Vol.2022-DPS-190 No.50
Vol.2022-CSEC-96 No.50
2022/3/11

Deep One Class Neural Networkを用いて作成した偽の署名に対する攻撃耐性の評価

中嶋 優貴[1,a)]　新井 瑠月[1,b)]　宮島 将[1,c)]　金山 匠之介[1,d)]　西山 雄貴[1,e)]　宇田 隆哉[1,f)]

概要：筆跡を真似することで署名を偽造する攻撃があり，この攻撃によって，クレジットカードの不正利用などの被害が生じる可能性がある．既存研究に，アナログ手法で収集した筆跡の鑑定や，本人判定を入力リズムで行うもの等があるが，機械学習による判定やGANなどの機械によって生成された文字の鑑定は行っていない．本稿では筆跡のなりすまし防止の一環として，入力された文字が機械によるものか人間によるものかの判別を機械学習により行った．評価した結果，すべての被験者において分類精度が高いという結果になった．これは，筆跡の太さに依存していると考えられ，筆跡の太さを知らない攻撃者は筆跡のなりすましに成功しない可能性があることが分かった．

Evaluation of Tolerance against Fake Signatures by Deep One Class Neural Networks

1. はじめに

現在様々な書類などの電子化が日本で進められ，その中で電子はんこや手書きの電子サインなどのサービスが広りつつある．しかしこれらを元に単にConvolutional Neural Network（以下CNNと表記する）などの多クラス分類器を利用して筆跡鑑定などを行う場合，敵対的生成ネットワークなどをはじめとした生成アルゴリズムを利用して突破しようとされることが容易に想像できる．またそれらの分類をすることは難しくなることがトレーニングに足る十分な攻撃データを確保できないことなどから予想される．それだけでなく，新しいクラスへの対応などがしにくいなどの問題点も存在する．

本論文では，我々はone-stepで分類可能な1クラス分類器であるDeep One Class Neural Network（以下DONNと表記する）を利用した署名の分類を行い，ユーザの署名を機械学習の情報として使用せずに，悪意のあるユーザ（以下攻撃者と表記する）が生成したサインがどの程度特定のユーザに分類されるのか評価する．

2. 関連研究

2.1 入力リズムの違いで手書きの署名により本人確認する研究

松本らによる研究として，手書きの署名を入力リズムの違いで本人確認する手法の提案と評価がある[1]．概要としては，タブレットコンピュータを用いて署名を入力する際の時系列変化を波形化し，本人であるかなりすました他人であるかを分類するというものである．

この研究の判定方法では，実際の筆記画像を真似して筆記する攻撃者を想定しているが，文字画像データの入手が難しいことなどがあげられ実用的な攻撃手法ではないと考えられる．また，類似の研究として，上田らによる日本字筆跡から筆者を特定し，恒常性を評価するために変動エントロピーによる評価と，希少性を評価するためにパターンマッチング法による筆者照合を行った研究もある[2]．しかし，この研究にも実用化の際には机の高さや入力デバイスの位置などにより，ユーザが毎回同じ姿勢で書くことが難しいという問題がある．

[1] 東京工科大学
東京都八王子市片倉町1404-1, 192–0982
[a)] c0119225bb@edu.teu.ac.jp
[b)] c0119015fc@edu.teu.ac.jp
[c)] c011930665@edu.teu.ac.jp
[d)] c0119078b7@edu.teu.ac.jp
[e)] c0119238ec@edu.teu.ac.jp
[f)] uda@stf.teu.ac.jp

2.2 ニューラルネットワークを用いて手書き数字の筆圧から個人を識別する研究

前川らによる，ニューラルネットワークを用いて手書き数字の筆圧から個人を識別する研究がある [3]．この研究では，圧力センサ上で字を書くことで筆圧変化データを取得し正規化している．取り込んだ筆圧変化データに関する個人別特徴の学習と判断をニューラルネットワークを用いて行っている．

2.3 マニューシャマッチングと DP マッチングを組み合わせた筆跡認証に関する研究

西内らによる，マニューシャマッチングと DP マッチングを組み合わせた筆跡認証がある [4]．これはタブレット端末に筆記する形式の，手書きでの電子署名の認証における，既存技術を組み合わせた新たな認証方式の提案である．具体的に従来の筆跡認証に用いられた DP マッチングと，指紋認証に用いられるマニューシャマッチングとの組み合わせによって照合を行っている．

3. 提案手法

3.1 準備

2 章で述べたように，既存研究では機械学習による 1 クラス分類による判定や攻撃者によって生成された偽造文字画像と本人文字画像の分類及びその組み合わせは行っていない．本章では攻撃データの敵対的生成ネットワークを利用した生成方法及び，筆記データの収集方法を説明し，それらを分類するための解決策となる手法である DONN による分類について提案する．

DONN は multi-step の分類器である One Class SVM の代替にあたる Neural Network（以下 NN と表記する）であり，カーネル関数及び分離平面を one-step で獲得しようとする NN である．DONN は特徴抽出に学習済みの NN を用いることで学習コストを大幅に削減することが可能であり，これによりカーネル関数にあたる，特徴量空間への写像を獲得し，最後の全結合層で分離平面の獲得を行い本人か本人ではないかの 1 クラス分類を行う分類器である．今回利用するトレーニング済みの NN は，Imagenet でトレーニング済みの VGG16 にガウシアンノイズの入力を加えた全結合層を 3 層加えたものである．

まず，攻撃用の画像である偽造文字画像データを用意するために利用する敵対的生成ネットワークを用意した．攻撃者が攻撃者本人及びなりすます対象の手書き文字画像の情報無しで生成するとして，名前の情報のみから生成した偽造文字画像によりどの程度攻撃が通るかを調べる．今回利用する画像のサイズは 64*64*1 とし，偽造文字画像の生成には Deep Convolutional Generative Adversarial Network[5]（以下 DCGAN と表記する）を利用した．DCGAN の大まかな構造と画像生成の流れは以下の**図 1** の通りである．主にノイズデータを元に Generator により画像データを生成しそれを本物の画像で学習した Discriminator に渡すことで，本物か偽物かを判断し Generator は Discriminator を騙すように学習し，Discriminator は Generator に騙されないように学習することで本物らしい画像を生成できるように学習していく．今回は batch size を 32 とし，epoch は 5000 回として 5000 回目に出力された画像の中で文字として識別できるもののみを利用した．

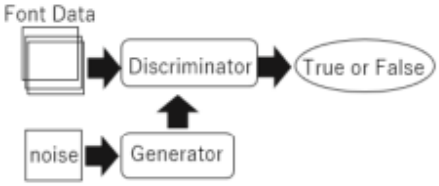

図 1 DCGAN

このような DCGAN を利用して一様乱数によって生成された 100 次元ノイズデータから文字を生成した．Generator と Discriminator のモデルの詳細はそれぞれ**図 2** の通りである．

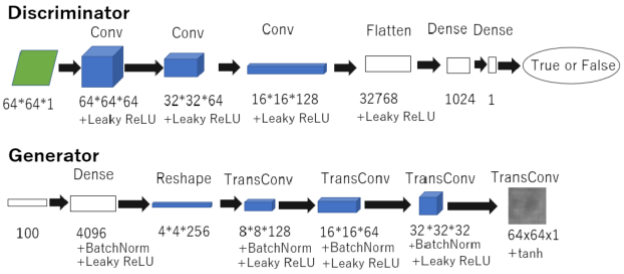

図 2 モデルの詳細

Generator では転置畳み込みを 4 回行い 4*4*256 から 2 倍ずつスケーリングを行い 64*64*1 にしている．これは今回利用する文字画像データを 64*64*1 としたためであるため，生成する画像のサイズによって任意に変更する必要がある．また学習の際，安定化のためにバッチ正規化を全結合と転置畳み込みの全ての層で利用している．また学習の際に画像のデータは 0 から 255 のグレースケール画像であったものを-1 から 1 の範囲に変換するため，活性化関数は出力層以外は勾配を消失させにくくするため 0 以下の領域でも傾きを持つ Leaky ReLU を利用している．一方出力層では tanh を利用している．

Discriminator では 64*64*1 から畳み込みを三回行い 16*16*128 にした後平坦化を行い，全結合層により生成された画像が騙されたか騙されていないかを出力する．Discriminator でも Generator と同様の理由から活性化関数に Leaky ReLU を利用した．また出力として利用するのは Generator で生成された画像のみである．

また生成する際に生成対象のデータとして利用したデータは文字画像作成機 [6] で作成されたフォントを画像にしたデータである．こうすることにより生成された画像は攻撃者本人の情報を含むことはない．実験には**表 1** のフォントを今回実験に参加した学生の名前の文字全てにおいて 5 枚ずつ利用した．

表 1 利用したフォント
Table 1 Used fonts.

	フォント
1	IPA P 明朝
2	たぬき油性マジック
3	はなぞめフォント
4	あくあ P フォント

3.2 筆記データの取得方法

今回実験をする上で実用上様々なデバイスで運用されることを想定して，ウェブサイトを VMware 上の仮想マシンである，Ubuntu 上に用意することで筆記画像の収集を行った．HTML で作成したキャンバスに文字の描画を行い，描画中の座標を Python CGI に送信し画像データの作成を行った．データ取得にあたって使用したプログラムは次のように作成した．

3.2.1 HTML

まず canvas を用いて横 1280，縦 256 で背景が黄色になる長方形を作成した．この長方形に入力のあった座標に描画を行いながら座標を取得する．文字の描画に際して，左クリックが押されている間だけカーソルの軌道が記録される機能とキャンバスを 5 等分する線を目安として描画する機能を実装した．これにより一文字ずつ文字画像を取得する時に比べ文字間のつながりの特徴も情報として抽出可能である．カーソルの座標が記録される機能は，左クリックが押されている時だけ座標を取得するように設定した．また取得した座標をリストにして，Python CGI に送信後にその座標を元に画像化する処理を行った．作成したキャンバスの HTML を**図 3** に示す．なお背景は書き込み可能な部分とそれ以外の部分を明確に示すために黄色の背景が用いられているが，これは画像化する際には排除される．

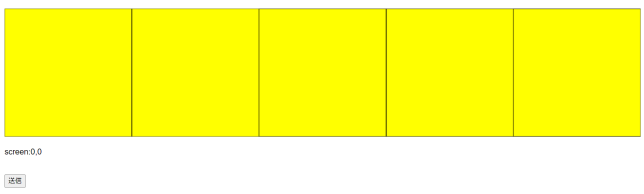

図 3 実際のキャンバス
Fig. 3 Actual canvas.

3.2.2 Python CGI

Python CGI 内では 1280*256 の長方形から 256*256 の正方形を 5 つに分けるために取り出して，それぞれを一文字として扱うため x，y 座標をそれぞれの文字画像に対応する範囲で x，y 座標を入力があった文字数に対応した 2 次元リスト内に格納するようにする．これに加えて，作成する画像サイズが 64*64*1 であるため，x 座標を 0〜255 になるように剰余演算を行った後に 4 で割り，y 座標は取得したデータを 4 で割る処理を行い Numpy を用いて 64*64*1 のグレースケール画像にしてから格納する．

4. 評価実験

3 章で提案した手法に関して評価実験を行った．本実験の参加者は東京工科大学内の学部生 7 名であり仮想マシン，あるいは設定済みのコンピュータを共有することでデータの収集を行った．評価実験では**図 4** のような DONN[7] を利用した．

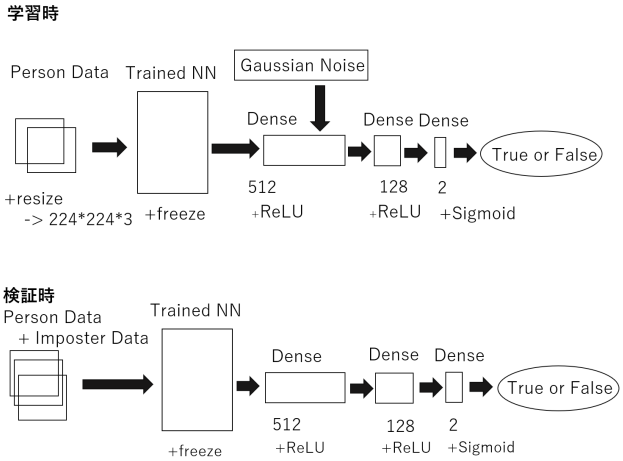

図 4 Deep One Class NN

なお，実験に利用した画像の枚数は次の通りである．7 名の被験者から，それぞれ約 20 枚の名前の文字数分の画像を作成し，それを Gimp を用いて水増しを行った．手書き文字，水増し画像，自動生成した文字それぞれの画像の例を**表 2** に示す．

表 2 使用画像の例
Table 2 Examples of images for evaluation.

手書き文字	水増し画像	偽造文字画像
		

評価実験では図 4 の DONN を利用した．

トレーニングデータとして筆記データのうち本人のもののみを利用する．これを VGG16 の特徴抽出能力に従って全結合層の 1 層目でガウシアンノイズと共に特徴量空間に写像し残りの 2 層で分離した結果，本人か本人ではないのかを判断している (図 4)．なお VGG16 部ではネットワークの重みは Imagenet をトレーニングしたときものをそのまま利用している．

テスト時にはガウシアンノイズを入力している部分を取り除き，本人が筆記した本人文字画像データと DCGAN よって生成された偽造文字画像データを同数入力して，それが本人のものかあるいは本人ではないかを判断して混同行列を出力することで性能を評価する．テストには 20epoch のトレーニングを行ったモデルを用いた．

また，7 人のトレーニング時に使用した本人文字画像と本人以外の文字画像の枚数をそれぞれ**表 3** に示す．テストには 5 分割交差検証を用いた．方法としては検証であるが，検証データがトレーニングに影響を与えないようにし，事実上のテストとなるよう工夫している．名前の先頭の文字だけの筆記データを 5 分割した上で水増しした画像との対応を取り，検証データにある画像の水増し画像がトレーニングデータに混在しないようにし，またその先頭の文字と同時に書かれた他の文字についても同様に検証データとして分離した．その際ちょうど 5 分割できないときは枚数の偏りがないように分離した．その後検証データと同数の偽造文字画像データを検証データに対応させた．

表 3 各被験者ごとの利用画像数
Table 3 Number of images for each examinee.

被験者	本人文字画像
1	360 枚
2	320 枚
3	340 枚
4	300 枚
5	475 枚
6	320 枚
7	300 枚

4.1 結果

評価実験の結果を以下に示す．今回の実験で分類を行った結果は混同行列として**表 4** に示す．混同行列の縦軸は DONN によって予測されたラベルを示し，横軸は実際のラベルである．表 4 から，一貫して偽造文字画像の分類の精度が高いという結果になった．その中で比較した場合，被験者 6 の本人文字画像の分類精度が最も低く，被験者 4, 7 の精度が最も高い．また偽造文字画像に関しては一貫して分類の精度が高かったことから，本人文字画像について分類精度の低かった被験者の筆記の特徴について注目していく．

表 4 分類結果の混同行列
Table 4 Confusion matrix of classification results.

被験者 1(平均)				被験者 1(標準偏差)			
		predict				predict	
		本人	偽造			本人	偽造
True	本人	0.988	0.012	True	本人	0.194	0.194
	偽造	0.000	1.000		偽造	0.000	0.000
被験者 2(平均)				被験者 2(標準偏差)			
		predict				predict	
		本人	偽造			本人	偽造
True	本人	0.994	0.006	True	本人	0.012	0.012
	偽造	0.000	1.000		偽造	0.000	0.000
被験者 3(平均)				被験者 3(標準偏差)			
		predict				predict	
		本人	偽造			本人	偽造
True	本人	0.998	0.002	True	本人	0.004	0.004
	偽造	0.000	1.000		偽造	0.000	0.000
被験者 4(平均)				被験者 4(標準偏差)			
		predict				predict	
		本人	偽造			本人	偽造
True	本人	0.996	0.004	True	本人	0.008	0.008
	偽造	0.000	1.000		偽造	0.000	0.000
被験者 5(平均)				被験者 5(標準偏差)			
		predict				predict	
		本人	偽造			本人	偽造
True	本人	0.988	0.012	True	本人	0.194	0.194
	偽造	0.001	0.990		偽造	0.000	0.000
被験者 6(平均)				被験者 6(標準偏差)			
		predict				predict	
		本人	偽造			本人	偽造
True	本人	0.980	0.020	True	本人	0.025	0.025
	偽造	0.000	1.000		偽造	0.000	0.000
被験者 7(平均)				被験者 7(標準偏差)			
		predict				predict	
		本人	偽造			本人	偽造
True	本人	0.996	0.004	True	本人	0.008	0.008
	偽造	0.000	1.000		偽造	0.000	0.000

4.2 考察

実験の結果から，共通して偽造文字画像の分類精度が高くなっている理由と，本人文字画像の誤分類が起きた被験者についてなぜそのような結果になったのかを考察する．

まず，精度が高くなった理由としては次のようなものが考えられる．今回 DCGAN を利用して攻撃データを作成したが，十分なフォント画像データがなく攻撃データとして用意した偽造文字画像が特定のフォントに依存した特徴を持っていた．また，データ取得プログラムには補完や筆圧が用いられていないため，線で描かれた文字というより，点の集合として描画されている．これにより，偽造文字画

像と本人文字画像を特徴量空間にマッピングした際大きく距離が開いてしまい，分離平面を厳密に取らずとも分類ができるようになったと考えられる．実際ユーザ間の分類ではそれほどの精度はなかった．これは画像取得プログラムの特徴を大きく受けていることを示唆している．

また，今回特徴量抽出に ImageNet でトレーニングした VGG16 を利用したが，224*224*3 の RGB 画像に対して最適化された構造と物体の画像で学習した NN のため，64*64*1 の NN を新しく設計しまたその重みのトレーニングも ETL 文字画像データセット [8] などの手書き文字のデータセットで行えば，特徴量の抽出がより厳密にできるのではないかと考えられる．

次に，誤分類が発生した被験者について，このような結果になった理由は次のようなものが考えられる．

被験者から収集したデータの一部の文字の特徴にバラつきが見られたことである．被験者 6 と 5 においては特にひらがななどの簡単な文字での大きさや位置，線同士の交差の有無などの筆記の際の特徴にバラつきが認められ，分離平面内に収まりきらないデータが発生したと考えられる．

さらに被験者 1 と 4 においては筆記された文字が角ばっていて大きいという一定の特徴を持っていたため，トレーニングした分離平面が厳しいものとなり，そこから 1 枚，2 枚程度はずれた画像が発生したものと考えられる．

以上の考察から，今回の目的として偽造文字画像の分類という目的としては DONN の性能は十分にあると言える．

また，筆跡を画像化する際の情報を知らない攻撃者は，たとえ筆跡のなりすましによる攻撃を行っても，類似した画像の生成が難しく成功しにくいという結果になった．これについては，筆跡を画像化する際の情報を攻撃者が知っている場合との比較実験を行えなかったため，どの程度の差があるか今後検証を行う必要がある．

5. おわりに

本論文では，DONN を利用した分類を行い，ユーザの署名と攻撃者が生成した署名を分類する評価を行った．

評価を行った結果，一貫して分類の精度が高いため，DONN の分類性能は十分発揮されたという結果になった．しかし一方で，DCGAN で生成した攻撃用の偽造文字画像データに不十分な点があることが，結果に影響を及ぼしている可能性を排除しきれていない．これは，十分なフォントデータの確保や実際の手書き文字データの不足により，DCGAN による自然な筆記された文字を生成する能力が低かったことが挙げられる．

今後は，攻撃者の文字を生成する技術がより高い性能を有していた場合や，DCGAN のトレーニングに人間の筆跡データセットを用いる，あるいはスタイル変換などを利用して自然な筆記画像データの生成に成功した場合にそれでも十分な分類が可能な性能があるか，その際，本人の文字画像の分類精度などがどのように変化するかの検証を行う必要がある．

また，画像収集プログラムについても，補完により線が自然につながるように描画されるようにしたり，筆圧を線の太さや濃さなどで表現したりすることで，紙に書かれた文字のように筆者の特徴が現れるようにすることも考えている．

今後の展望として，より正確な実験を行い，攻撃への耐性の再確認とユーザ間の分類が実用的であるかの検証を行っていきたい．

参考文献

[1] 松本憲幸，杉森眞二，浮川初子，浮川和宣：手書きの署名を入力リズムの違いで本人判定できる可能性の評価，情報処理学会，第 81 回全国大会講演論文集，Vol.2019, No.1, pp.391–392, (2019).
[2] 上田勝彦，松尾賢一：日本字筆跡の変動解析と筆跡個性に関する基礎的検討，情報処理学会研究報告，Vol.CH63, pp.1–6, (2004).
[3] 前川佳徳，井俣利昭，大西成也：ニューラルネットワークを用いた手書き数字の筆圧による個人識別，情報処理学会，第 51 回全国大会講演論文集，メディア情報処理，pp.161–162, (1995).
[4] 西内信之，吉田裕一，シムコフスキー・ピーター，サイード・カリード：マニューシャマッチングと DP マッチングを組み合わせた筆跡認証，電子情報通信学会技術研究報告，Vol.121, No.55, BioX2021-8, pp.36–40, (2021).
[5] Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, *Under review as a conference paper at ICLR 2016*, (2016). arXiv:1511.06434 [cs.LG].
[6] 文字画像作成機，`https://www.nin-fan.net/tool/string_image.html` (2022 年 01 月 03 日参照).
[7] Oza, P. and Patel, V. M.: One-Class Convolutional Neural Network, IEEE, *Signal Processing Letters*, (2018).
[8] etlcdb, `http://etlcdb.db.aist.go.jp/?lang=ja` (2022 年 02 月 07 日参照).