CK-006

重み一定符号を用いた DNN 電子透かしの検出法
Detction of DNN Watermark Encoded by Constant Weight Code

安井 達哉[1)]　アサド マリック[2)]　栗林 稔[3)]
Tatsuya Yasui　Asad Malik　Minoru Kuribayashi

1　はじめに

近年，計算機の性能向上やビッグデータによって深層学習 (DNN) 技術の研究開発が盛んに行われている．深層学習では，モデルのアーキテクチャに対する重みパラメータを計算機リソースと大量の学習データで学習する．そのため，生成されたモデルには大きな価値があるとして，DNN モデルの権利を電子透かしを用いて保護する研究が行われている．電子透かしとは，画像や音楽などのデジタルデータに対して秘密裏に情報を埋め込む技術である．画像や音楽の場合は，元のコンテンツの品質が劣化しないように情報を埋め込む一方で，DNN 電子透かしの場合は，モデルの精度が可能な限り低下しないように埋め込む必要がある．

Uchida らの研究 [1, 2] では，畳み込みニューラルネットワークの特定の層に埋め込みを行う手法が提案されており，初めて DNN モデルに対して電子透かし (DNN 電子透かし) 技術を適用した研究である．この手法では，オリジナルタスクの損失関数とは別に埋め込みのための損失関数を導入し，秘密行列と重みパラメータの行列積が透かし情報のベクトルに近づくように更新を行う計算を行っている．学習は，２つの損失関数の和を最小化するように重みパラメータを更新する．Rouhani らの研究 [3, 4] では，Uchida らの手法 [1, 2] で紹介されたこれらの手法の特徴ベクトルの選択を改良している．さらに，Choromanska らの論文 [5] や Dauphin らの論文 [6] では，ほぼ全ての局所解は大域解と同等であることが報告されている．経験的に見ても，より深いモデルや大きなモデルでは損失値が近い値を取るため，大域解の代わりに局所解を用いても十分であることが示されている．この特性を利用して，いくつかの選択された重みパラメータの値に電子透かしを埋め込む手法が提案されている [7, 8].

重みパラメータに埋め込むタイプの DNN 電子透かしは，意図的であるか否かに関わらず重みパラメータが変更されることに対してロバストである必要がある．重みパラメータが変更される一例としてモデルの刈り込み (プルーニング) がある．これは，DNN モデルの実行にかかる計算コストを削減するために，精度を落とさずに冗長なニューロンを刈り込むというものである．プルーニングの目的は，損失への寄与が小さい，つまり重要度の低い重みパラメータを DNN モデルから削除することである．もし，DNN 電子透かしがそのような重要度の低い重みパラメータに埋め込まれていた場合，プルーニングによって簡単に消失または変更されてしまう．したがって，DNN 電子透かしは，元のタスクにおける精度を保証しつつ，プルーニングのような攻撃に対してロバストであることである．Uchida ら [1, 2] は，65%の重みパラメータを刈り取るプルーニング攻撃を受けても，電子透かしが消失しないことを実験的に示している．

我々の従来研究 [9] では，重み一定符号 (CWC)[10] を用いることで，重みパラメータへのプルーニング攻撃に対してロバストな電子透かしを実現している．電子透かしの埋め込み操作時に 2 つの閾値を設け，プルーニング攻撃によるロバスト性を制御している．従来研究では，DNN モデルから選択された重みパラメータに CWC 符号語を埋め込む操作および，選択された重みパラメータから CWC 符号語を抽出する操作について具体的な手法を提案している．しかし，選択された重みパラメータに透かし情報が存在するか否かの確認，すなわち電子透かしの検出については言及はされているものの，その具体的な手法については検討がされていなかった．電子透かしが埋め込まれていない場合であっても，系列が抽出できるため，事前に DNN 電子透かしの検出ができれば誤った CWC 符号語の抽出が行われることがなくなる．

本研究では，従来研究において実現した CWC を用いたプルーニング攻撃に耐性のある DNN 電子透かしに対して，重みパラメータが従う分布を理論的に解析し，電子透かしの検出を行う検出器を設計する．DNN モデルの重みパラメータは，初期値として一様分布に従うランダムな系列であることを仮定し，学習が進んだ後においても概ねその分布は等価であることを前提とする．一方で，透かし情報の埋め込みのために選出される重みパラメータは，学習時に埋め込み処理が行われるため，その分布は CWC 符号語に応じて偏りが生じる．提案手法では，その偏りをうまく利用して，選択された重みパラメータが CWC 符号語であるかランダム系列であるかを判別する．シミュレーションにより，その仮定の妥当性と理論的に導出した期待値の正当性を確認することができた．

2　関連技術

本章では，DNN 電子透かしと DNN 電子透かしの脅威の一つであるプルーニング攻撃，および従来研究で使われている重み一定符号について述べる．

2.1　DNN 電子透かし

DNN 電子透かし技術は，大きくホワイトボックス電子透かしとブラックボックス電子透かしの 2 種類に分類することができる [4]．ホワイトボックス電子透かしは，内部の構造やパラメータが公開されるため，重みパラメータや活性化関数に対して透かし情報を埋め込むことができる．一方で，ブラックボックス電子透かしでは，内部の構造やパラメータが秘匿されているため，入力を与えた DNN モデルの出力から透かし情報を検出する．ブラックボックス電子透かしでは，基本的に最終層の出力にしかアクセスできないが，与えられた入力に対して意図的に間違った出力をするようにモデルを学習し，それを電子透かしとみなす研究も行われて

1)　岡山大学大学院自然科学研究科
2)　Department of Computer Science, Aligarh Muslim University, India
3)　岡山大学大学院学術研究院自然科学学域

いる [11, 12]．最初のホワイトボックス電子透かしは，Uchida らによって発表された手法 [1] であり，畳み込みニューラルネットワーク (CNN) の畳み込み層の重みパラメータに対して電子透かしを埋め込む手法である．埋め込み方法として，埋め込み損失関数を用いて透かしの埋め込みとモデルの学習が同時に行われる．この手法は後に，Rouhani ら [3] によって改良され，DNN モデルを不正に利用するユーザを追跡するための電子指紋技術へ応用の検討がされた．

2.2 DNN モデルのプルーニング

多くの重みパラメータを有する DNN モデルから冗長な重みパラメータを取り除くことで，メモリ容量の削減や計算時間の削減等が期待できる [13]．このような学習手法は，プルーニングと呼ばれている．

DNN のモデルは，多くの重みパラメータを有しているため，中には，モデルのパフォーマンスに寄与しない冗長な重みパラメータも存在する．このような冗長な重みパラメータを刈り込むことで，モデルサイズが小さくなり，計算コストを削減することができる．しかし，DNN モデルが有する数百万の重みパラメータの中から，刈り込むべき重みパラメータの最適な組み合わせを見つけることは NP 困難な問題である [14]．そこで，重みパラメータの中でも比較的重要度の低い重みパラメータを刈り込む場合が多い．一般的な方法として，重みパラメータの絶対値が小さいものを刈り込む (0 にする) 方法が提案されている．重みパラメータが刈り込まれた後は，調整のためにモデルを再学習してプルーニングによる精度への影響が少なくなるようにする．

2.3 重み一定符号

重み一定符号 (CWC) とは，その符号語の重みが一定になるように設計される符号である．符号長 L でハミング重みが α の重み一定符号 $C(\alpha, L)$ の符号語 $\mathbf{c} = (c_0, c_1, ..., c_{L-1}), c_i \in \{0, 1\}$ は，以下の式を満たす．

$$\sum_{i=0}^{L-1} c_i = \alpha \tag{1}$$

CWC 符号語のハミング距離の最小距離は 2 であることから，誤り検出は 1 か所のみしかできない．そのため，事実上誤り訂正能力を持たない．

本研究，および従来研究では CWC として最小距離 2 の符号化および復号化に Schalkwijk のアルゴリズム [10] を使用している．

3 プルーニング耐性を有する DNN 電子透かし

従来研究では，CWC を用いることで，重みパラメータへのプルーニング攻撃に対してロバストな電子透かしを実現している．CWC 符号語に使用されるシンボル “1” の個数は一定であり，可能な限り少なくなるように設計される．そのため，CWC 符号語に使用されるシンボルは，ほとんど “0” である．CWC 符号語の埋め込みでは，2 つの閾値を設定して DNN モデルの学習に制約をかける．CWC 符号語のシンボル “1” に該当する重みパラメータの値が，大きい方の閾値より大きく，また CWC 符号語のシンボル “0” に該当する重みパラメータの値が小さい方の閾値より小さくなるように学習される．CWC が埋め込まれた DNN モデルに対してプルーニング攻撃が行われた場合，シンボル “0” に該当する重みパラメータは値が小さくても抽出が可能であるためプルーニングの影響を受けない．従来研究では，符号化時のパラメータを適切に設定することで，あらかじめ設定したレベルのプルーニング率でロバスト性を保証している．

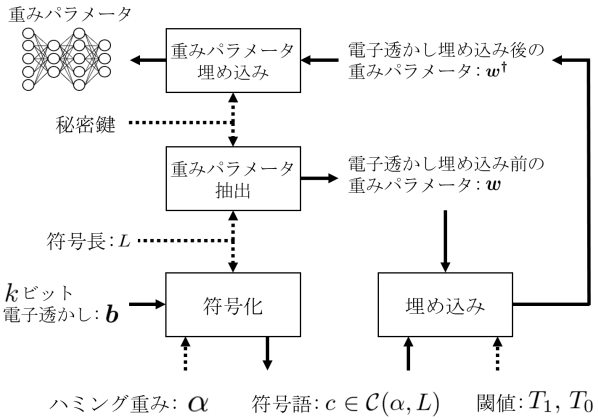


図 1 CWC 符号語の埋め込み処理の流れ

3.1 CWC 符号語の埋め込み

電子透かしの埋め込みは，CWC に基づいて行われる．埋め込み処理の流れは，図 1 に示す．初めに，k ビット電子透かし $\mathbf{b}$ を符号長 L の CWC $\mathbf{c}$ に符号化する．符号化アルゴリズムは，schalkwijk の手法 [10] を用いる．このとき，CWC 符号語の重み α と符号長 L は，以下の条件を満たす必要がある．

$$2^k \leq \binom{L}{\alpha} = \frac{L!}{\alpha!(L-\alpha)!} < 2^{k+1} \tag{2}$$

CWC 符号語 $\mathbf{c}$ と閾値 T_1,T_0 を用いて，以下の条件で DNN モデルから選択された重みパラメータ $\mathbf{w}$ を $\mathbf{w}^\dagger = (w_0^\dagger, w_1^\dagger, ..., w_{L-1}^\dagger)$ に修正する．

$$w_i^\dagger = \begin{cases} w_i & (c_i = 1) \cap (|w_i| \geq T_1) \\ sgn(w_i) \times T_1 & (c_i = 1) \cap (|w_i| < T_1) \\ w_i & (c_i = 0) \cap (|w_i| \leq T_0) \\ sgn(w_i) \times T_0 & (c_i = 0) \cap (|w_i| > T_0) \end{cases}, \tag{3}$$

ここで，$sgn(x)$ は，符号関数である．

$$sgn(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \tag{4}$$

図 2 に CWC 符号語の埋め込み処理の概略を示す． 図 2 の通り，電子透かし埋め込み後の重み $|w_i^\dagger|$ は，必ず T_1 以上，もしくは T_0 以下の値となる．

3.2 CWC 符号語の復号

秘密鍵に基づいて DNN モデルから選択された重みパラメータ $\mathbf{w}' = (w_0', w_1', ..., w_{L-1}')$ が雑音や攻撃によって変更されていない理想状態である場合，w_i' の値に基づいて埋め込まれた CWC 符号語のシンボルを抽出することができる．CWC 符号語の復号処理の流れは，図 3 に示す． 図 3 の抽出ブロックにおける CWC 符号語の抽出は，検出者があらかじめどの CWC 符号語が埋め込まれているかを知っている前提で抽出を行っている．もし，透かしが埋め込まれていない場合には，その後の復号処理において正しく復号を行うことはできない．次章の本研究による提案では，抽出前に CWC 符号

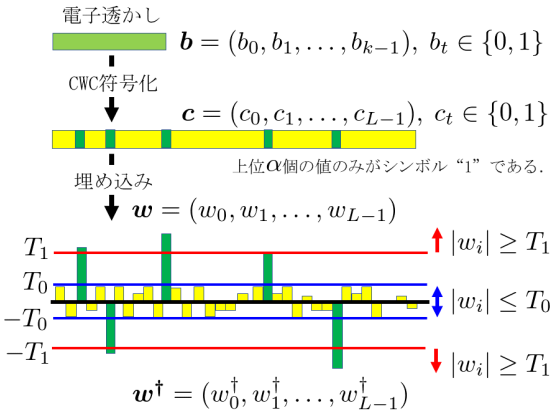


図 2　CWC 符号語の埋め込み処理の概略

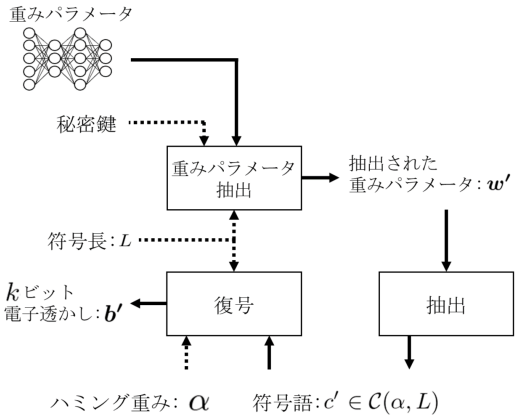


図 3　CWC 符号語の復号の流れ

語が埋め込まれているかどうかを検出する検出器を設計している．$\mathbf{w}'$ では，埋め込まれた CWC 符号語の重みパラメータの α 個のみが閾値 T_1 以上の値をとる．したがって，$\mathbf{w}'$ の値の上位 α 個を抽出後の CWC 符号語 $\mathbf{c}' = (c'_0, c'_1, \ldots, c'_{L-1})$ のシンボル $c'_i = 1$ に，また，上位 α 個を除く $L-\alpha$ 個をシンボル $c'_i = 0$ とすればよい．符号語は，符号化アルゴリズムの逆処理 [10] によって復号後の電子透かし $\mathbf{b}'$ に復号することができる．選択された重みパラメータ $\mathbf{w}'$ のうち，CWC 符号語のシンボル $c'_i = 0$ に該当する重みパラメータ，つまり上位 α 個を除く $L-\alpha$ 個のシンボルに関しては，プルーニング攻撃によって重みパラメータの値が 0 になっても抽出の影響を受けない．このことから，この手法によって保証できるプルーニング率 R を以下のように定義することができる．

$$R < \frac{L-\alpha}{L} = \overline{R} \tag{5}$$

3.3　閾値の設計

CWC 符号語は，2 つの閾値 T_1,T_0 を用いて埋め込まれる．閾値 T_1 に関しては，大きな値であればあるほど，より高いプルーニング率によるプルーニングによる刈り取りを防ぐことができる．一方で，小さければ小さいほど修正後の重みパラメータとの差が少なくなり，精度への影響が少なくなる．閾値 T_0 に関してはその逆で，大きな値であればあるほど精度への影響が少なくなる一方で，小さな値であればあるほど，より高いプルーニング率による刈り取りを防ぐことができる．したがって，2 つの閾値の設定にはトレードオフの関係があるが，任意のプルーニング率に関して適切な閾値を設定できれば都合が良い．そこで，DNN モデルの重みパラメータの値の分布に基づいて閾値を定める手法を提案している．一般的に，DNN モデルの重みパラメータは，モデルの学習前に任意の分布に従う乱数で初期化される場合が多い．重みパラメータの初期化については，ガウス分布や一様分布などの分布の選択よりも，スケールパラメータの選択が学習の収束と汎化性能に大きな影響を及ぼすことが知られている [15]．ガウス分布を選択した場合の分散は，Glorot らの研究 [16] で行われており，Xavier 初期化と呼ばれている．後に Xavier 初期化は，活性化関数である RELU との相性が悪いことが明らかとなり [17]，非線形の活性化関数に対応するように修正された [18]．従来研究 [9] では，DNN モデルの重みパラメータがガウス分布と一様分布によって初期化される場合の閾値を計算している．

重みパラメータの分布が $[-U, U]$ の一様分布である場合は，確率密度関数をもとにプルーニング率 R と閾値 T_1 について以下の不等式を得る．

$$R \leq 2T_1 \times \frac{1}{2U} \tag{6}$$

式 6 を変形して，任意のプルーニング率 $\overline{R}$ に対する適切な閾値 T_1 は以下の式で導出することができる．

$$T_1 = \overline{R}U. \tag{7}$$

3.4　CWC とプルーニング率の例

表 1 は，k ビットの電子透かしに対する CWC(α, L) の計算例を示している．例えば，128 ビットの電子透かしを $\alpha = 20$，符号長 $L = 722$ で符号化した場合，理論上プルーニング率 $R < 0.972$ のプルーニングに対してロバストであることを示している．電子透かし $\mathbf{b}$ の符号長が非常に大きい場合であっても，電子透かしを複数のブロックに分割して，ブロック単位で個別に符号化することで，同様に埋め込むことができる．

表 1　CWC の構成例

k	α	L	$\overline{R}$
64	8	972	0.992
	9	583	0.985
	10	393	0.975
	11	288	0.962
128	16	1757	0.991
	18	1063	0.983
	20	722	0.972
	22	533	0.959
256	32	3307	0.990
	36	2011	0.982
	40	1373	0.971
	43	1090	0.961

4　提案手法

従来研究では，電子透かしを CWC で符号化し DNN モデルに対する埋め込む方法および，抽出した CWC 符号語の復号方法について提案を行った．しかし，秘密鍵に基づいて DNN モデルから選択される重みパラメータ

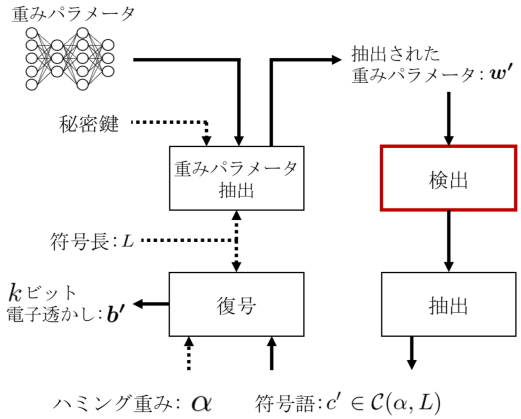
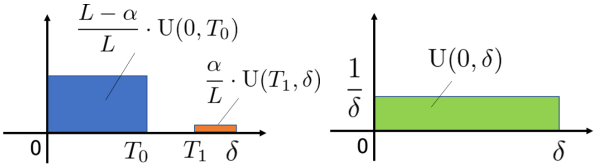


図 4　検出器を導入した CWC 符号語の復号処理の流れ



(a) CWC 符号語　　(b) ランダム系列

図 5　抽出した重みパラメータの分布

に電子透かしが埋め込まれているか否かの電子透かしの検出方法に関しては，理論的な解析が未検討であった．本研究では，従来研究によって電子透かしが埋め込まれた重みパラメータと埋め込まれていない重みパラメータの分布の特徴を理論的に説明し，電子透かしの存在を適切に検出する検出器を提案する．従来研究における CWC 符号語の復号処理は，本研究の検出器を加えることで図 4 の流れとなる． なお，本提案における重みパラメータは，単純化のために範囲 $[-\delta, \delta]$ の一様分布で初期化されているものとし，重みパラメータの絶対値を扱うこととする．

4.1　重みパラメータの値の分布

秘密鍵に基づいて DNN モデルから選択される重みパラメータ $\mathbf{w}'$ に電子透かしが埋め込まれている場合，値の上位 α 個とそれ以外では，異なる分布に従う．図 5(a) は，重みパラメータに電子透かしが埋め込まれている場合の値の分布を示したものである．値の上位 α 個は，閾値 T_1 に基づいて重みパラメータの値の修正が行われているため，範囲 $[T_1, \delta]$ の一様分布に従う．一方で，上位 α 個を除く重みパラメータでは，T_0 に基づいて重みパラメータの修正が行われているため，範囲 $[0, T_0]$ の一様分布に従う．電子透かしが重みパラメータに埋め込まれていない場合は，図 5(b) に示すように，元の分布，すなわち範囲 $[0, \delta]$ の一様分布に従う．

4.2　検出器の設計

電子透かしの有無による重みパラメータがなす分布の違いに着目することで，検出器を設計することができる．重みパラメータに電子透かしが埋め込まれている場合は，図 5(a) に示すような分布であるため，上位 α 個の重みパラメータの平均値は，$(\delta + T_1)/2$ であり，上位 α 個を除く $L - \alpha$ 個の重みパラメータの平均値は，$T_0/2$ である．一方，電子透かしが無い場合には，図 5(b) に示すような分布となり，重みパラメータの平均値は $1/\delta$ である．提案検出器では，対象となる DNN モデルから抽出した重みパラメータ $\mathbf{c}'$ の値に対して，前者に示した平均値からのばらつきを指標として分布の違いを判別する．そのばらつきの指標は以下に示す平均二乗誤差 (Mean Square Error: MSE) によって求める．

$$MSE = \frac{1}{L} \sum_{i=0}^{L-1} d_i, \tag{8}$$

ただし

$$d_i = \begin{cases} \left(c'_i - \frac{\delta + T_1}{2}\right)^2 & (0 \le i < \alpha) \\ \left(c'_i - \frac{T_0}{2}\right)^2 & (\alpha \le i < L) \end{cases} \tag{9}$$

DNN モデルから抽出した重みパラメータに電子透かしが埋め込まれている場合，MSE の値は小さくなり，また，埋め込まれていない場合は，MSE の値は大きくなる．

電子透かしの検出は，以下の手順で行う．

1. DNN モデルから秘密鍵に基づく L 個の重みパラメータを抽出．
2. MSE を計算．
3. MSE の値が検出用閾値を超えれば電子透かしありとして検出．

4.3　MSE の期待値

DNN モデルから抽出した重みパラメータが CWC 符号語である場合とランダム系列である場合において，MSE の期待値の理論値 (E_C, E_N) は，以下の式で求めることができる．

CWC 符号語の場合:E_C

$$\begin{aligned} E[MSE] &= \frac{1}{L}\left\{\alpha \cdot \frac{1}{\delta - T_1} \int_{T_1}^{\delta} \left(x - \frac{\delta + T_1}{2}\right)^2 dx \right. \\ &\quad \left. + (L - \alpha) \cdot \frac{1}{T_0} \int_0^{T_0} \left(x - \frac{T_0}{2}\right)^2 dx \right\} \qquad (10) \\ &= \frac{1}{L}\left\{\frac{\alpha}{\delta - T_1}\left[\frac{1}{3}x^3 - \frac{\delta + T_1}{2}x^2 + \left(\frac{\delta + T_1}{2}\right)^2 x\right]_{T_1}^{\delta} \right. \\ &\quad \left. + \frac{L - \alpha}{T_0}\left[\frac{1}{3}x^3 - \frac{T_0}{2}x^2 + \left(\frac{T_0}{2}\right)^2 x\right]_0^{T_0} \right\} \qquad (11) \end{aligned}$$

ランダム系列の場合 E_N

$$\begin{aligned} E[MSE] &= \frac{1}{L}\left\{\alpha \cdot \frac{1}{\delta} \int_0^{\delta} \left(x - \frac{\delta + T_1}{2}\right)^2 dx \right. \\ &\quad \left. + (L - \alpha) \cdot \frac{1}{\delta} \int_0^{\delta} \left(x - \frac{T_0}{2}\right)^2 dx \right\} \qquad (12) \\ &= \frac{\alpha}{\delta L}\left[\frac{1}{3}x^3 - \frac{\delta + T_1}{2}x^2 + \left(\frac{\delta + T_1}{2}\right)^2 x\right]_0^{\delta} \\ &\quad + \frac{L - \alpha}{\delta L}\left[\frac{1}{3}x^3 - \frac{T_0}{2}x^2 + \left(\frac{T_0}{2}\right)^2 x\right]_0^{\delta} \qquad (13) \end{aligned}$$

上記の検出器は，検出時にどの CWC 符号語 $\mathbf{c}$ が埋め込まれているかを知っている前提である．重みパラメータの値が上位 α 個は，電子透かしの有無に関わらず，そ

の値が δ に近い値となることが期待される．それゆえ，上位 α 個を除いた $L-\alpha$ 個の重みパラメータで MSE を計算した方が，電子透かしの有無によりその差をより正確に示すことができる．そのため，MSE の計算を次のように修正する．

$$\tilde{MSE} = \sum_{i=0}^{L-\alpha-1} \left(\tilde{c}_i - \frac{T_0}{2}\right)^2 \tag{14}$$

ただし，$\tilde{\mathbf{c}} = \mathrm{sort}(\mathbf{c}')$ である．

以下に，修正版の電子透かしの検出方法を示す．

1. DNN モデルから秘密鍵に基づく L 個の重みパラメータ $\mathbf{c}'$ を抽出
2. $\mathbf{c}'$ の L 個の重みパラメータのうち，上位 α 個を除いて，残りの $L-\alpha$ 個の重みパラメータで $\tilde{MSE}$ を計算
3. $\tilde{MSE}$ の値が検出用閾値を超えれば電子透かしありとして検出

修正後の $\tilde{MSE}$ の期待値の理論値 $(\tilde{E}_C, \tilde{E}_N)$ は，以下の式で求めることができる．

CWC 符号語の場合:$\tilde{E}_C$

$$E[\tilde{MSE}] = \frac{1}{L}\left\{(L-\alpha)\cdot\frac{1}{T_0}\int_0^{T_0}\left(x-\frac{T_0}{2}\right)^2 dx\right\} \tag{15}$$

$$= \frac{L-\alpha}{T_0 L}\left[\frac{1}{3}x^3 - \frac{T_0}{2}x^2 + \left(\frac{T_0}{2}\right)^2 x\right]_0^{T_0} \tag{16}$$

ランダム系列の場合:$\tilde{E}_N$

$$E[\tilde{MSE}] = \frac{1}{L}\left\{(L-\alpha)\cdot\frac{1}{\delta}\int_0^{\delta}\left(x-\frac{T_0}{2}\right)^2 dx\right\} \tag{17}$$

$$= \frac{L-\alpha}{\delta L}\left[\frac{1}{3}x^3 - \frac{T_0}{2}x^2 + \left(\frac{T_0}{2}\right)^2 x\right]_0^{\delta} \tag{18}$$

5 実験結果

本章では，提案手法によって CWC 符号語が正しく検出できることをシミュレーションによって評価する．

5.1 実験設定

実験設定として，CWC 符号語が埋め込まれた系列 $\mathbf{c}'$ と CWC 符号語が埋め込まれていないランダムな系列 $\mathbf{n}'$ をそれぞれ $M = 100000$ 個用意し，MSE を使って検出を行う．それぞれの系列の重みパラメータ α と符号長 L は，CWC$(16, 1757)$ として設定する．シミュレーションの単純化のために，各値は絶対値をとり，範囲 $[0, \delta]$ の一様分布によって初期化される．このとき，一様分布の上限 δ は，$\delta = 0.026650$ とする．MSE の期待値の理論値 (15)(17) によると検出には閾値 T_1 を必要としない．したがって，閾値 T_0 のみを変動させて結果を確認する．T_0 は，$T_0 = \{0.010, 0.015, 0.020\}$ でそれぞれ検証を行う．

5.2 MSE と検出用閾値

5.1 節で設定したパラメータにおける MSE の期待値の理論値 $(\tilde{E}_C, \tilde{E}_N)$ とシミュレーションによる実験値を表 2 に示す．ただし，CWC 符号語が埋め込まれた系列 $\mathbf{c}'$ と CWC 符号語が埋め込まれていないランダムな系列 $\mathbf{n}'$ は，$M = 100000$ 個の平均値である． また，図 6 に T_0 の違いによる MSE の頻度分布を示す． この結果から，実験値が理論値通りの値であることを確認できたことに加え，T_0 が大きいほど CWC 符号語が埋め込まれた系列 $\mathbf{c}'$ と CWC 符号語が埋め込まれていないランダムな系列 $\mathbf{n}'$ の MSE の差が小さくなることが分かった．これは，$T_0 < T_1$ の制約の下で T_0 を T_1 に近づけることで，CWC 符号語の分布がランダムな系列の分布に近づくからである．ランダムな系列は，$T_0 = T_1 = \delta$ の特殊な CWC 符号語であるとも言える．CWC 符号語の分布がランダムな系列に近づくにつれて，分散が大きくなり誤検出の懸念がある．しかし，通常はそのような極端なパラメータを設定することはないため，例えば $T_0 < T_1/2$ となるように T_0 と T_1 で十分なマージンを設ければ，単純な検出用の閾値であっても十分な検出が期待できる．最も単純な検出用閾値 T_d は，CWC 符号語とランダムな系列の各 MSE の期待値の理論値 (15)(17) の平均値:$T_d = (\tilde{E}_C + \tilde{E}_N)/2$ である．表 3 に，各 T_0 に対する検出用閾値 T_d を示す． さらに，T_0 と T_1 で十分なマージンを設けることは，重みパラメータが雑音等で歪んだ場合の復号誤り率を抑制する利点もある．

表 2 MSE の理論値と実験値の比較

	CWC 符号語		ランダム系列	
T_0	理論値 $\tilde{E}_C$	実験値 $\mathbf{c}'$	理論値 $\tilde{E}_N$	実験値 $\mathbf{n}'$
0.010	0.0000083	0.0000083	0.0001273	0.0001254
0.015	0.0000186	0.0000188	0.0000923	0.0000907
0.020	0.0000330	0.0000333	0.0000696	0.0000684

表 3 閾値 T_0 に対する検出用閾値 T_d

α	L	δ	T_0	T_d
16	1757	0.02665	0.010	0.0000678
			0.015	0.0000554
			0.020	0.0000513

6 おわりに

本研究では，DNN モデル内の重みパラメータを対象としたプルーニング攻撃に対する耐性を考慮して，重み一定符号を用いた符号化手法を提案した従来研究において，DNN モデルから選択された重みパラメータに符号語が埋め込まれているか否かを検出する検出器を提案した．検出は，符号語とランダムな系列の分布の違いに着目し，平均二乗誤差を計算して理論値に基づいて判断する．実験の結果，符号語とランダムな系列の理論的に導出した MSE の期待値の正当性を確認することができた．今後の課題として，ガウス分布のような一様分布とは異なる分布を用いて初期化された DNN の重みパラメータに対する理論的な検出器の設計が挙げられる．

謝辞

本研究は，JSPS 科研費 19K22846, JST SICORP, JPMJSC20C3, JST CREST JPMJCR20D3 の支援を受けたものである．

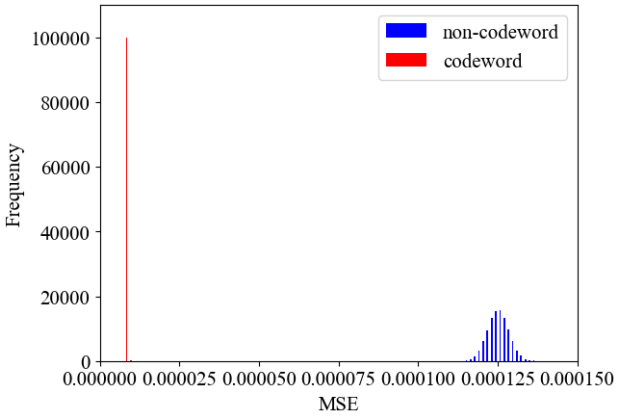
参考文献

[1] Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh, “Embedding watermarks into deep neural networks,” in *Proc. ICMR’17*, pp. 269–277, 2017.

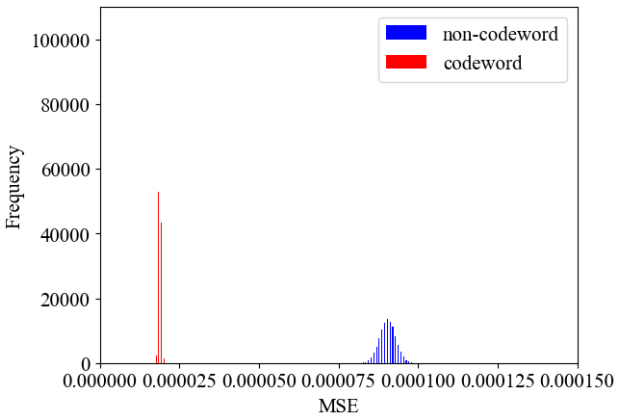
[2] Y. Nagai, Y. Uchida, S. Sakazawa, and S. Satoh, “Digital watermarking for deep neural networks,” *International Journal of Multimedia Information Retrieval*, vol. 7, pp. 3–16, 2018.

[3] B. D. Rouhani, H. Chen, and F. Koushanfar, “DeepSigns: An end-to-end watermarking framework for ownership

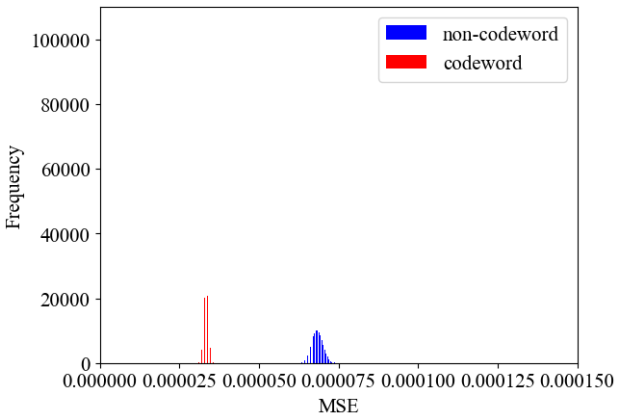
protection of deep neural networks," in *Proc. ASPLOS'19*, pp. 485–497, 2019.
[4] H. Chen, B. D. Rouhani, X. Fan, O. C. Kilinc, and F. Koushanfar, "Performance comparison of contemporary DNN watermarking techniques," *CoRR*, vol. abs/1811.03713, 2018.
[5] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, *The loss surfaces of multilayer networks*. Artificial Intelligence and Statistics, 2015.
[6] Y. N. Dauphin, R. Pascanu, Ç. Gülçehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Proc. NIPS'14*, pp. 2933–2941, 2014.
[7] Y. Kong and J. Zhang, "Adversarial audio: A new information hiding method and backdoor for DNN-based speech recognition models," *CoRR*, vol. abs/1904.03829, 2019.
[8] Y. Wang and H. Wu, "Protecting the intellectual property of speaker recognition model by black-box watermarking in the frequency domain," *Symmetry*, vol. 14, no. 3, p. 619, 2022.
[9] T. Yasui, T. Tanaka, A. Malik, and M. Kuribayashi, "Coded dnn watermark: Robustness against pruning models using constant weight code," *Journal of Imaging*, vol. 8, no. 6, 2022.
[10] J. P. M. Schalkwijk, "An algorithm for source coding," *IEEE Trans. Information Theory*, vol. IT-18, no. 3, pp. 395–399, 1972.
[11] E. Le Merrer, P. Perez, and G. Trédan, "Adversarial frontier stitching for remote neural network watermarking," *Neural Computing and Applications*, vol. 32, no. 13, pp. 9233–9244, 2020.
[12] H. Wu, G. Liu, Y. Yao, and X. Zhang, "Watermarking neural networks with watermarked images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2591–2601, 2021.
[13] N. Denil, B. Shakibi, L. Dinh, M. A. Ranzato, and N. D. Freitas, "Predicting parameters in deep learning," in *Advances In Neural Information Processing Systems*, pp. 2148–2156, 2013.
[14] Y. Guo, A. Yao, and Y. Chen, "Dynamic network surgery for efficient DNNs," in *Advances In Neural Information Processing Systems*, pp. 1379–1387, 2016.
[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[16] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. PMLR'10*, vol. 9, pp. 249–256, 2010.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. ICCV'15*, pp. 1026–1034, 2015.
[18] S. K. Kumar, "On weight initialization in deep neural networks," *CoRR*, vol. abs/1704.08863, 2017.



(a) $T_0 = 0.010$



(b) $T_0 = 0.015$



(c) $T_0 = 0.020$

図 6　$\tilde{MSE}$ の頻度分布