深層学習に対する欺き画像の生成技術の振る舞い解析と耐性向上

大島　辰之輔†　　佐藤　真一††
† 東京大学大学院　情報理工学系研究科　　†† 国立情報学研究所

1. はじめに

深層学習の一つである畳み込みニューラルネットワーク(CNN)は，一般物体認識の分野においてその認識精度の高さから近年多く研究されている．しかしその動作原理は深く探求されておらず，人間の認識と異なる点もあると言われる．その一つが，人間に知覚できないが深層学習モデルが高い確信度で特定のクラスへと認識してしまうような欺き画像が存在することである．このような欺き画像が簡単に生成できる[1]ことを Nguyen らは示したが，本稿ではその欺き画像生成技術における振る舞いを解析し，さらに深層学習モデルが欺き耐性を獲得できることを実験により証明する．

2. 欺き画像生成の振る舞い解析

Nguyen らの直接コード化と間接コード化との 2 種類の手法を用いて，MNIST 手書き数字データセットを認識できる Lenet[2]深層学習モデルに対し欺き画像を生成する．結果，図 1 のように生成された画像に対し99.99%以上の確信度で認識することを確認した．

欺き画像に対して Lenet の中間層における出力を特徴ベクトルとして LIBLINEAR による線形分離で交叉検定を行うと，90%以上分類に成功した．これは欺かれた深層学習モデルも，特徴抽出段階では欺き画像を識別する潜在的能力を持つことを示唆している．

3. 欺き画像に対する深層学習の耐性向上

欺き画像を数字クラスとは別の第 11 クラス目と判別させる深層学習モデルを生成し，そのモデルに対して更に 2 と同様に欺き画像を生成する．これを 15 世代分繰り返し，欺き画像を学習した深層学習モデルが欺き画像に対する耐性を獲得できるのかを検証する．

生成された欺き画像に対して深層学習モデルが出した確信度の平均と中央値を図 2 に示す．深層学習モデルの世代の進化とともに欺き画像に対しての確信度が下がっている．深層学習が欺き画像を学習したことで欺きにくくなった結果であり，耐性を獲得できたと言える．

また，欺き画像の生成時に用いている進化的アルゴリズムの世代数を増やしたところ，高い確信度で認識されてしまう画像が生成された．しかし，図 3 に示したように実際に生成された画像を観察すると，人間の知覚でも判別できる画像が生成されており，欺き画像と言えない画像であることが判明した．

図1．欺き画像生成例，左から 1, 2, 3, 4, 5 と識別される

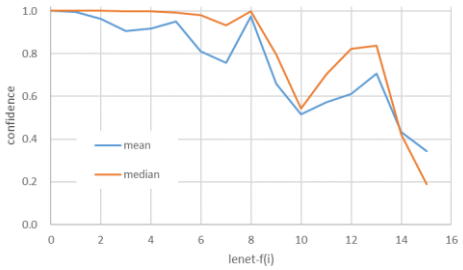

図 2．欺き耐性深層学習モデルの世代と欺き画像確信度推移

図 3．耐性モデルに対して高い確信度を持つ欺き画像生成結果
左から 1, 2, 3, 4, 5 と認識される．

4. 深層学習における欺き耐性の汎化

深層学習が欺き耐性を獲得し，その脆弱性を解決するには未知の手法により生成された欺き画像に対してもその耐性が示されなければならない．そこで，間接コード化による欺き画像を学習させたモデルに，未知の手法と呼べる直接コード化による欺き画像を認識させる実験を行った．

その結果，耐性を持つ深層学習モデルは未知の手法により生成した欺き画像をほぼ 100%の精度で認識できた．このことから，欺き画像を持つ深層学習モデルは，その生成手法が未知の欺き画像に対してもその耐性を発揮できることを証明した．

5. まとめ

欺き画像に対して，深層学習モデルが認識できる能力を持つことを振る舞い解析によって示し，実際に欺き画像を学習させたモデルを用いて実験することで欺き耐性向上を証明した．さらに欺き画像を学習したモデルが未知の欺き画像にも耐性を持つことを実証した．

参考文献

[1] A. Nguyen, et al., Proceedings of the IEEE Conference on CVPR. pp. 427-436, 2015.

[2] Y. LeCun, et al., Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278-2324, 1998.