

電子透かしを用いた学習済みモデルの世代管理に関する一検討

松本晴輝¹ 酒澤茂之¹

概要： DNN（ディープニューラルネットワーク）の需要が拡大している。この DNN の需要の拡大に応えるにはモデルを初期状態から構築するのではなく、既存の学習済みモデルから個別の案件に調整した派生モデルを生成することが効率的である。DNN モデルの権利管理においては、これらの派生モデルに対応可能な電子透かしの技術が必要である。しかし、現時点での研究動向は違法な利用者による不正利用への対策が主であり、正当な利用者による派生モデルの世代管理等の二次利用に適した電子透かし技術は未検討である。そこで本論文では、初期検討として、学習済みの親世代のモデルに子世代の情報を追加的に埋め込む実験的検証を行ったので報告する。

キーワード： DNN, 二次利用, 電子透かし

1. はじめに

現在、ディープニューラルネットワーク (DNN)の発展により様々な分野に应用が広がっている。この DNN の需要の拡大に応えるにはモデルを初期状態から構築するのではなく、既存の学習済みモデルから個別の案件に調整した派生モデルを生成することが効率的である。一方で、学習済みモデルには知的財産としての側面があり、その著作権保護のためにモデル内に電子透かしを埋め込む技術が注目されている[1]。しかし、現時点での研究動向は違法な利用者による不正利用への対策が主であり、正統な利用者による派生モデルの世代管理等の二次利用に適した電子透かし技術は未検討である。

そこで、本論文では学習済みモデルの原型に対して、著作権者の電子透かし情報を埋め込み、その親世代モデルを基に二次利用者が派生モデル(子世代モデル)を作成した際に、派生モデルから元の著作権者と派生モデルの作成者を識別できるようにすることを最終目標とする。すなわち、派生モデル作成時にすでに埋め込まれている電子透かしの情報とは別の二次利用者の情報を埋め込み、この電子透かしの情報を検出する技術が必要である。この最終目標の初期検討として、親世代の透かし入りのモデルに対して、子世代の透かしを入れ実験を行う。

本論文では、電子透かし方式として既存研究[2]の方式を利用する。既存の研究では、特別な加工を施した画像を集めてデータセットを作り、本来のタスクを行うためのデータセットと合わせて学習しモデルを作成する。そのモデルに特別な加工を施した画像を入力すると特別なロゴが検出できるという研究がある。しかし、この既存研究では、埋め込まれている電子透かしが上書きされた場合元の電子透かしが消えてしまっは意味がないという課題がある。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という研究に対して、さらに上から別の電子透かしを埋め込む。そして、元の電子透かしが消えないこ

とで学習済みモデルの世代管理をする方式を提案する。

2. 既存研究と問題点

既存研究[2]では既存研究[1]の問題点であった非専門家からの了解性の低さに焦点を当てた研究である。本来のタスクとしては、100 種類の画像分類としており、データセットは Cifar-100 を使用している。本来のタスクを行うための Cifar-100 のデータセットと、特別な加工を施した鍵画像だけで構成されたデータセットで通常の学習を行う。特別な加工を施した鍵画像を入力し、その鍵画像に対して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、モデルの権利者を見ただけで直感的に理解できるという研究である。既存研究[2]の特別な加工を施した鍵画像は図 1 示す。この研究の問題点は、モデルに初期作成者の電子透かしを埋め込んでいても、二次利用者の情報を別の電子透かしとして追加で埋め込んだ場合に初期モデル作成者の電子透かしが消えてしまっはモデルの世代管理ができないという点である。

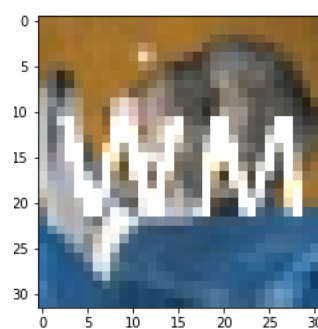


図 1 既存研究[2]の特別な加工を施した鍵画像

¹ 大阪工業大学
Osaka Institute of Technology

3. 提案方式

3.1 提案方式の概要

提案方式は既存研究[2]に基づき black-box setting を用いる。black-box setting とは、入力データと出力データのみ閲覧可能な場合のことである。そのため、入力データと出力されるラベル値のみが観測できる。タスクとしては 100 種類の画像分類とし、学習する画像データセットは Cifar-100 を用いる。既存研究[2]で Wide Residual Networks(WRN)が使われていたので、本論文でも使用する。使用するモデルは WRN を使用したすでに電子透かしが埋め込まれているものを使用する。

※以降初めから埋め込まれている電子透かしを「透かし A」後から埋め込む電子透かしを「透かし B」と呼ぶ。

透かし A の鍵画像が入力された場合と透かし B の鍵画像が入力された場合とでそれぞれ、各ラベルの確率値を 10×10 のマトリクスに並べる。確率値に 255 を掛けて輝度値に変換し異なるロゴ画像を出力するようにする。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

本実験では、100 種類の画像分類のモデルに対し電子透かしが埋め込まれたモデルを使用する。このモデルは Cifar-100 の学習データセット 5 万枚に透かし A の鍵画像で構成された、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。この親世代モデルに対して元のタスクを行うための学習に使用した Cifar-100 の画像 10000 枚と、透かし A の鍵画像 500 枚と、透かし B の鍵画像 500 枚の合計 11000 枚で fine-tuning を行う。この工程のフローを図 3 に示す。

透かし A に使用されている鍵画像は、Cifar-10 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 WM の文字を白画素として重畳したものである。透かし A の鍵画像は図 3 に示す。

透かし B に使用している鍵画像は ImageNet 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 MH の文字を白画素として重畳したものである。透かし B の鍵画像は図 4 に示す。鍵画像のラベルについては、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 10×10 画素とする。 10×10 画素に固定するのは、100 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。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 10×10 の二次元配列に格納し、複数の出力結果の総和が 1 になるようにする。本論文では、前景部分のラベル値を 0.05 としている。

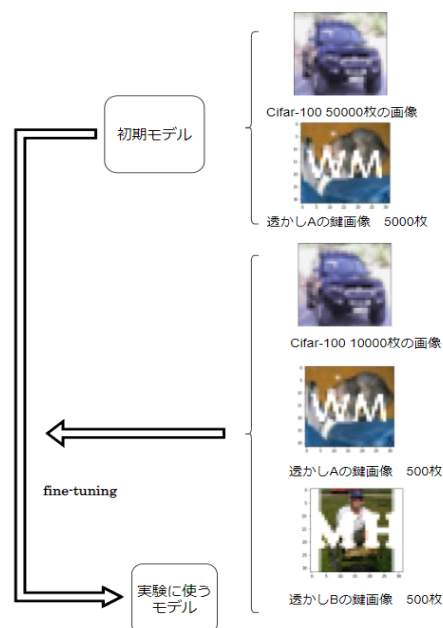


図 2 モデル作成フロー

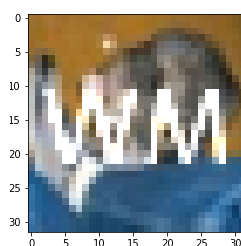


図 3 透かし A の鍵画像例

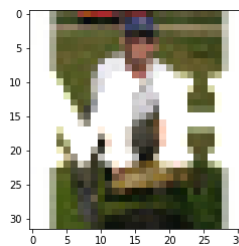


図 4 透かし B の鍵画像例

3.3 電子透かしの埋め込み方法

実験では、本来のタスクである 32×32 画素の Cifar-100 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。learning rate は $1e-7$ とする。本実験で使用するモデルは、3.3 節で述べた Cifar-100 の学習データセット 50000 枚に電子透かし埋め込み用のデータセット 5000 枚を加えた学習済みモデルに対し、Cifar-100 の学習データセット 10000 枚と透かし A の鍵画像 500 枚と透かし B の鍵画像 500 枚で fine-tuning を行ったモデルである。

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透かし A の鍵画像、透かし B の鍵画像それぞれをモデルに入力し、判定結果として 100 分類の確率値を得る。これを、 10×10 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、背景部分は黒色で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 100 分類の確率値を CSV ファイルに格納し，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．

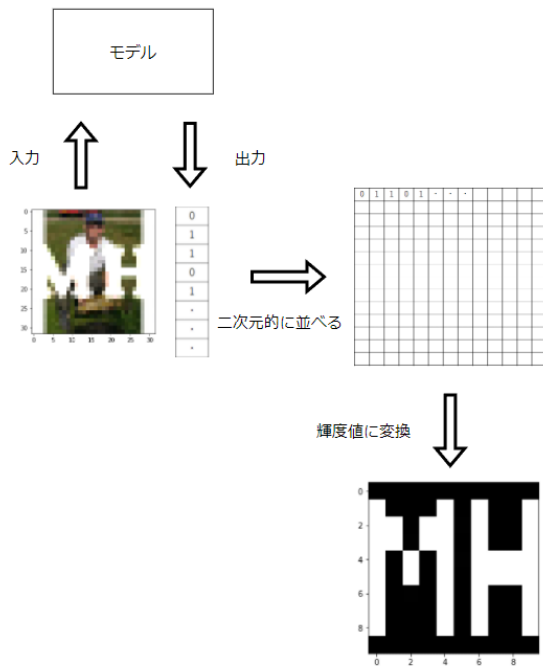


図5 透かし検出のフローチャート（透かし B）

4. 実験

4.1 実験概要

二次利用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，初期モデル作成者が埋め込んだ電子透かしにどのような影響を与えるかを評価する．加えて，元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．

$$\text{コントラスト} = (L1 + 0.05) / (L2 + 0.05) \quad (1)$$

(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，透かし B が埋め込まれた子世代のモデルと，透かし A だけが埋め込まれた親世代のモデルの精度を比較する．

4.2 実験結果

コントラスト値を表 1 に，観測した電子透かしを図 5，図 6，図 7 に示す．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．

表 1 コントラスト

	図 6	図 7	図 8
コントラスト	38.6	53.8	97.1

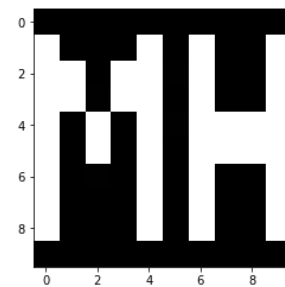


図 6 透かし B のロゴ

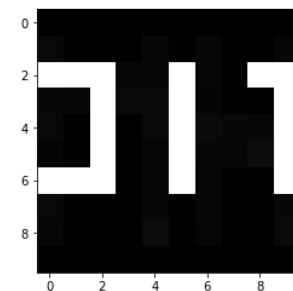


図 7 透かし B を埋め込んだ後の透かし A のロゴ

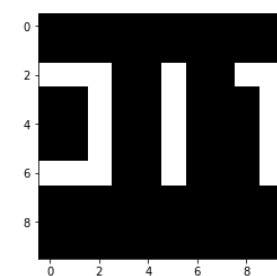


図 8 透かし B を埋め込む前の透かし A のロゴ

次に，元のタスクへの影響を表 2 に示す．親世代と子世代のモデルのタスク精度を比較すると，差は 2.73%であった．

表 2 精度比較

	透かし A	透かし B
精度	0.7392	0.7119

5. 考察

5.1 透かし B が本来のタスクに及ぼす影響

透かし B が及ぼす影響は、透かし A のみモデルの精度と比較して、約 2.7%の劣化が確認できた。電子透かしを埋め込んだことにより、精度が劣化する原因は本来のタスクつまり Cifar-100 分類に関係がない画像を学習させることだと考える。電子透かしを合計で 2 回埋め込んでいることもあり精度の劣化は少ないとは言えない。

5.2 電子透かしの検出

透かし B を埋め込み検出を試みると、問題なく検出できているのが視覚的にもわかる。図 7、図 8 のコントラストに差が出ている事実を視覚的にわかりやすくするため、明瞭度の調整など加工を行った画像の比較を図 9 に示す。図 9 を見るとわかる通り、透かし B を埋め込んだことにより、透かし A にコントラストの差が出ていることが視覚的にもわかる。さらに、図 7 を加工した画像を注視すると薄っすらとではあるが、透かし B の表現ロゴである MH が見える。

次にコントラストを比べる。条件をそろえるためロゴ画像作成に用いる鍵画像を 100 枚に統一しているが、コントラストの値はすべて異なっていた。コントラストが最も高いのは図 8 である。図 8 のロゴ画像はコントラストの値は他 2 つと比べ最も高いが、視覚的に確認した場合には他の 2 つのロゴとは違いを感じられない。OIT の文字は明瞭に視認可能である。図 6、図 7 のロゴもコントラスト値が 1 を大きく超えていて、視覚的にも十分視認可能であるという結果になった。図 7、図 8 にコントラストの差が出ている原因は透かし B を埋め込んだからと考えられる。本実験では透かしの埋め込みは合計 2 回であるが、モデルの世代管理をするとするとこれ以上に電子透かしを上書きすることになるため 1 番初めの電子透かしに大きな影響が出ると考えられる。

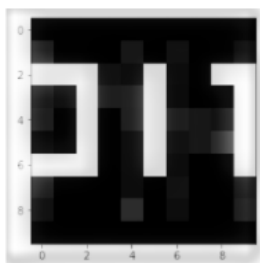


図 7 を加工した画像

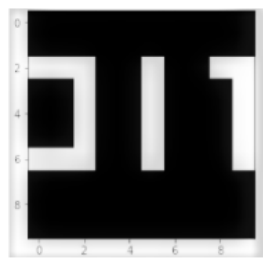


図 8 を加工した画像

図 9 加工した画像の比較

6. おわりに

本論文では、既存研究の問題点である、学習済みの親世代モデルに対し、子世代の情報を電子透かしとして埋め込み、親世代と子世代の情報を検出することで、学習済みモデルの世代管理をするために実験を行ってきた。実験の結果から、後から電子透かしが埋め込まれた場合に元の電子透かしが問題なく検出できることを示した。

しかし、本実験では透かしの埋め込みは合計 2 回であるが、モデルの世代管理をするとなるとこれ以上に電子透かしを上書きすることになるため 1 番初めの電子透かしに大きな影響が出ると考えられる。この初めの電子透かしを守るための方法も考える必要がある。さらに、電子透かしを埋め込むたびに本来のタスクの精度は少しずつ落ちていくと考えられる。学習時に本来のタスクのためのデータセットも加えて学習していても精度は劣化していたので、この点についても対策する必要がある。加えて、本実験は 10×10 の画素でロゴを表現したが、より多い画素数で表現することで多くの種類の文字や絵を表現できるのでこれについても今後研究する必要がある。

謝辞 本研究は JSPS 科研費 JP21K11896 の助成を受けたものである。

参考文献

- [1] J Zhang, Z Gu, J Jang, H Wu, "Protecting Intellectual Property of Deep Neural Networks with Watermarking", ASIACCS '18, pp 159–172, 2018.
- [2] 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” vol. 120, no. 42, EMM2020-1, pp. 1-6, 2020 年 5 月
- [3] S Zagoruyko, N Komodakis, "Wide Residual Networks", arXiv:1605.07146
- [4] 酒澤, “深層学習モデルに対する視覚復号可能な電子透かし方式”, PCSJ/IMPS2019, P-3-01 (2019)