.20814267	0.17491722	0.15962824
80	0.00167863	0.00143894	0.14278906
96	0.03636950	0.03134882	0.13804616

前後、平均では 31%の改善率を記録しており、RCGA を用いてブラックボックス上の敵対的サンプルを改善することに成功した。

6. まとめ

本稿では、ブラックボックス上の Non-Targeted 攻撃である疑似 FGSM を用いて生成した敵対的サンプルを、RCGA によって改善する手法を提案し実験と評価を行った。実験を行った全ての敵対的サンプルにおいて、元の推論値との割合で平均 30%の改善に成功した。しかし、実験に用いた RCGA の交叉手法、世代交代モデルは既存のものであり、より適した手法の探索については課題が残った。

本研究を通じて攻撃側の手法を改善することができた。これに対して防御側も改善されることによって、深層学習のセキュリティのさらなる向上が期待できる。

7. 付録

実験 2 で用いた 38 枚全ての結果を表 6 に示す。

表 6: 実験 2 の結果

番号	元の推論値	改善後の推論値	推論値の差
2	0.02532866	0.01360973	0.01171893
4	0.01786746	0.01050556	0.00736189
5	0.05554723	0.04103282	0.01451441
6	0.00325899	0.00232766	0.00093133
7	0.25756496	0.20722783	0.05033714
8	0.00005589	0.00004578	0.00001011
12	0.07231835	0.05471672	0.01760163
15	0.00036194	0.00028543	0.00007651
18	0.00164694	0.00112658	0.00052036
20	0.01585304	0.01167306	0.00417997
24	0.00274958	0.00160349	0.00114609
29	0.30210915	0.21881489	0.08329426
33	0.00983654	0.00692383	0.00291271
36	0.00516322	0.00430762	0.00085560
38	0.01179016	0.00419930	0.00759086
40	0.02275388	0.01500602	0.00774786
41	0.00136108	0.00103604	0.00032504
42	0.10007794	0.07328338	0.02679456
44	0.28052419	0.18281995	0.09770425
48	0.11692267	0.09548074	0.02144193
53	0.20814267	0.17491722	0.03322545
57	0.01374103	0.00940024	0.00434079
58	0.02783094	0.02172517	0.00610578
59	0.00008243	0.00003746	0.00004497
62	0.00022515	0.00013687	0.00008827
63	0.01977429	0.01336875	0.00640554
65	0.00152916	0.00096261	0.00056655
66	0.21957387	0.17853768	0.04103619
67	0.00096763	0.00070001	0.00026763
72	0.29179689	0.14393593	0.14786096
73	0.00107557	0.00076898	0.00030659
78	0.01289239	0.00664908	0.00624331
80	0.00167863	0.00143894	0.00023969
89	0.00154832	0.00095513	0.00059319
92	0.00006806	0.00003844	0.00002962
93	0.05559240	0.03695535	0.01863705
95	0.01112692	0.00814395	0.00298297
96	0.03636950	0.03134882	0.00502067

参考文献

[1] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. ”Explaining and Harnessing Adversarial Examples” arXiv preprint arXiv:1412.6572, 2014.

[2] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, Rujun Long, Patrick McDaniel. ”Technical Report on the CleverHans v2.1.0 Adversarial Examples Library” arXiv preprint arXiv:1610.00768

[3] 先崎 佑哉, 大畑幸矢, 松浦幹太. ”深層学習に対する効率的な Adversarial Examples 生成によるブラックボックス攻撃とその対策”, 2018 念暗号と情報セキュリティシンポジウム (SCIS2018), 予稿集 USB メモリ, 3F1-4. 新潟, 1 月, 2018 年