1R-08

遺伝的アルゴリズムに基づく DNN 敵対的サンプルの生成

李　文韜　　馬　雷　　趙　建軍
九州大学 システム情報科学研究院

1. はじめに

ディープニューラルネットワーク(DNN)は近年機械学習の分野で最も人気のある課題である。DNN の流行は、伝統的な機械学習分野における多くの仕事の識別率を大幅高めたことに成功しする。DNN は多くの分野で急速に流行してきたので、DNN の堅牢さは重要である。DNN の堅牢性を向上すると DNN の安全問題を発見するために、テストが重要な手段である。DNN のテストには 3 つの重要点がある，テストのカバレッジ基準，DNN テストのテストデータ自動生成、DNN テストのテストデータの品質の評価である。今行なっている研究では，新たな遺伝的アルゴリズムに基づく深層学習システムの自動テスト方法を提案する。この方法の一番重要なステップはテストサンプルの生成することである。本論は，実験を行うと、少し修正すると特別なテストサンプル--敵対的サンプル生成することができる。入力に小さい変化を追加して、DNN の出力に間違って分類させる可能性がある。そして本論も将来に一般的なテストサンプルを生成する方法を構想する。

2. 提案手法

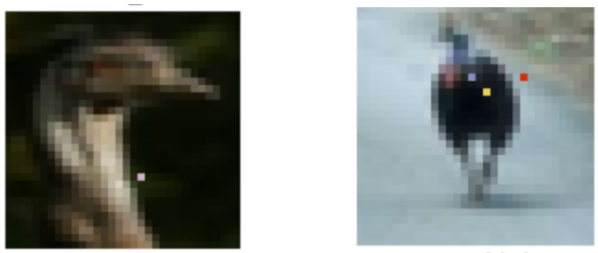
2.1 敵対的サンプル

自然画像に対する人工的な妨害は、DNN に間違えて分類を行わせやすく、これら生成のサンプルは敵対的サンプルと呼ばれている。敵対的サンプルを生成するしことで、DNN の脆弱性を発見できる。図 1 に示すように、妨害を加えて、画像を間違えて分類する。これは敵対的サンプルである。

敵対的サンプルを生成する正しい方法は、元の画像に小さい妨害を加える。この妨害は人間にとっては気づけないが、DNN に全く異なる分類として識別させることができるものはずである。でも今多く敵対的サンプルの生成方法は、サンプルの変更部分は限りがあるはずだと考えていない。即ち、変更されたピクセルの数が人間に気づかれてしまうほど多すぎである。

本研究では、以前の研究[1]を参考し、少しピクセルのみ(1,3,5)を変更しでも、効果の良い敵対的サンプルを生成できる方法を提案する。

図 1 ：敵対的サンプル



2.2 遺伝的アルゴリズム

遺伝的アルゴリズムはデータ（候補解）を遺伝子で表現した個体を複数用意し、適応度の高い個体を選択して組み換え・突然変異などの操作を繰り返しながら解を探す。その中に、適応度は適応度関数によって与えられる。

この方法の利点は、評価関数の微分可能性と単峰性に関する知識がなくても適用できることである。 必要条件は評価関数の全順序であり、探索空間はトポロジーを有する。

図 2 の遺伝的アルゴリズムは、あるソフトウェアテストの研究[2]で利用した遺伝的アルゴリズムである。この遺伝的アルゴリズムはソフトウェアデータスイトを生成するときの表現が良い。ここで利用した時に、アルゴリズムの適応度関数(Fitness Function)は目標分類の信頼度を設定する。

2.3 DNN 自動テストの構想

今の構想は、遺伝的アルゴリズムを利用して、DNN の自動テストすることができるようになる。従来のソフトウェアと同様に、カバレッジはテストの十分さを判断するための良い方法である。すなわち、DNN テストの最も重要な部分は、多くのネットワーク部分をカバーするテストサンプ

Generating Adversarial Examples Based on Genetic Algorithm
Wentao Li,Lei Ma,Jianjun Zhao
Faculty of Information Science and Electrical Engineering,Kyushu University

図２：利用された遺伝的アルゴリズム

```
current_population ← generate random population
repeat
  Z ← elite of current_population
  while |Z| ≠ |current_population| do
    P1,P2 ← select two parents with rank selection
    if crossover probability then
      O1,O2 ← crossover P1,P2
    else
      O1,O2 ← P1,P2
    mutate O1 and O2
    fP = min(fitness(P1),fitness(P2))
    fO = min(fitness(O1),fitness(O2))
    lP = length(P1) + length(P2)
    lO = length(O1) + length(O2)
    TB = best individual of current_population
    if fO < fP ∨ (fO = fP ∧ lO ≤ lP) then
      for O in {O1,O2} do
        if length(O) ≤ 2 × length(TB) then
          Z ← Z ∪ {O}
        else
          Z ← Z ∪ {P1 or P2}
    else
      Z ← Z ∪ {P1,P2}
  current_population ← Z
until solution found or maximum resources spent
```

ルを生成することである。今の構想は遺伝的アルゴリズムの Fitness Function を DNN のカバレッジを設定する、最大化ために最適化する。ある研究は DNN のカバッレジを定義した[3][4]。今後の研究としでは、最も適した DNN カバッレジを選択し、自動的なテストツールを開発する。

3. 実験

遺伝的アルゴリズムを利用し、cifar10 のデータセットで攻撃を行う。cifar10 データセットには、10 つ分類の画像データが含まれている。

攻撃は二部に分ける：目的ある攻撃と目的なし攻撃である。目的ある攻撃は、攻撃したいサンプルを間違う分類がある、そして攻撃の適応度 F は、攻撃しているサンプルを目標の分類信頼度を 50%になるために最適化する。一度目標分類の信頼度が 50%になると、進化が止める。目的なし攻撃は、攻撃したいサンプルが特に目標分類がしなくて、今の正しい分類を低くになるため最適化にする。停止の条件は：一度正しい分類の信頼度が 50 以下になる時、または遺伝的アルゴリズムの繰り返す数が 100 になる時である。

実験は４つの普通な DNN モデルの上で行う。モデルのパラメーター数と出力の精度は表 1 に示す。

実験を行なった時、各モデルが目的なし攻撃で 1000 つ画像を選択された。そして、変わるピクセルに 1、３、５を設定する。目的ある攻撃は、各モデルが 100 つの画像を選択し、各画像は別の９分類を１、３、５ピクセルに変更し、攻撃を行う。

表１　実験を行う DNN モデル

モデル	パラメンタ	精度
LeNet	62K	74.9%
ResNet	470K	92.3%
pureCNN	1.4M	88.8%
DenseNet	850K	94.7%

表２　実験結果

モデル	ピクセル	目的なし攻撃成功率	目的ある攻撃成功率
LeNet	1	57.0%	21.9%
	3	94.6%	64.3%
	5	98.2%	78.0%
ResNet	1	26.2%	0.80%
	3	64.7%	27.2%
	5	81.0%	45.5%
PureCNN	1	8.08%	1.15%
	3	48.5%	10.1%
	5	65.1%	19.0%
DenseNet	1	20.1%	4.04%
	3	69.5%	25.1%
	5	72.1%	30.5%

4. まとめ

今研究では、遺伝的アルゴリズムで敵対的サンプルを生成する。少しだけピクセルを変更しでも敵対的サンプルが生成できる。

今後の研究では、この遺伝的アルゴリズムを利用し、DNN のカバレッジ基準を決めて、カバレッジが高いテストサンプルを生成する。

参考文献

[1] Jiawei Su, One pixel attack for fooling deep neural networks,arViv.org,2017.

[2] G. Fraser,Whole Test Suite Generation, IEEE Transactions on Software Engineering, vol. 39, iss. 2, pp. 276-291, 2013.

[3] Lei Ma, DeepGauge, ACM/IEEE ASE,2018.

[4] Kexin Pei, DeepXplore, SOSP'17,20