

JPEG 圧縮による画像分類器の識別結果の変動解析に基づく 敵対的事例の検知法

東 亮憲^{1,a)} 栗林 稔¹ 船曳 信生¹ HUY H. NGUYEN² 越前功²

概要: CNN を用いた画像分類器の研究が盛んに進められている一方で、その画像分類器を欺くことができる敵対的事例の存在が指摘されている。敵対的事例とは、意図的に造られた小さな摂動を元の画像に加えることで、画像分類器の識別結果が元の画像と異なるクラスを指す画像である。人間の目では元の画像との差異を判別することが困難であるため、敵対的事例を検知することが一般的には難しい。本研究では、敵対的事例と正常な画像を識別することを目的とする。画像分類器が誤った結果を出力する理由は、敵対的事例の微小なノイズが影響しているためである。提案手法では、ノイズ除去を目的として JPEG 圧縮によるフィルタ処理をすることで、画像分類器の識別結果の変動を解析する。JPEG 圧縮の品質パラメータの変化に伴う識別結果の違いを特徴成分として、単純なしきい値を用いることにより敵対的事例と正常な画像を高精度で識別することができた。

キーワード: 敵対的事例, JPEG 圧縮, CNN, 画像分類器

Detecting Adversarial Example Based on Changes of CNN Classification Results caused by JPEG Compression

AKINORI HIGASHI^{1,a)} MINORU KURIBAYASHI¹ NOBUO FUNABIKI¹ HUY H. NGUYEN²
ISAO ECHIZEN²

Abstract: While the research on image classifiers using CNN has been actively investigated, the study of adversarial example which can fool the image classifiers has been a potential threat in this research field. The adversarial example is created by intentionally adding small perturbations to an image so that an image classifier identifies a different class from the original class. It is not easy to detect adversarial example because the human eye cannot distinguish the difference from the original image. The objective of this study is to distinguish adversarial examples from natural images. The reason why an image classifier outputs an incorrect result is the tiny noise added to an original image. In this study, we analyse the fluctuation of the classification result by filtering with JPEG compression for the purpose of noise removal. By using a simple threshold, it is possible to discriminate adversarial examples and natural images with high accuracy from the characteristics of the fluctuations.

Keywords: Adversarial Example, JPEG compression, CNN, Image classifier

1. はじめに

コンピュータの計算能力向上に伴い、画像処理技術に深層学習を取り入れることが可能になってきた。畳み込みニューラルネットワーク (CNN: convolutional neural network) を用いた画像認識においては、代表的な VGG [1]

¹ 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology,
Okayama University

² 国立情報学研究所
National Institute of Informatics

a) p8m25ybd@s.okayama-u.ac.jp

や ResNet [2] などが発表されており、顔認証、自動運転、マルウェアの分類など多くの研究分野で応用されている。しかし、CNN を用いた画像分類器は敵対的事例に対して脆弱であることが報告されている [3]。敵対的事例とは、CNN を用いた画像分類器が誤分類するように意図的に造られた小さな摂動が加えられた画像のことである。この攻撃が懸念される例を挙げると、自動運転に使用される標識の分類器が、敵対的事例となった一時停止の標識を速度制限表示と誤認識したという報告がある。これから多くのシステムで CNN を用いた画像分類器が組み込まれることが予想されており、敵対的事例の存在は CNN を用いた画像分類器が実際に普及する上で、致命的な脅威となり得るため、敵対的事例の検知は喫緊の課題である。敵対的事例が報告されて以来、対抗する手法の開発が始まったものの、短期間で破られてしまうことが多く、日進月歩な研究分野となっている。

先行研究に敵対的事例に対する代表的な防御手法として、敵対的事例を認知できるように CNN を用いた画像分類器の強化を試みる Adversarial Training [4] と、敵対的事例を構築する上で必要な勾配の情報を隠す Gradient Masking [5] という 2 つの手法が挙げられる。前者は予防接種と同じ発想で、ネットワークモデルの学習時に敵対的事例も同時に学習させるという手法である。この結果、エラー率が 89.4% から 17.4% に下がったが、FGSM 攻撃 [4] に対してのみ防御できたデータであるため、ほかの攻撃手法に対しては脆弱であると考えられる。この手法で学習されたモデルは、敵対的事例だけでなく悪意のないランダムノイズにも反応するようになり、元々のモデルの識別精度が下がるという問題点を持っている。そのため、複数の起こり得る攻撃に対応させつつモデルの識別精度を維持することは非常に困難である。また、後者は攻撃の多くがモデルの勾配を利用して敵対的事例を探す手法であることから、モデルの勾配の情報を読み取れないようにする手法である。しかし、これに対するより高度な攻撃手法が開発されており、その一つに手元で操作可能な代理のモデルに学習させ、そのモデルの勾配を利用して敵対的事例を探す手法がある。あるモデルで作られた敵対的事例は、その他のモデルにおいても誤分類をさせることができるという性質を持っているため、元のモデルについての情報が分からずとも攻撃が可能となる。

本稿では、敵対的事例に加わったノイズを JPEG 圧縮で除去することによる画像分類の結果の変動を利用した敵対的事例の検知法を提案する。誤分類を引き起こす原因である敵対的事例に加えられたノイズは、人間の目で判別することが一般的には困難である。このノイズ削除を目的として JPEG 圧縮を施して、画像分類の結果の変動を本研究では解析する。予備実験において、代表的な 5 種類の攻撃手法を用いて原画像から敵対的事例を作成し、ノイズ削除を

目的として JPEG 圧縮を施したところ、画像分類の結果が正常なクラスに戻ることが確認された。そこで、複数候補の品質パラメータ (QF) で敵対的事例を JPEG 圧縮して、それらの分類結果の変動を解析したところ、その変動に顕著な特徴が現れることが分かった。この特徴を用いて、しきい値を用いた敵対的事例と正常な画像を分類する手法を提案した。提案手法の利点は、従来手法のようにネットワークモデルの変更は行っていないため、元々の識別精度を下げることはなく、対抗できる攻撃手法も限定されない点である。計算機シミュレーションにおいて、画像サンプルとその敵対的事例を使用して解析を行った結果、正常な画像の場合の品質パラメータによる識別結果の変動よりも、敵対的事例の場合での変動の方が大きいことが認められた。

2. 敵対的事例

CNN に基づく分類器が正しく分類できる画像に対して、人の目では判別できない程度のノイズを加えることで、人為的に分類器の判断を誤らせることができる。このようにノイズを加えて生成された画像を敵対的事例と呼び、その生成方法を分類器に対する攻撃と呼ぶ。

2.1 勾配に基づく攻撃

画像分類器によりあるクラスに分類される入力画像ベクトル \mathbf{x} に対して、ノイズベクトル $\boldsymbol{\eta}$ を加えることで、分類器が $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\eta}$ を異なるクラスに分類させることを考える。画像分類器のニューラルネットワークの訓練において、重みの更新に用いる損失関数の勾配を、入力への改変に利用することで、分類器が正しく動作しないようにこの $\boldsymbol{\eta}$ を求める攻撃法がある。

2.1.1 FGSM 攻撃 [4]

$\boldsymbol{\theta}$ を分類器モデルのパラメータとし、 y は \mathbf{x} に対する正解のクラスラベルとする。訓練のために使う損失関数を $J(\boldsymbol{\theta}, \mathbf{x}, y)$ として、この損失関数を \mathbf{x} で微分して損失が大きくなるように、微小な値 ϵ の正負の符号を調整したベクトルを $\boldsymbol{\eta}$ として扱う。

$$\boldsymbol{\eta} = \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\boldsymbol{\theta}, \mathbf{x}, y)) \quad (1)$$

ただし、 $\text{sign}()$ は正負の符号を返す関数である。このように $\boldsymbol{\eta}$ を求める手法が FGSM (Fast Gradient Sign Method) 攻撃である。FGSM 攻撃では、 $\boldsymbol{\eta}$ を計算する際に、指定したラベル y に対応する損失関数の値が増加するため y であると推論されにくくなり、学習器の認識を誤らせている。

2.1.2 LBFGS 攻撃 [6]

数値最適化において、非制限非線形最適化問題に対する反復的解法の一つである BFGS (Broyden-Fletcher-Goldfarb-Shannon) 法 [7] があり、多くの変数を扱えるように変更された L-BFGS 法や単純な矩形拘束を扱う BFGS-B 法があ

る。LBFGS 攻撃では、これらの手法を用いて入力画像と敵対的事例との距離を最小化するだけでなく、敵対的事例の分類クラスとターゲットとして与えられたクラスとのクロスエントロピを最小化している。入力画像に対して、誤認識させたいターゲットクラスに分類されやすくなるように η を求める手法となっている。

2.1.3 BIM 攻撃 [8]

BIM(Basic Iterative Method) 攻撃は、 x に対して要素ごとに切り抜いたクリップに FGSM 攻撃を繰り返し適用させる攻撃であり、特定のターゲットクラスに分類させることが可能である。

2.1.4 PGD 攻撃 [8]

PGD(Project Gradient Descent) 攻撃は、FGSM 攻撃を繰り返し適用させて、摂動された例を有効な例として複数回投影することにより、敵対的事例を生成する。

2.1.5 L1, L2 距離最小化攻撃

BIM 攻撃において、L1 距離や L2 距離を最小化させるように修正された攻撃である。

2.2 防御手法

Adversarial Training [4] では、学習プロセスの中で敵対的事例を生成し、教師データとして学習に利用することで、敵対的事例も正しく元のクラスに分類できるようにする手法である。学習フェーズにおいて、毎回その時点でのネットワークに対する敵対的事例を生成し、それに対して計算した損失関数を通常の損失関数と混ぜた新しい損失関数 $\tilde{J}(\theta, x, y)$ を用いる。

$$\tilde{J}(\theta, x, y) = \alpha \cdot J(\theta, \tilde{x}, y) + (1 - \alpha) \cdot J(\theta, x, y) \quad (2)$$

ただし、 α は混合割合を表すパラメータである。

入力に対する勾配を抑制する手法として、学習フェーズでの条件を変える Distillation [9] がある。この手法では、正しいラベルに関する情報だけでなく、正しくないラベルの確信度を学習に利用することで、敵対的事例による誤認識を防いでいる。

しかし、Adversarial Training や Distillation による手法を回避して、誤認識を引き起こさせる敵対的事例の生成手法が報告されている [10]。

2.3 敵対的事例の検出法

敵対的事例に画像処理を施すことで、CNN に基づく分類器の分類結果が変化しやすい性質を利用する Feature Squeezing と呼ばれる手法が提案されている [11]。その基本アイデアは、CNN が出力した推論値における各クラスの確信度のベクトルにおいて、通常分類結果といくつかの画像処理を施した分類結果の距離の大小で、与えられた画像が敵対的事例であるかを判別することである。文献 [11] 中では、画像のカラーのビット数削減と平滑化フィルタを

用いている。しかし、FGSM 攻撃と BIM 攻撃に対しては、あまり高い精度が得られていない。更には、その後に改良された LBFGS 攻撃や PGD 攻撃、L1, L2 距離最小化攻撃に対しても、カラーのビット数削減と平滑化フィルタでは高い精度を得られないと考えられる。

3. 敵対的事例の提案識別器

敵対的事例は、画像にノイズ（摂動）を加えることで、CNN に基づく画像識別器を誤検知させている。そのため、ノイズ除去フィルタを画像に対して適用させれば、その識別結果が変動する可能性が高い。本研究では、ノイズ除去フィルタとして、JPEG 圧縮を用いてその挙動を調べる。

JPEG 圧縮のアルゴリズムは、人間の目の特性を利用して、見た目の劣化を防ぎつつ圧縮を行っている。RGB 空間の 24 ビットのカラー画像に対して、YCrCb 空間の色成分に変換し、輝度成分と色差成分に分けて、 8×8 画素のブロック単位で、周波数成分を求める。ここで、ブロック内における緩やかな変化のある箇所は低周波成分を多く含んでおり、比較的画像の重要な情報を有する。一方、変化の激しい箇所は高周波成分を多く含んでおり、ノイズのようなパターンとなることから、その劣化は知覚されにくい。この特徴に基づいて、周波数成分は低周波成分はなるべく細かく量子化し、高周波成分になるほど粗く量子化するように、量子化テーブルが作成されている。JPEG 圧縮の圧縮率を変化させる品質パラメータ (QF) は、この量子化テーブル全体を一定の規則で変化させており、QF 値が小さくなればなるほど、高い圧縮率となり、画像が歪んでいく。

QF 値は、0~100 の値を取ることが可能であり、その値を減少させていくと、画像中の重要でない成分を削減する量が増えていく傾向がある。本研究では、この特徴を利用して、敵対的事例の生成において加えられたノイズ η を QF 値を変化させて、徐々に除去することを考える。ノイズの除去に伴って、誤って分類されていた分類結果が、正しい分類結果に戻ることが期待される。つまり、JPEG 圧縮を用いることで画像の特徴を保持しながら敵対的事例に含まれるノイズを除去することができるため、QF 値を変化させながら圧縮した場合、原画像と敵対的事例で識別結果の変動が異なると考えられる。

図 1 は、(a) に示す原画像に対して、FGSM 攻撃を用いて作成した敵対的事例の画像、および異なる QF 値で JPEG 圧縮した画像を示している。CNN を用いた画像分類器として VGG-19 を利用しており、原画像の分類結果は”Egyptian cat” であるが、敵対的事例では”tabby, tabbycat” と誤認識している。図 1(b) の画像を、QF = 75% として JPEG 圧縮した場合には、依然として”tabby, tabbycat” と誤認識しているが、QF = 65% の場合には元の”Egyptian cat” に正しく認識できるようになっている。この画像では、QF

値が 69~100 の場合には誤認識したが、QF 値が 68 以下の場合にはすべて正しく認識できた。画像によって、この特徴にはばらつきがあるが、勾配に基づく敵対的事例作成の手法を用いて作成された場合に、少なくとも JPEG 圧縮による認識結果に正常な画像とは異なる顕著な影響があることが予備実験により確認された。

この特徴を利用して、敵対的事例の識別器を、単純なしきい値を用いる手法により提案する。提案する敵対的事例の識別器では、敵対的事例かどうか分からない未知の入力画像に対して、異なる QF 値で圧縮して画像分類したそれぞれの結果と、無圧縮で画像分類した結果が一致した個数によって識別結果の変動を表し、設定したしきい値 T を下回った場合は敵対的事例と識別する。図 2 に、提案方式における敵対的事例を識別するための処理の流れを示す。

次章で示すシミュレーションでは、QF 値を 25 から 100 まで変化させて圧縮した原画像と敵対的事例をそれぞれ画像分類器で分類し、その識別結果と圧縮前の識別結果が一致した回数を求め、識別結果の変動を確認するとともに、各攻撃について同じしきい値を設定した場合の検知率を算出している。

4. シミュレーション結果

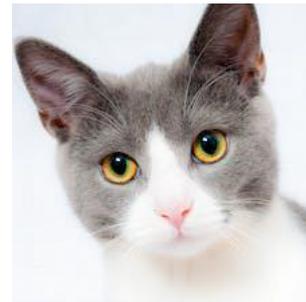
4.1 環境設定

本研究では、Foolbox*1を用いて原画像から敵対的事例を生成する。TensorFlow, Keras, Pytorch など様々なライブラリに対応している。攻撃手法は先の述べた勾配に基づく攻撃である FGSM 攻撃, LBFGS 攻撃, BIM 攻撃などをはじめ、多くの攻撃手法を使用することができる。これらの攻撃は、特定の CNN による画像分類器のモデルに対して、敵対的事例を作成することを意味する。Foolbox では、各ライブラリが提供する学習済みモデルが使用可能であるため、任意のモデルに対して敵対的事例を作成することができる。ただし、敵対的事例は必ずしもすべての画像で作成できるとは限らず、画像の特徴によって難しい場合もある。攻撃の手法に応じて、攻撃パラメータとしてノイズ付加重みを示す ϵ 値を適宜設定することで、敵対的事例の生成率を高めることができる。本実験では、 ϵ は Foolbox のライブラリ中にデフォルトで設定されてある値を用いた。

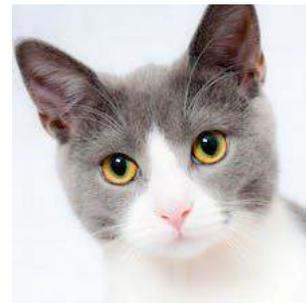
ImageNet*2では、CNN 分類器の学習用・検証用の画像とそのクラスを示すラベルが伴ったデータセットが提供されている。本研究で用いた画像のデータセットは、ImageNet の ILSVRC2012 で使用された検証用データである。ImageNet には数多くのクラスの画像があるが、本研究で使用する学習済みモデル VGG-19 は 1000 のクラスに分類することができ、使用するデータセットはこのクラス数に対応している。このデータセットのうち、各攻撃手法で敵



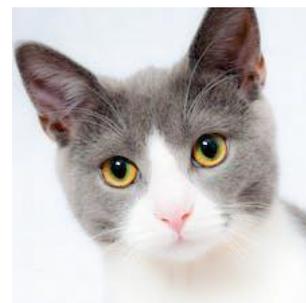
(a) 原画像
(分類結果: Egyptian cat)



(b) 敵対的事例
(分類結果: tabby, tabbycat)



(c) 敵対的事例の JPEG 圧縮後の画像 (QF=75%)
(分類結果: tabby, tabbycat)



(d) 敵対的事例の JPEG 圧縮後の画像 (QF=65%)
(分類結果: Egyptian cat)

図 1

対的事例が見つかった画像サンプルの中でランダムに 100 枚選んだものを検証で用いた。

4.2 識別結果の挙動

正常な画像と敵対的事例の圧縮前の結果と圧縮後の結果

*1 <https://foolbox.readthedocs.io/en/stable/>

*2 <http://www.image-net.org/>

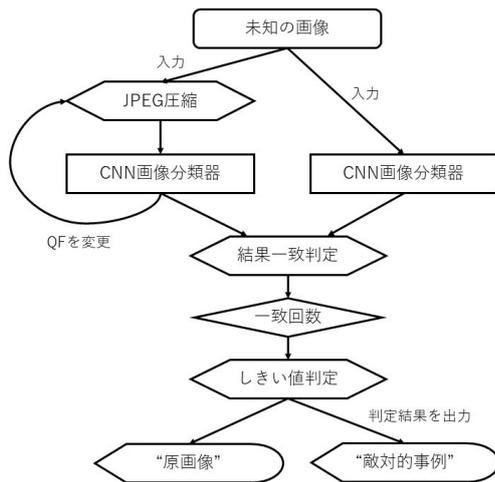


図 2 提案手法による敵対的事例の識別フローチャート

の一致数を示したグラフを図 3 に示す．今回試した 5 種類の攻撃すべてにおいて，正常な画像と敵対的事例との識別結果の変動の特徴に差異のあることが認められた．正常な画像では QF 値の変動に対して識別結果が不変なサンプルが多く見られた．それに対して敵対的事例では，今回の実験で用いた画像すべてにおいて，攻撃の種類に関わらず識別結果が大きく変動した．特に，L1, L2 距離最小化攻撃では複数の画像において圧縮前クラスとの一致数が高くなっている．しかしながら，概ね JPEG 圧縮による識別結果の挙動に特徴が現れることが確認できる．

各攻撃について設けるしきい値 T を 10, 15, 20 としたときの誤検知率を表 1 に示す．ただし，FP(False Positive) 率は正常な画像を誤って敵対的事例と判定した割合であり，FN (False Negative) 率は敵対的事例を誤って正常な画像と判定した割合を示している．また，TP(True Positive) 率と TN(True Negative) 率は，それぞれ $100 - FP$ と $100 - FN$ で求められ，正確度 (Accuracy), 精度 (Precision), 再現率 (Recall) は次の式で求められる．

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TN}{TP + FN} \quad (5)$$

L1, L2 距離最小化攻撃が予期通りの変動結果を示し，しきい値が $T = 10$ の場合には Accuracy が両方とも 98.5% と高い精度で検知できていることが分かる．しきい値が $T = 20$ の場合にこれらの攻撃の Accuracy は 94.5% と

表 1 しきい値 T による検知率 [%]

(a) $T = 10$

| | L1 | L2 | BIM | PGD | L1 | L2 |
|-----------|-------|------|-------|------|------|----|
| FP | 0 | 9 | 0 | 30 | 28 | |
| FN | 3 | 3 | 3 | 4 | 3 | |
| Accuracy | 98.5 | 94.5 | 98.5 | 83.0 | 84.5 | |
| Precision | 100.0 | 91.0 | 100.0 | 70.0 | 72.0 | |
| Recall | 97.1 | 96.8 | 97.1 | 94.6 | 96.0 | |

(b) $T = 15$

| | L1 | L2 | BIM | PGD | L1 | L2 |
|-----------|-------|------|-------|------|------|----|
| FP | 0 | 4 | 0 | 14 | 16 | |
| FN | 7 | 7 | 7 | 10 | 7 | |
| Accuracy | 96.5 | 94.5 | 96.5 | 88.0 | 88.5 | |
| Precision | 100.0 | 96.0 | 100.0 | 86.0 | 84.0 | |
| Recall | 93.5 | 93.2 | 93.5 | 89.6 | 92.3 | |

(c) $T = 20$

| | L1 | L2 | BIM | PGD | L1 | L2 |
|-----------|-------|------|-------|------|------|----|
| FP | 0 | 1 | 0 | 10 | 11 | |
| FN | 11 | 11 | 11 | 15 | 11 | |
| Accuracy | 94.5 | 94.0 | 94.5 | 87.5 | 89.0 | |
| Precision | 100.0 | 99.0 | 100.0 | 90.0 | 89.0 | |
| Recall | 90.1 | 90.0 | 90.1 | 85.7 | 89.0 | |

なった．一方，L1, L2 距離最小化攻撃の場合で予期通りでない敵対的事例の識別結果の変動を示したサンプル数があったように，しきい値が $T = 20$ においても Accuracy は 90% を下回った．

4.3 考察

多くの画像サンプルにおいて，どの攻撃手法でも原画像と敵対的事例の QF 値の変化による画像分類器の識別結果の変動に違いが見られたことから，JPEG 圧縮によって敵対的事例の検知が可能であることが分かる．

しかし，ノイズ除去フィルタとして JPEG 圧縮を用いた場合には，攻撃手法によって敵対的事例に含まれるノイズの影響を抑えることができないサンプルがあることも分かった．その原因として，L1, L2 距離最小化攻撃の両方で識別結果の変動が予期通りでない画像サンプルが共通している点から，ある特徴を持つ画像サンプルと攻撃手法の組み合わせによって，ノイズ除去にロバストな敵対的事例が生成された可能性が考えられる．

提案した手法では，ノイズ除去フィルタとして JPEG 圧縮を用いることで，モデルの強化・変更をせずに，敵対的事例の検知ができることが示された．本手法を前処理として用いれば，学習済みモデルの本来の識別精度を下げることなく，敵対的事例による誤分類を未然に防ぐことができる．更に検知精度を高めるためには，JPEG 圧縮だけでなく他のノイズ除去フィルタと併用するなどの対応が考えられる．

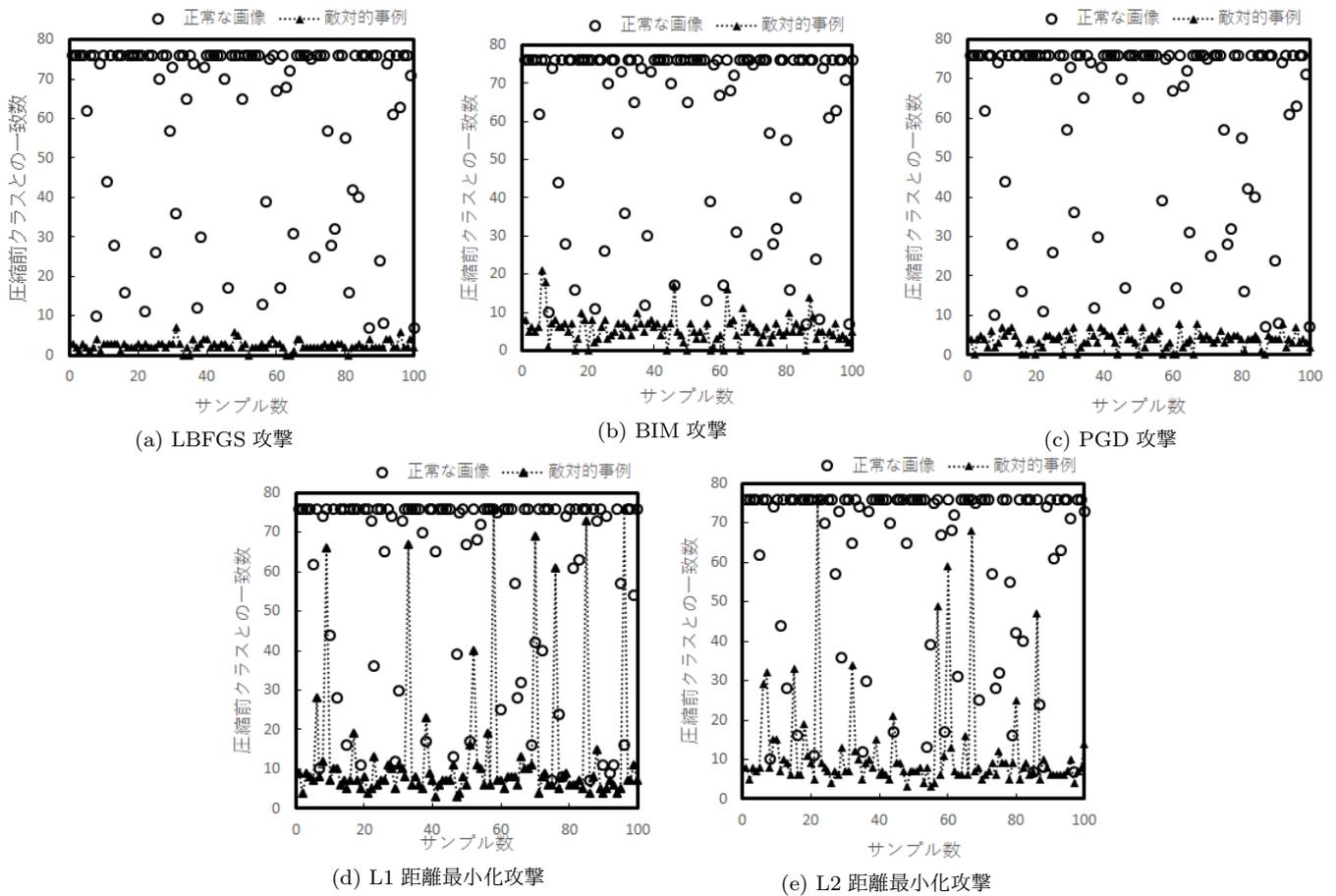


図 3 JPEG 圧縮の各 QF に対する分類結果の変化

5. おわりに

本稿では、JPEG 圧縮による CNN を用いた画像分類器の識別結果の変動を解析することで敵対的事例を検知する手法を提案した。今回の検証において、予期通りでない結果を示した攻撃手法があったように、他にも JPEG 圧縮の影響を受けにくい攻撃があると考えられる。従って、敵対的事例の検知精度を向上するためには、JPEG 圧縮以外の他のフィルタ処理で効果的なものを探ることが必要となる。また、検知したい画像を QF 値ごとに識別結果を出力する分類器も VGG-19 以外の学習済みモデルで同様の検証を行い、攻撃対象となるモデルが違う場合の識別結果の変動についても、確認する必要がある。本研究では最多ラベルのみを用いて識別結果の変動を観察したが、識別結果の確信度上位 5 つのラベルやその確信度の値を各 QF 値において調べることで、敵対的事例と正常な画像との識別結果の変動データを増やすことを考えている。将来的には、これらを含めた総合的な解析に基づいて敵対的事例の検知を行うニューラルネットワークの作成を検討したい。

謝辞 本研究は JSPS 科研費 JP19K22846 の助成を受けたものである。

参考文献

- [1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” Proc. ICLR2015, 2015.
- [2] G. Huang, Z. Liu, K.Q. Weinberger, and L. Maaten, “Densely connected convolutional networks,” Proc. CVPR2017, 2017.
- [3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I.J. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” Proc. ICLR2014, 2014.
- [4] I.J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” Proc. ICLR2015, 2015.
- [5] N. Papernot, P. McDaniel, I.J. Goodfellow, S. Jha, Z.B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” Proc. ASIACCS2017, pp.506–519, 2017.
- [6] P. Tabacof and E. Valle, “Exploring the space of adversarial images,” Proc. IJCNN2016, 2016.
- [7] R. Fletcher, Practical methods of optimization (2nd ed.), John Wiley & Sons, 2000.
- [8] A. Kurakin, I.J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” Proc. ICLR2017, 2017.
- [9] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” NIPS2014 Deep Learning Workshop, 2014.
- [10] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” Proc. IEEE Symposium Security and Privacy, pp.39–57, 2017.

- [11] W. Xu, D. Evans, and Y. Qi, “Feature squeezing: detecting adversarial examples in deep neural networks,” Proc. NDSS2018, 2018.