5P-01

遺伝的アルゴリズムを用いた敵対的サンプルの生成と調整

田中 大樹 †
東京理科大学 理工学部 情報科学科 学部 4 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 GA、後者は実数値型 GA などと呼ばれる。実数値型 GA 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。
以下に実数値型 GA で用いられている選択手法、交叉手法を簡単に述べる。

2.3.1 選択手法

1. Minimal Generation Gap
   現世代の個体の集団から親個体を選択し、その個体を交叉させてできた子集団に選択した親個体を含めた集団を生成し、その集団の最良個体とルーレット選択により得られた個体を最初に選んだ親個体と入れ替えて新世代とする手法である。

2. Just Generation Gap
   親個体の選択方法は MGG と変わらないが、交叉に参加した親個体群と、子集団から評価値最良順に選んだ子個体群を入れ替えて新世代とする手法である。

2.3.2 交叉手法

1. BLX-α
   親個体を 2 体選択し、それぞれのパラメータについて区間 d_i を計算して、区間 d_i を両側に $\alpha \times d_i$ だけ拡張した区間から子個体のパラメータを一様乱数に従ってランダムに生成する手法である。

2. Simplex
   シンプレックスとは n 次元空間において、n+1 個の点からなる図形のことである。例えば 2 次元平面の場合は三角形がシンプレックスとなる。同様に、遺伝子長を n としたとき、親個体を $n+1$ 体選択して、親個体を内部に含むようなシンプレックスから子個体のパラメータを一様乱数に従ってランダムに生成する手法である。

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

3. 関連研究

敵対的サンプルの生成方法として一般的に使われているものに、Fast Gradient Sign Method(FGSM)[1] が挙げられる。対象となる学習器の損失関数の勾配の正負を用いて与える摂動を計算しており、これ以前に提唱されていた方法より高速で敵対的サンプルを生成することが可能となった。これは学習器の内部情報を用いているためホワイトボックス攻撃に分類されるが、FGSM で得られる敵対的サンプルは各ピクセルの値に定数 ϵ を加えるか減らすかのいずれかであることを利用してブラックボックス条件下で疑似的に再現する方法も提案されている [2]。
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[3] 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。

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、求められた摂動に対して実数値型 GA を用いてより適した解を探索することについて実験し、その効果を検討する。

5. 実験と結果

5.1 実験

MNIST を学習させた CNN を攻撃対象として敵対的サンプルを生成する。学習器の構造とパラメータを表 1, 表 2 に示す。

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

このモデルに対して、疑似的に再現した FGSM を用いて敵対的サンプルを生成する。敵対的サンプルが生成出来たら、実数値型 GA を用いてそのサンプルの最適化を行う。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、それを最終的に組み合わせて出力としている。

表 2: 対象となる学習器のパラメータ

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

表 3: 実数値型 GA のパラメータ

パラメータ	手法または値
選択方法	MGG
交叉方法	BLX-α
評価関数	学習器の正解ラベルの推論値
遺伝子長	28
集団数	20
世代数	50

5.2 結果

余白の問題上、学会にて発表する。

6. まとめ

本稿では、疑似 FGSM を用いて作られた敵対的サンプルを、実数値型遺伝的アルゴリズムを用いて最適化する方法について実験し、その効果を検討した。
FGSM 以外の敵対的サンプルの生成手法についても遺伝的アルゴリズムを用いて最適化することを今後の課題とする。

参考文献

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

[2] 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

[3] 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