Vol.2021-AVM-112 No.3
2021/2/26

31×31 画像ロゴを表現できる深層学習電子透かし方式の一検討

衣川 晃弘 [1]　酒澤 茂之 [1]

概要：近年，DNN （ディープニューラルネットワーク)の著作権保護を目的とした，電子透かし技術が注目されている．画像などのマルチメディアコンテンツ向けの電子透かしと同様に，DNN 電子透かし技術では，DNN の本来のタスクに対する影響をできるだけ抑制しつつ，著作権者の情報を埋め込むことが求められる．既存の DNN 電子透かし方式では，モデルの重み係数に対する計算結果を表示するものや，加工を施した車の画像を飛行機と誤認識させるものなどがあるが，それらは非専門家に対する了解性の低さが課題である．そこで，本論文では DNN 電子透かしの検出結果が権利者のロゴ画像となることで，了解性を高める方式を提案する．ImageNet を分類できるモデルに対して，電子透かしを埋め込んだ結果，本来のタスクへの影響は少なく，電子透かし検出用の鍵画像を入力した場合に，ラベル値から 31×31 画素ロゴ画像を観測できることを示す．

キーワード：　電子透かし，著作権保護，DNN モデル

A Study on a Deep Learning Watermarking Method Capable of Representing 31x31 Image Logos

AKIHIRO KINUGAWA [†1]　SHIGEYUKI SAKAZAWA [†1]

Abstract: A trained DNN (Deep Neural Network) is a kind of intellectual property, and its copyright protection technology is required. Watermarking in DNN is one of solutions and there are many researches. Almost all the existing researches have a common problem on persuasion to non-experts, because some methods just output the watermark information by a “black-box” calculation, other methods insist that misclassification of a specific image means watermark existence. Hence, we propose a novel method which can represent a Logo image of copyright owner in an intuitive way. It is applicable for 1000 class image classification DNN, and 1000 elements of DNN outputs are organized as 31x31 pixels to show Logo image. Computer simulation results show that our watermarking method can clearly represent the Logo image when specific key images are inputted to DNN, while it has low impact on the image classification task for normal images.

Keywords: Watermark, copyright protection, Deep neural network

1. はじめに

現在，ディープニューラルネットワーク（DNN)の発展により車の自動運転や医療などの様々な分野に応用が広がっている．DNN の学習には，多量の学習データ，高性能な計算機器，DNN の専門知識などの多くのコストがかかる．その一方で，何らかの理由で流出した際の複製や複製したモデルを第三者に配布することは極めて容易である．又，複製された学習済みモデルを元に転移学習などで別の用途に特化した派生モデルを一から学習させるよりも少ない学習データで高性能なモデルを作成することができる．このことから，学習済み DNN モデルは知的財産であり，保護しなければならない．

学習済み DNN モデルの知的財産権については，知的財産戦略本部の「新たな情報財検討委員会 報告書」において言及されている[1]．一方，派生モデルと元のモデルとの関連性が分からなくなるため，類似性を立証が難しいと述べられている．そこで，本論文では元のモデルと複製されたモデルとの同一性を示すため電子透かしを利用し DNN モデルの著作権保護を行う．DNN 電子透かし技術では，DNN の本来のタスクに対する影響をできるだけ抑制しつつ，著作権者の情報を埋め込むことが求められる．

既存の DNN 電子透かし方式では，モデルの重み係数に対する計算結果や，加工を施した車画像を飛行機と誤認識させるものなどがあるが，それらは非専門家に対する了解性の低さが課題である．

そこで，本論文では DNN 電子透かしの検出結果を権利者ロゴ画像として表現することで，了解性の高い電子透かし方式を提案する．

2. 既存研究と問題点

2.1 既存研究

DNN 電子透かし技術は”white-box setting”と” black-box setting”に分類される．内部パラメータを全て閲覧可能な場合を white-box setting と呼び，入力データと出力データのみ閲覧可能な場合を black-box setting と呼ぶ．

white-box setting 版の電子透かしでは，DNN の選択し

1 大阪工業大学　情報科学部
Osaka Institute of Technology, Faculty of Information Science and Technology

たモデルパラータに対して電子透かしを埋め込む．したがって，電子透かしの検出は，透かし入りの重み係数を取り出してその性質を分析することで行う[2]．black-box setting 版では, 特別な加工を施した鍵画像群を意図的に誤認識(ex. 加工した車画像を飛行機と認識する)される[3]．加工としては，本来のタスクとは別の画像群に”WM”の文字パターンを白画素として重畳を行っている[4].

2.2 既存研究の問題点と解決のフレームワーク

既存研究のように加工した画像の誤認識を電子透かしとした場合，非専門家から見ればその誤認識を電子透かしではなく，単純な誤認識として認識される．既存研究の電子透かしの例を図 1 に示す．このような非専門家に対する了解性の低さが課題である．
そのため，本論文では出力される確率値を輝度とみなしロゴ画像として表現することで了解性を高める電子透かしを提案する．

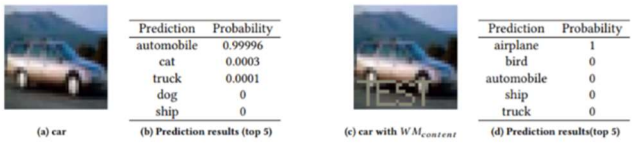

図 1 既存研究の電子透かしの例[3]

3. 提案方式

3.1 提案方式の概要

提案方式は既存研究に基づき画像分類の DNN モデルに対し電子透かしを埋め込み，black-box setting を用いて実験を行う．そのため，観測できるデータは入力データと出力される確率値のみである．タスクとしては 1000 種類の画像分類とし, 学習する画像データセットは ImageNet を用いる[5]．ImageNet とは，学習用のデータセット約 120 万枚と検証用データセット約 5 万枚から構成されるカラー写真の教師ラベル付き画像データベースである．解像度は 32×32 画素とする．ImageNet データセットを学習するにあたって，高精度を出すために Patryk らの論文を参考とした[6]. その論文では Wide Residual Networks（WRN)[7]が使われていたので，本論文でも採用する．
鍵画像を入力した場合の確率値を 31×31 の行列に並べ，輝度値と見なすことで画像として表現する．ロゴ画像作成のプロセスを 3.4 節で詳述する．

3.2 使用するデータセット

提案方式では, 本来のタスクである ImageNet とは別に電子透かし埋め込みの為のデータセットを構築する．

まず，このデータセット作成のため本来のタスクに類似している CIFAR-10 の学習データセットから 5 万枚の画像データを用意する．次に，この 32×32 画素の画像データに対し，視覚的に鍵画像であると容易に認識できるように星マークを白画素として重畳する. この画像を鍵画像と呼ぶ. 学習に用いる画像の例を図 2 に示す.

学習データセットは画像と，それが 1000 種類のクラスのどれに当たるのかを示すラベルのセットで構成されている．したがって，ImageNet のデータの場合は，正解のラベルの場所のみが高くなるように設定する．

一方，鍵画像では，ロゴ画像の作成に各ラベルの確率値を使用するため，表現可能なロゴを 31×31 画素とする．31×31 画素に固定するのは，1000 個のラベルの確率値を行と列が等しい行列にする場合の最大値を示すからである[a]. ロゴの背景部分の画素を 0，ロゴの前景部分を 1 とする．次に，出力される確率値を 31×31 の二次元配列一次元化して、961 元ベクトルにする. そこに、39 要素を加えて 1000 元ベクトルにする．そのあと，正規化している．本論文では,正規化後の前景部分のラベル値を 0.5 としている.このプロセスを図 3 に示す.

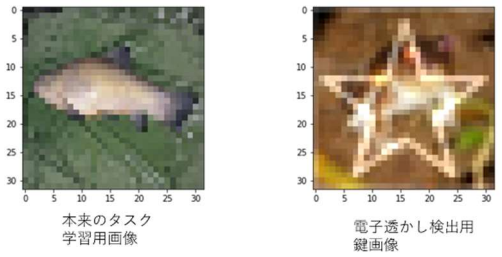

図 2 学習に用いる画像例

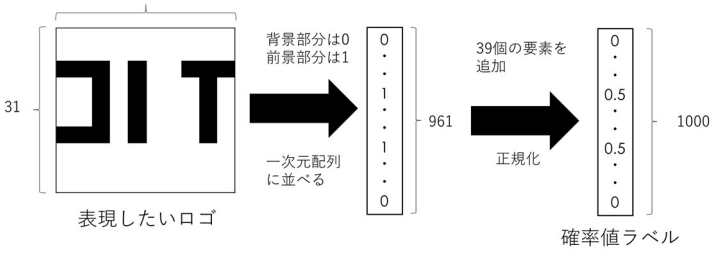

図 3 鍵画像のラベル作成

3.3 電子透かし埋め込み方法

提案方式では，本来のタスクである 32×32 画素の ImageNet の画像分類において高い精度を出すために WRN を使用する．

電子透かしの埋め込みはモデルの学習プロセスを通して行うが，モデルの損失関数などのモデルパラメータには変

a もちろん，横 50 画素 x 縦 20 画素としてもよい.

更を加えない.

学習に用いるデータセットは 3.2 節で述べた ImageNet の学習データセット 120 万枚に電子透かし埋め込み用のデータセット 5000 枚を加えたものである．なお，ImageNet のラベルは 1-hot ベクトル形式の 1000 元ベクトルである．

3.4 電子透かしの検出方法

電子透かしの検出は，電子透かし検出用のデータセットからの鍵画像をモデルに入力し，判定結果として 1000 分類の確率値を得る．これを，31×31 の行列に並べ，ロゴとして表現する為に 0 から 1 までの値をとる確率値に 255 を乗算することにより 0 から 255 の値をとる輝度値に変換する．よって，背景部分は 0，前景部分は 255 の値をとることで，背景部分は黒色，前景部分は白色に配色される．このプロセスを図 4 に示す.

ロゴ画像としての表現は，複数の鍵画像による判定結果の確率値を用いることにする．このとき，判定結果の確率値すなわちロゴ画像の輝度値を順次加算していく．輝度値が 255 を超えた場合, 255 に固定しオーバーフローを防ぐ.

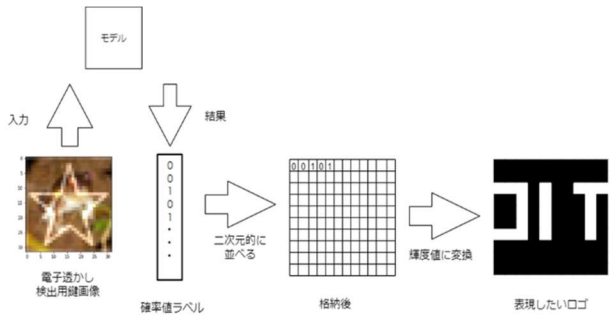

図 4 電子透かし検出プロセス

4. 実験と考察

4.1 無攻撃時の評価

権利者が電子透かしを埋め込んだ際，その電子透かしを検出できるか，そして，本来のタスクにどれだけ影響を与えるかを評価する.

電子透かしの検出については，ロゴ画像の背景画素と前景画素の輝度のコントラストを用いて視認性を評価する.

コントラスト= (L1 + 0.05) / (L2 + 0.05)　　(1)
(L1=前景部分の輝度値の最低値 , L2=背景部分の輝度値の最高値)

式 1 を用いてコントラストを求め，コントラストの値が 1 を超えていることが視認できる条件であると言える．この式で 0.05 を加算するのはゼロ除算を防ぐためである．一方の本来のタスクへの影響については，電子透かしが埋め込まれたモデルと ImageNet の画像分類のみを行う電子透かしが埋め込まれていないモデルとの精度比較を行う．精度比較のため Top1 精度と Top5 精度を用いる．まず，Top1 精度とは，対象ラベルがモデルのトップ予測である場合，モデルは与えられた画像を正しく分類したとみなされる手法である．一方，Top5 精度とは，対象ラベルがモデルのトップ 5 予測の 1 つである場合，モデルは与えられた画像を正しく分類したとみなされる手法である．これらの手法を用いる理由は Large Scale Visual Recognition Challenge（ILSVRC)における機械学習モデルのベンチマーク手法だからである.

コントラスト値を表 1 に，観測した電子透かしを図 5 に示す．また，学習する鍵画像の枚数を増やして実験を行った．条件をそろえるためロゴ画像作成に用いる鍵画像を 100 枚に統一している．学習する鍵画像が 5000 枚（Env1)と 10000 枚（Env2)とを比較すると，Env1 はコントラストが 1 を超え，明瞭に検出できるが，Env2 は前景部分にノイズを含んだロゴ画像になる．しかし，視覚的には図 3 に示すように，Env2 も Env1 と同様に権利者を理解することが可能である.

表 1 コントラスト

Env1	Env2
11.089589	0.4158416

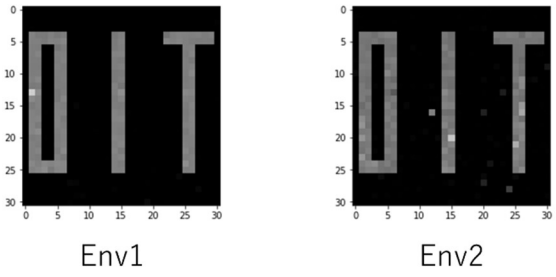

図 5 観測した電子透かし

次に，本来のタスクへの影響を表 2 に示す．透かしがないモデルと比較して 1%未満に収まっていることが分かる．そのため，電子透かしが及ぼす影響は少ないと考えられる.

表 2 精度比較

	透かしなし	Env1	Env3
Top1精度	52.5163	52.2694	52.2694
Top5精度	75.9204	75.5898	75.5898

単位%

4.2 攻撃時の評価

Env1 に対し攻撃を行い評価する. 想定される攻撃として

は

1. 電子透かし入り DNN モデルに対するモデル攻撃
2. 入出力する画像に対しての攻撃

の二つが考えられる．2 については，本実験では入力される鍵画像への攻撃のみを対象とする．

モデル攻撃としては, 不正利用者はタスクを ImageNet の判定から変更しないものと仮定し，モデルパラメータを変更する攻撃を行う．具体的には，ImageNet の学習用データセットの 10%である 12 万枚を用いた再学習である. 10%を用いる理由は，学習用データセット全て入手可能な場合はモデルの不正入手を行わずとも，ゼロから学習すればよいからである．

鍵画像への攻撃としては，ガウシアンノイズ付加とソルトアンドペッパーノイズ（s&p ノイズ)付加を行う．ガウシアンノイズとはノイズがとる値がガウス分布であるものである．s&p ノイズとはインパルスノイズとも呼ばれ，白と黒のピクセルがまばらに現れるノイズである．鍵画像への攻撃の例を図 6 に示す．

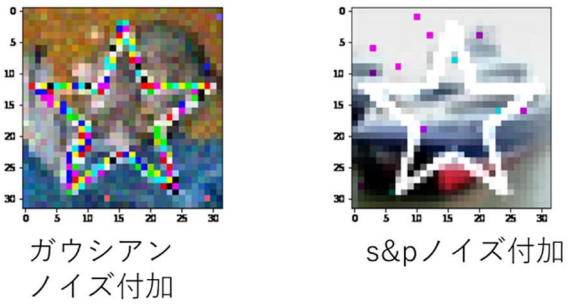

図 6 鍵画像への攻撃の例

各攻撃でのコントラストの結果を表 3，モデル攻撃に対する電子透かしを図 7，鍵画像への攻撃に対する電子透かしを図 8 に示す．

表 3 各攻撃でのコントラスト

ガウシアンノイズ	s&pノイズ	モデル攻撃
6.488	1.063	2.324

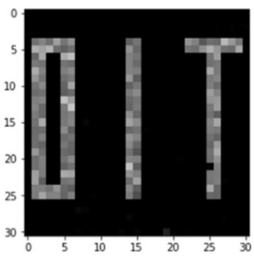

図 7 モデル攻撃に対する電子透かし

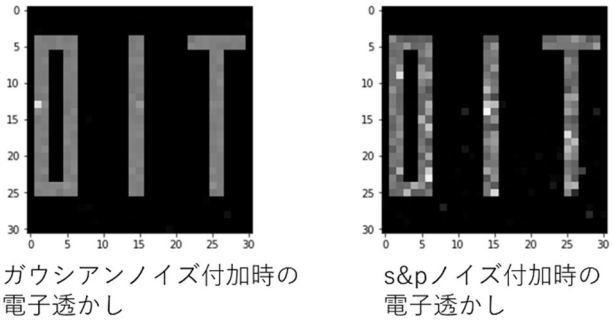

図 8 鍵画像への攻撃に対する電子透かし

まず，モデル攻撃の耐性の評価について述べる．前景部分の輝度値は安定しないが，OIT の文字が確認でき，コントラストの値も 1 を超えている．そのため，モデル攻撃に対しては耐性があると言える．

次に，入力する画像に対しての攻撃についての評価について述べる．ガウシアンノイズを施した鍵画像の場合のロゴ画像は無攻撃時コントラストの値より多少劣るが，OIT の文字は明瞭に視認可能である．そのため，ガウシアンノイズには耐性があると言える．一方，s&p ノイズを施した鍵画像の場合のロゴ画像は無攻撃時コントラストの値よりかなり劣るが，OIT の文字は視認可能である．しかし，文字部分の輝度値が安定せず，モザイクを掛けたようなロゴ画像になる．また，本来背景部分の一部が前景部分に干渉している. ノイズを含むが OIT の文字が確認できる為, s&p ノイズに耐性があると言える．

5. おわりに

今回は，学習済み DNN モデルに対する電子透かし技術を提案した．既存研究の問題点である非専門家に対する了解性の低さを改善する為，ロゴ画像として電子透かしを表出して，評価を行った．提案方式は，学習済み DNN モデルの要求条件である 1） 本来のタスクに対する影響をできるだけ抑制すること，2） モデルの加工に対して透かしが消えないこと，3）透かしの読み出しは，許可された者のみが可能であることを満たしている．2 についてはモデル攻撃や透かし検出のための鍵画像にノイズを掛ける攻撃に対して，十分な耐性があることを示した．

謝辞 本研究は JSPS 科研費 JP18K11309 の助成を受けたものである．

参考文献

[1] 知的財産戦略本部 新たな情報財検討委員会 ：
“新たな情報財検討委員会報告書” (2017)
[2] Uchida Yusuke, et al. "Embedding watermarks into deep neural networks." Proceedings of the 2017 ACM on International

Conference on Multimedia Retrieval, pp. 269-277, ACM, 2017.
[3] J Zhang, Z Gu, J Jang, H Wu, ” Protecting Intellectual Property of Deep Neural Networks with Watermarking”, ASIACCS '18, pp 159–172, 2018.
[4] 酒澤 茂之,” 10x10 画素ロゴを表現可能な深層学習電子透かし方式” 電子情報通信学会信学技報 EMM2020-1, pp.1-6, 2020.
[5] IMAGENET http://www.image-net.org 2021/1/28 アクセス
[6] P Chrabaszcz, I Loshchilov, F Hutter,” A Downsampled Variant of ImageNet as an Alternative to the CIFAR datasets”, arXiv:1707.08819
[7] S Zagoruyko, N Komodakis,” Wide Residual Networks”, arXiv:1605.07146