AT-1：組合せ論と情報理論
ー 最新動向と展望 ー

プルーニングに対する耐性を高めるための重み一定符号化による深層学習モデル保護用電子透かし

栗林　稔（東北大学）

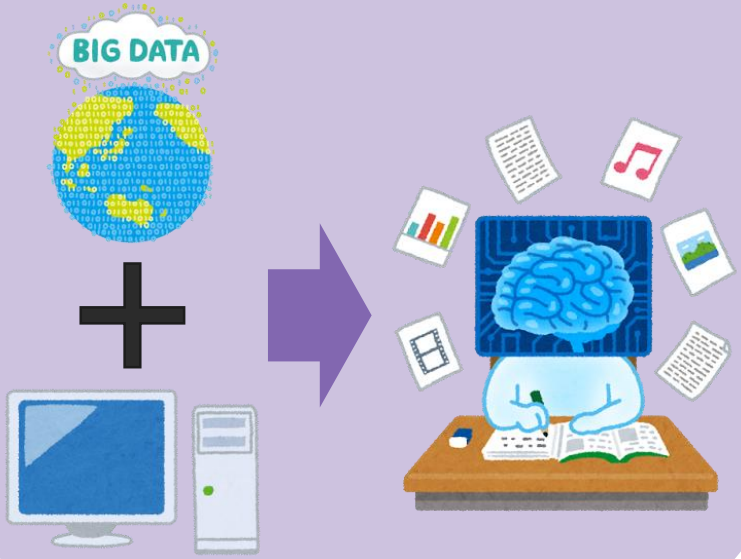
目次

- 研究背景
- DNNモデルへの電子透かし
- プルーニング攻撃
- 重み一定符号を用いた電子透かし
  - 重み一定符号
  - 埋め込み処理
- 透かし情報の検出と抽出
- まとめ

研究背景

計算機性能の向上 → 深層学習の研究が盛んに行われている

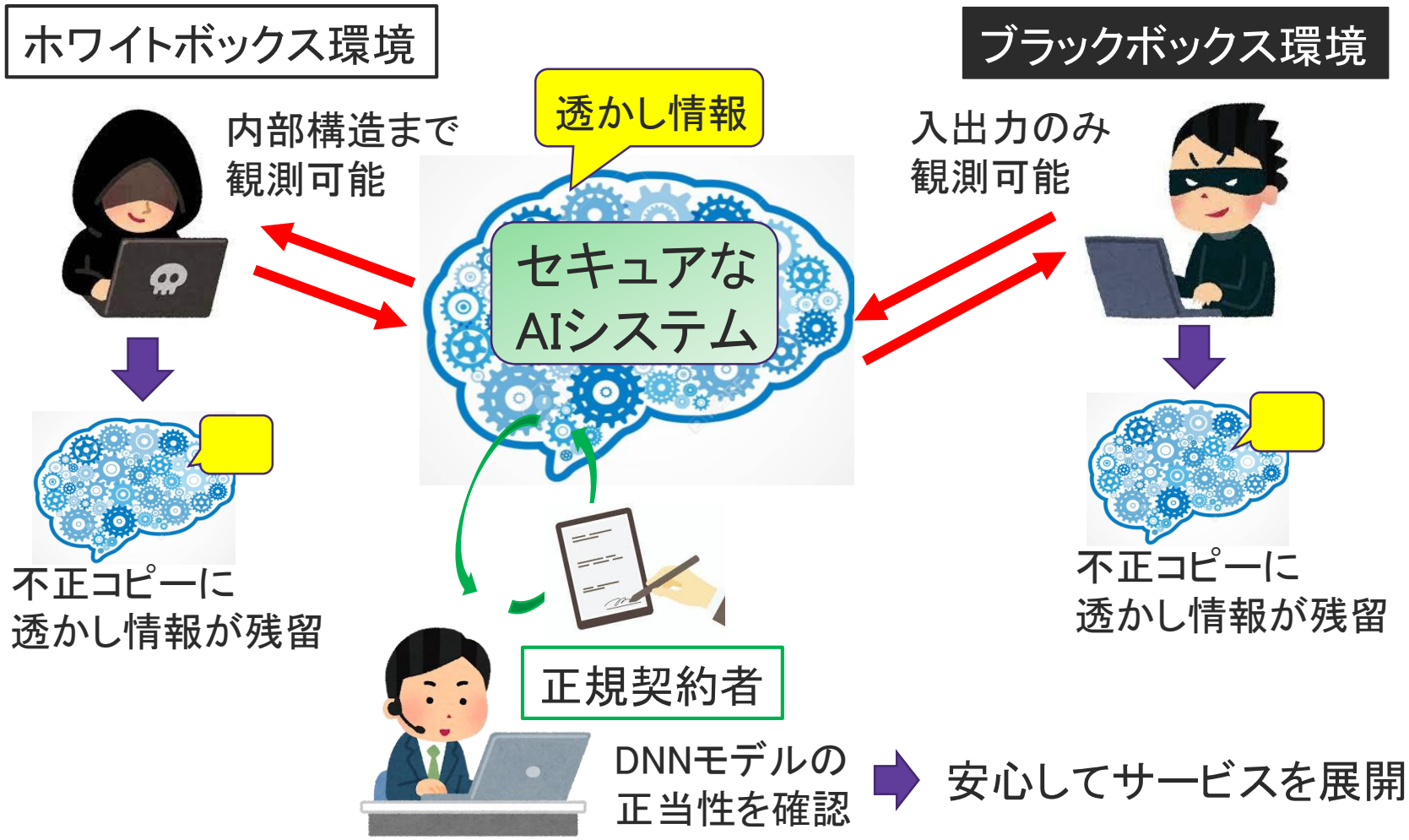
DNNモデルの作成



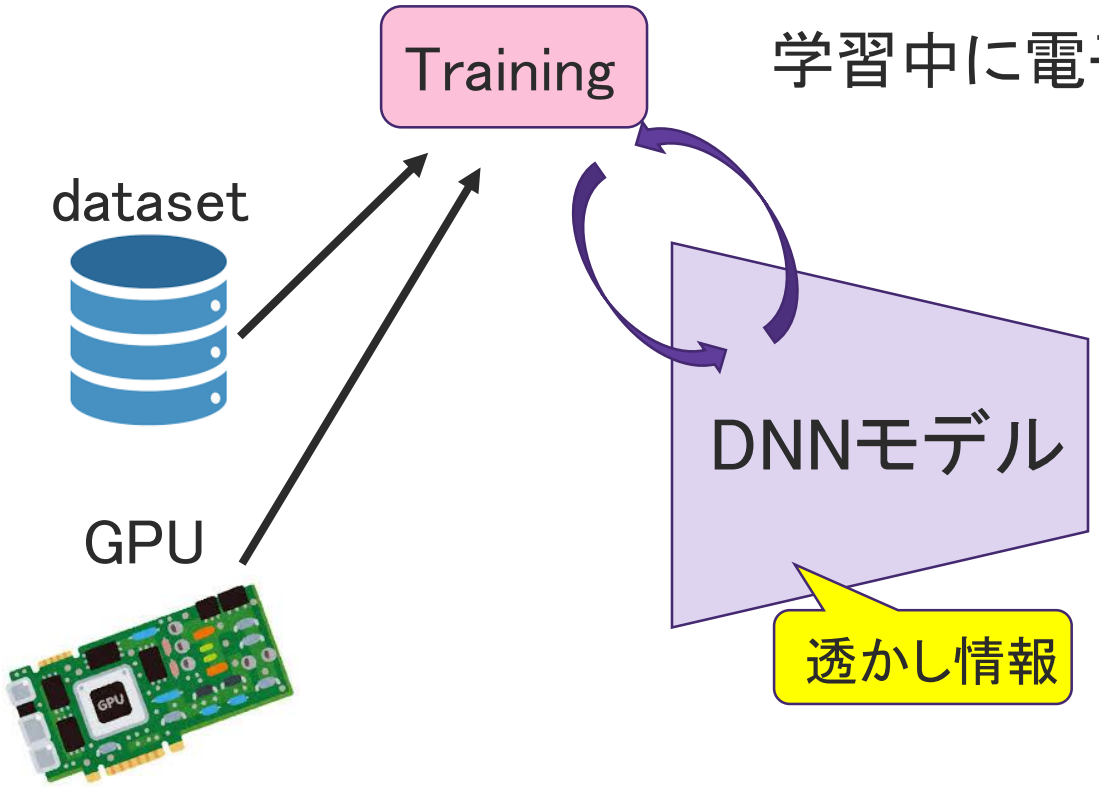
学習済みモデルの不正利用を防ぐ

電子透かし技術を用いて作成者の権利保護

深層学習モデル保護のための電子透かし



透かし情報の埋め込み



学習中に電子透かしを埋め込む

ホワイトボックス環境

モデルの内部情報を
直接書き換え

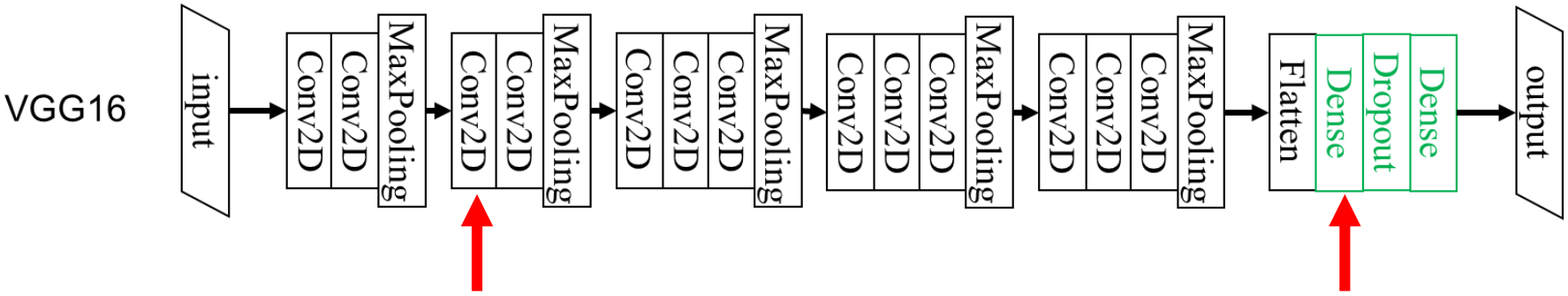
ブラックボックス環境

トリガー入力に対して
特異な出力をする
ように学習させる

CNNモデルへの適用例

- VGG16のファインチューニングモデルへの適用

  学習によって重みパラメータを調整する際に
  指定した箇所に透かし情報を埋め込む



畳み込みフィルタ
のパラメータ

全結合層
のパラメータ

莫大な数の重みパラメータがある

↳ 秘密鍵に基づいてサンプリング

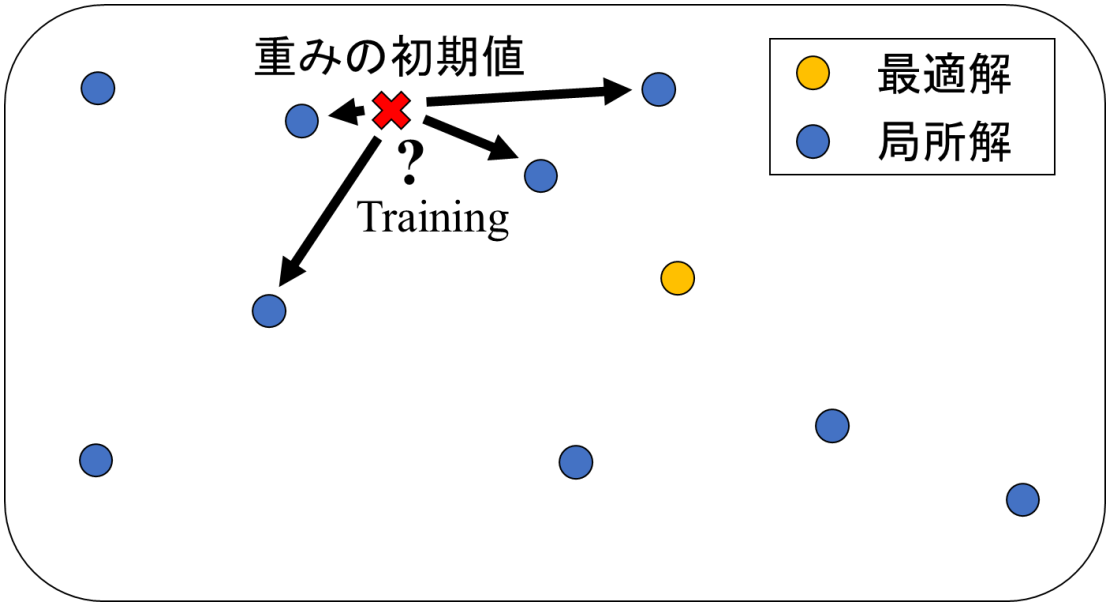
目次

- 研究背景
- **DNNモデルへの電子透かし**
- プルーニング攻撃
- 重み一定符号を用いた電子透かし
  - 重み一定符号
  - 埋め込み処理
- 透かし情報の検出と抽出
- まとめ

DNNモデルの学習の収束

- 学習：損失関数の出力が小さい局所解を探索

複数の局所解での性能≒最適解での性能

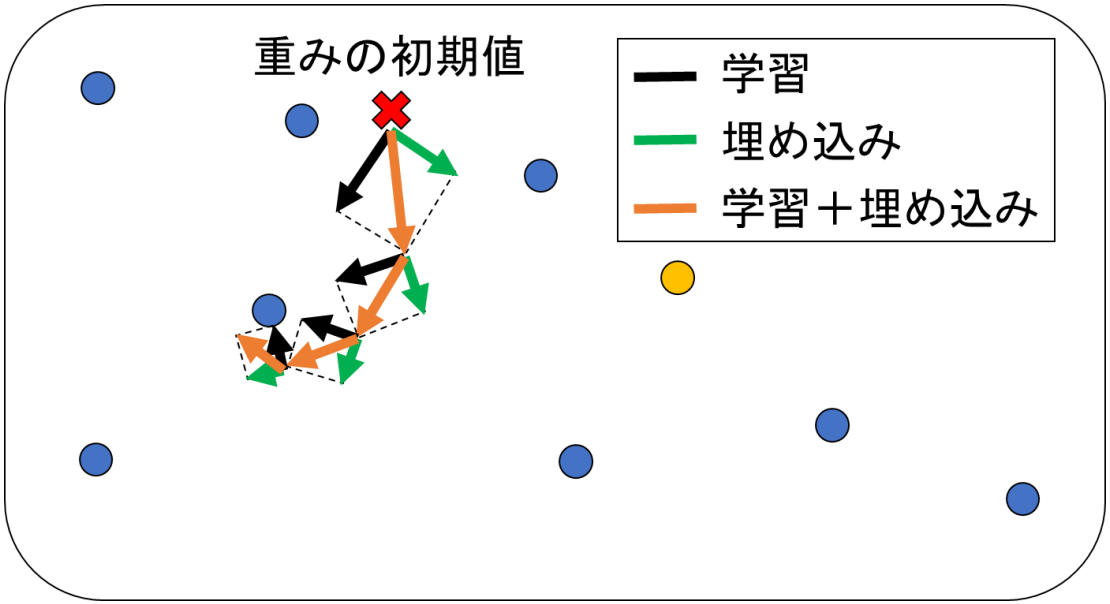


学習用と埋め込み用の損失関数の出力の和を小さくするように訓練させる

DNNモデルの学習の収束

- 学習：損失関数の出力が小さい局所解を探索

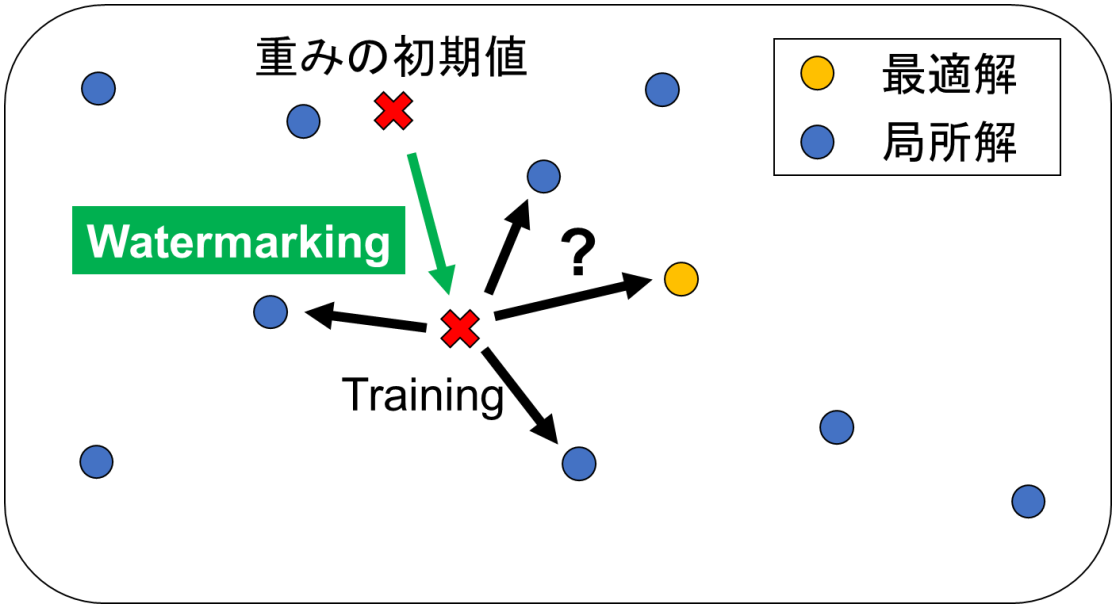
複数の局所解での性能≒最適解での性能



学習用と埋め込み用の損失関数の出力の和を小さくするように訓練させる

埋め込み処理の影響

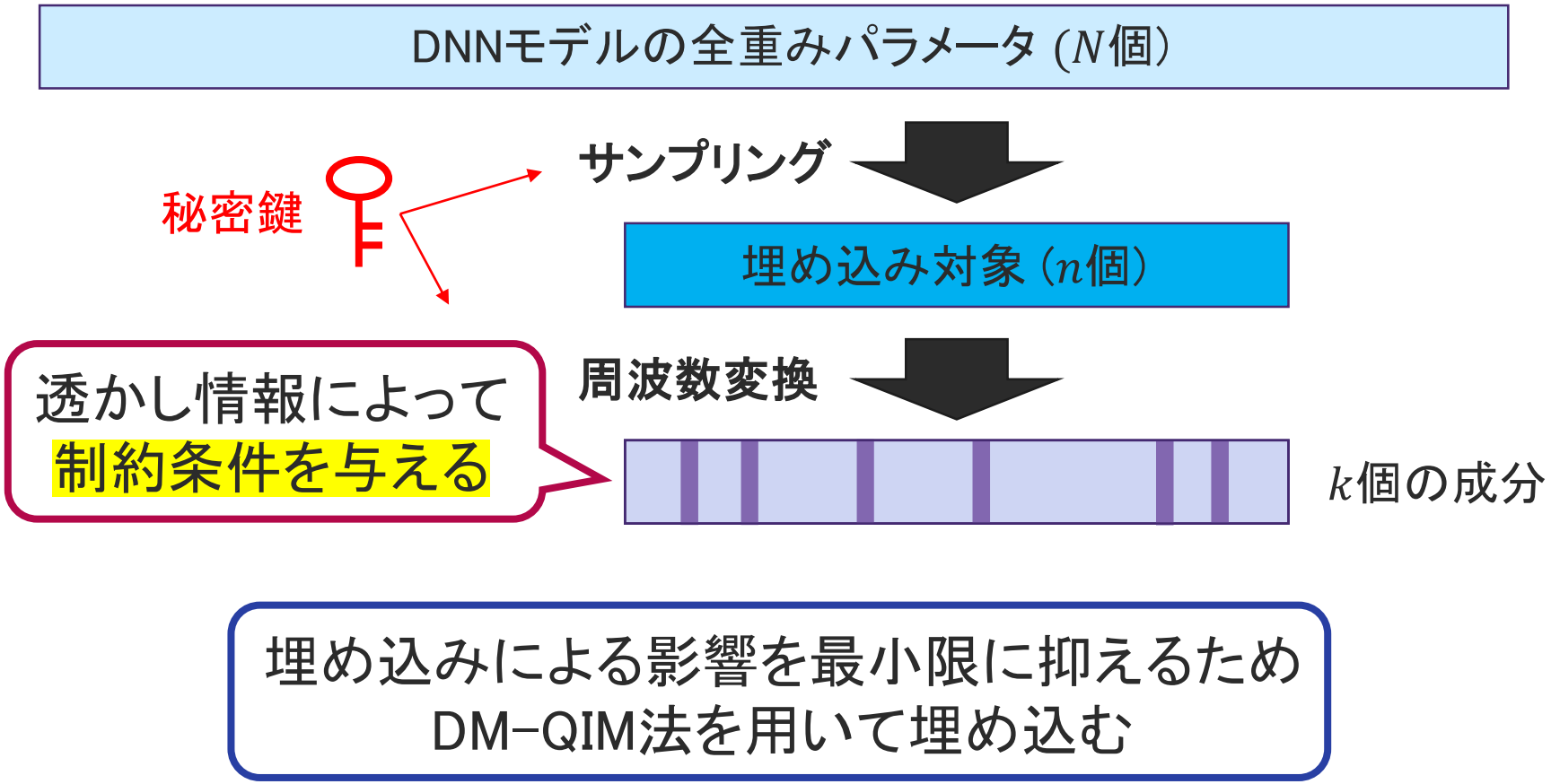
- 最初のエポックでの透かし情報埋め込みで重みの初期値を大きく変化させる



それぞれの局所解での性能≒最適解での性能であるため，最終的な性能には影響を与えない

埋め込み手法 (IH&MMSEC'21)

- 透かし情報は重みパラメータの一部に埋め込む

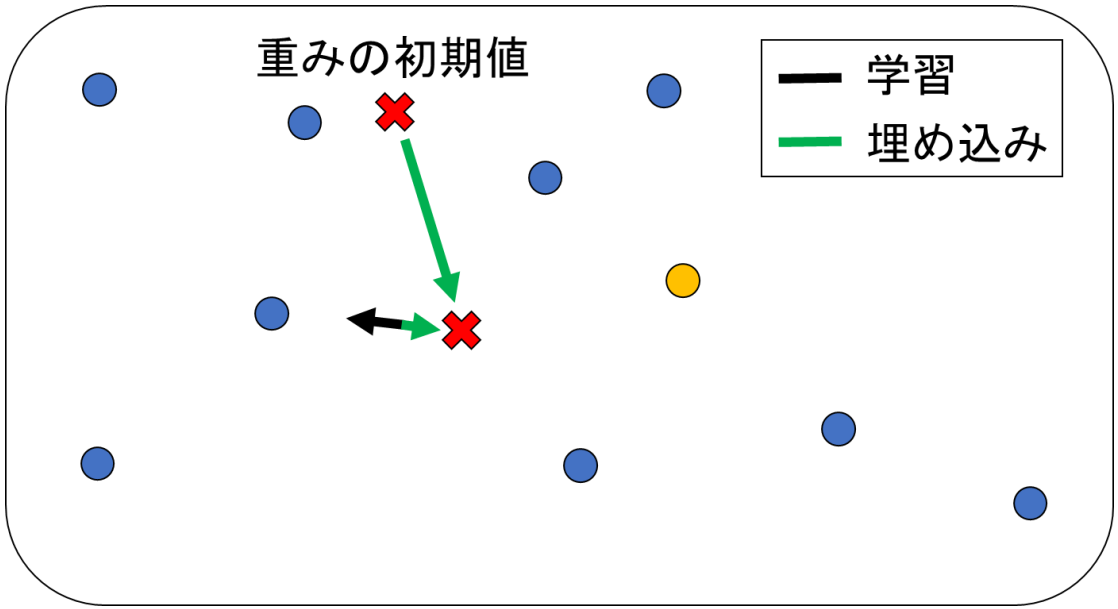


埋め込みによる影響を最小限に抑えるため
DM-QIM法を用いて埋め込む

M. Kuribayashi, T. Tanaka, S. Suzuki, T. Yasui, and N. Funabiki, "White-box watermarking scheme for fully-connected layers in fine-tuning model," Proc. IH&MMSec'21, pp.165-170, 2021.

歪み補正とその影響

- 2エポック目以降では，学習により変化したn個の重みパラメータに対して歪み補正を行う．



歪み補正が与える影響はn/Nとなるため，
$n \ll N$であれば，その影響は無視できる

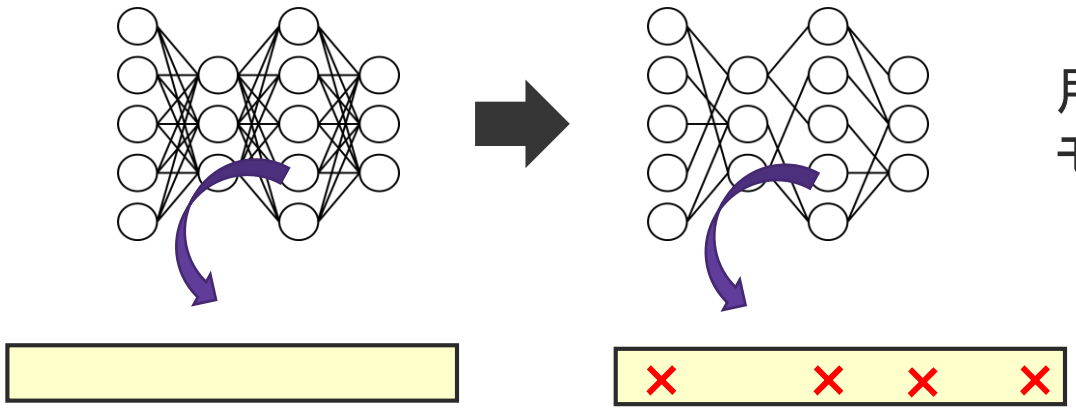
目次

- 研究背景
- DNNモデルへの電子透かし
- **プルーニング攻撃**
- 重み一定符号を用いた電子透かし
  - 重み一定符号
  - 埋め込み処理
- 透かし情報の検出と抽出
- まとめ

プルーニング攻撃

- 精度への寄与が少ない重みを消す(枝刈り)

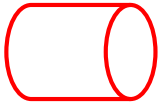
※一般的に値が小さい重みの値を0にする(昇順プルーニング)



用途:
モデルの圧縮(軽量化)

大部分が消失

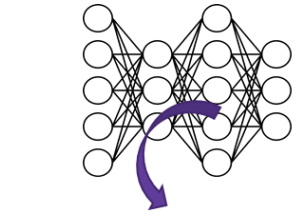
(重みの値がゼロ)



消失通信路と見なせる

消失の割合が大きいため、誤り訂正符号では対応が難しい

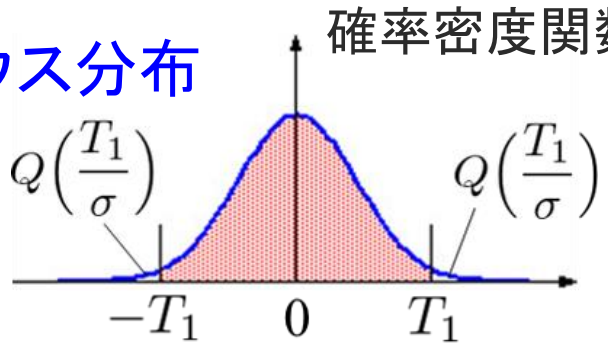
プルーニングのモデル化



重みの初期値

- 学習モデルの重みの分布

ガウス分布

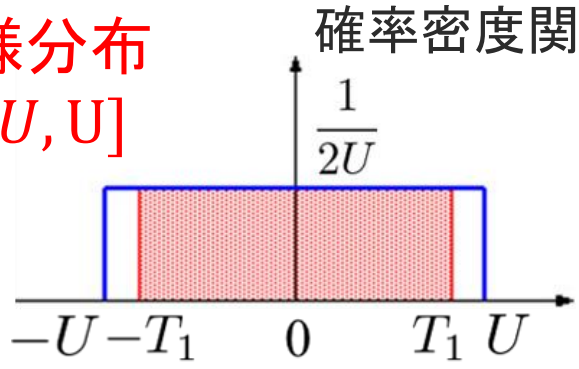


Q関数

$$Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} \exp\left(-\frac{x^2}{2}\right) dx$$

プルーニング対象

一様分布
$[-U, U]$



プルーニング攻撃の戦略

昇順	小さい順番に枝刈り
ランダム	対象の重みを ランダムに枝刈り

目次

- 研究背景
- DNNモデルへの電子透かし
- プルーニング攻撃
- **重み一定符号を用いた電子透かし**
  - **重み一定符号**
  - **埋め込み処理**
- 透かし情報の検出と抽出
- まとめ

重み一定符号

(CWC: Constant Weight Code)

- 符号語の重みが一定になるように設計された符号
- 符号長Lでハミング重みがαの重み一定符号$\boldsymbol{C}(\alpha, L)$

$$\boldsymbol{c} = (c_0, c_1, \ldots, c_{L-1}), \quad c_i \in \{0, 1\}$$

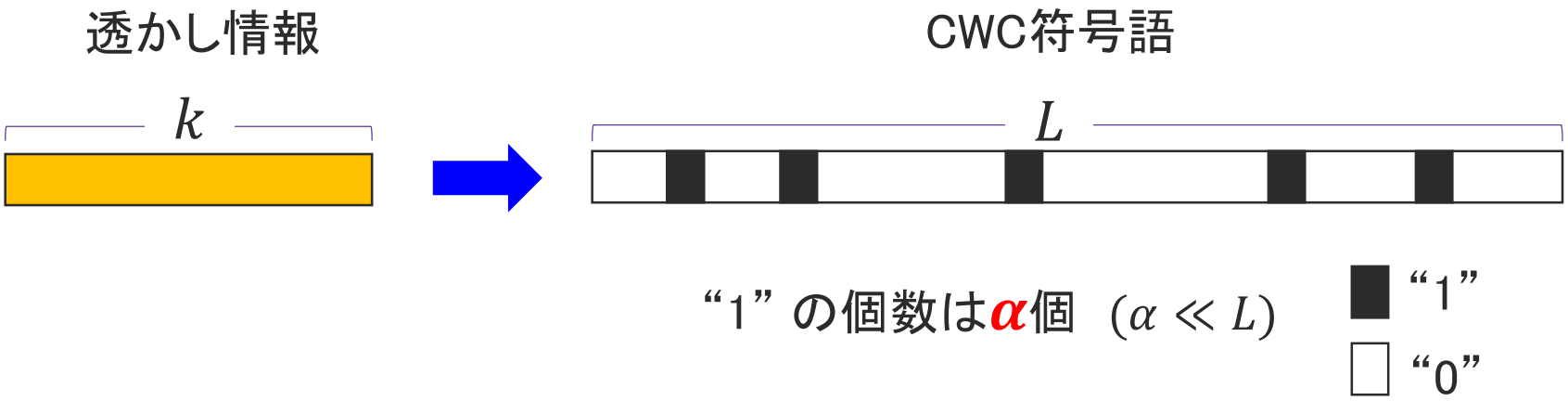
$$\sum_{i=0}^{L-1} c_i = \alpha$$

- kビットの情報を符号化する場合，次の条件を満足する

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

CWC符号化

- kビットの透かし情報をLビットのバイナリ符号語に符号化



プルーニング攻撃のイメージ

- 攻撃を受ける箇所をシンボル“0”に限定させる
- シンボル“1”の箇所が分かれば情報を復号可能

符号パラメータの例

k	α	L	$\bar{R}$
64	8	972	0.9918
	9	583	0.9846
	10	393	0.9746
	11	288	0.9618

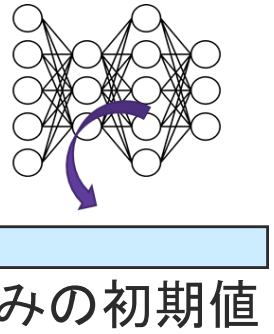
k	α	L	$\bar{R}$
256	32	3307	0.9903
	36	2011	0.9821
	40	1373	0.9709
	43	1090	0.9606

k	α	L	$\bar{R}$
128	16	1757	0.9909
	18	1063	0.9831
	20	722	0.9723
	22	533	0.9587

プルーニング率

$$\bar{R} = \frac{L - \alpha}{L}$$

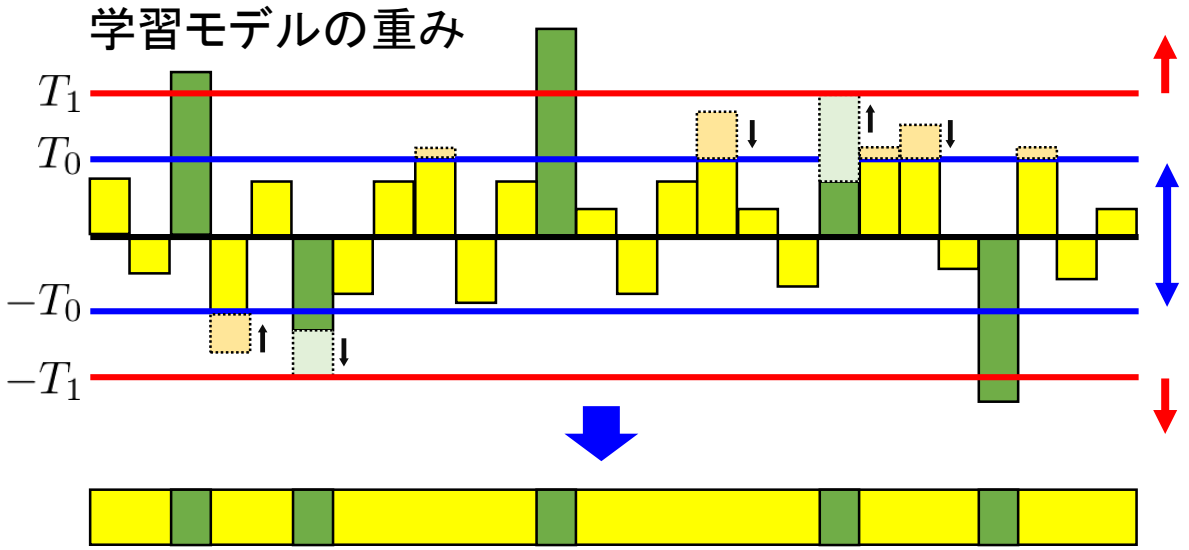
透かし情報の埋め込み



①CWC符号語を生成

CWC符号語

②学習モデルの重みをCWC符号語に基づいて修正



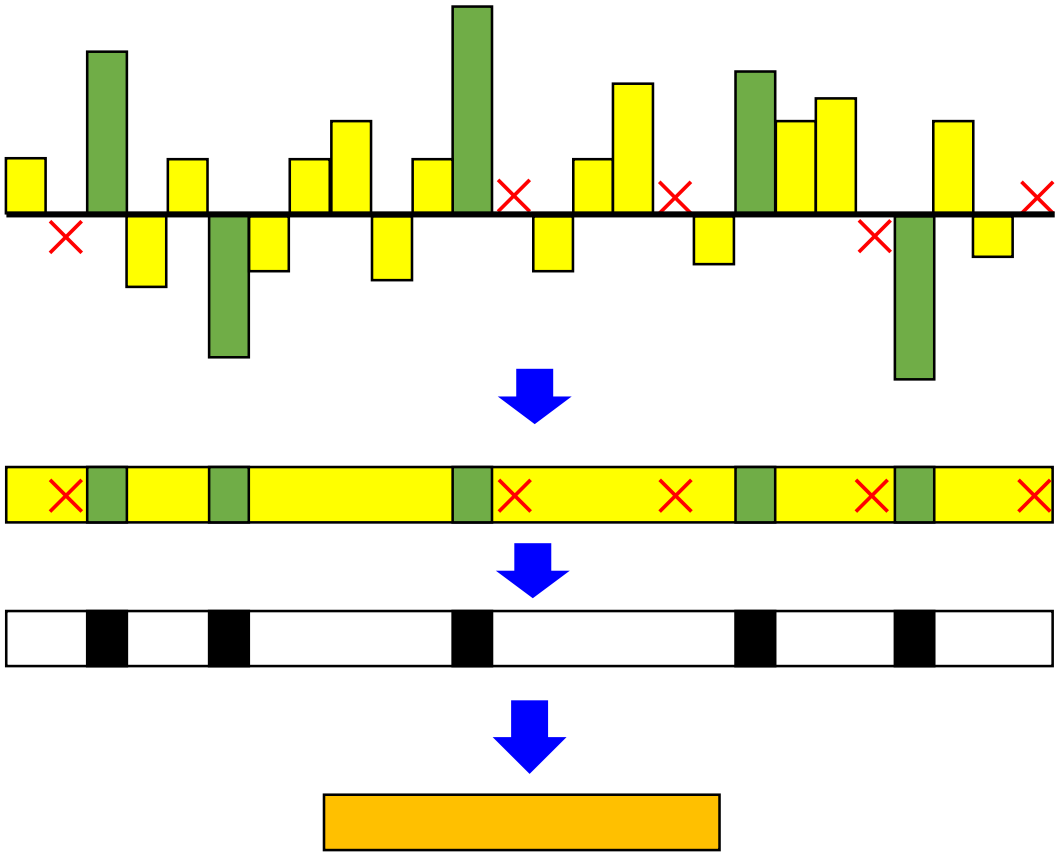
シンボルが“1”の箇所
閾値 T_1 以上になるよう修正

シンボルが“0”の箇所
閾値 T_0 以下になるよう修正

シンボルが“1”の箇所
閾値 T_1 以下になるよう修正

透かし情報の抽出

上位α個をシンボル”1”としてCWC符号語を抽出

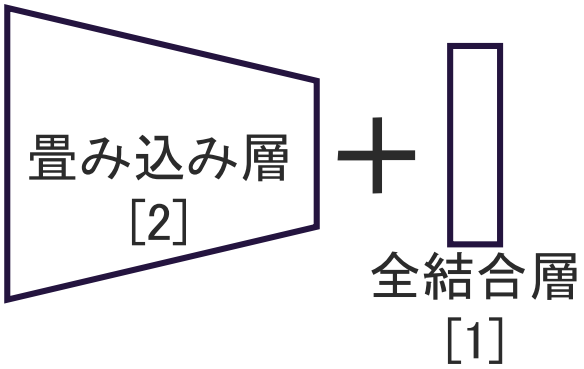


プルーニング攻撃後の
重みパラメータ

シンボル”1”は値が
大きいため影響しない

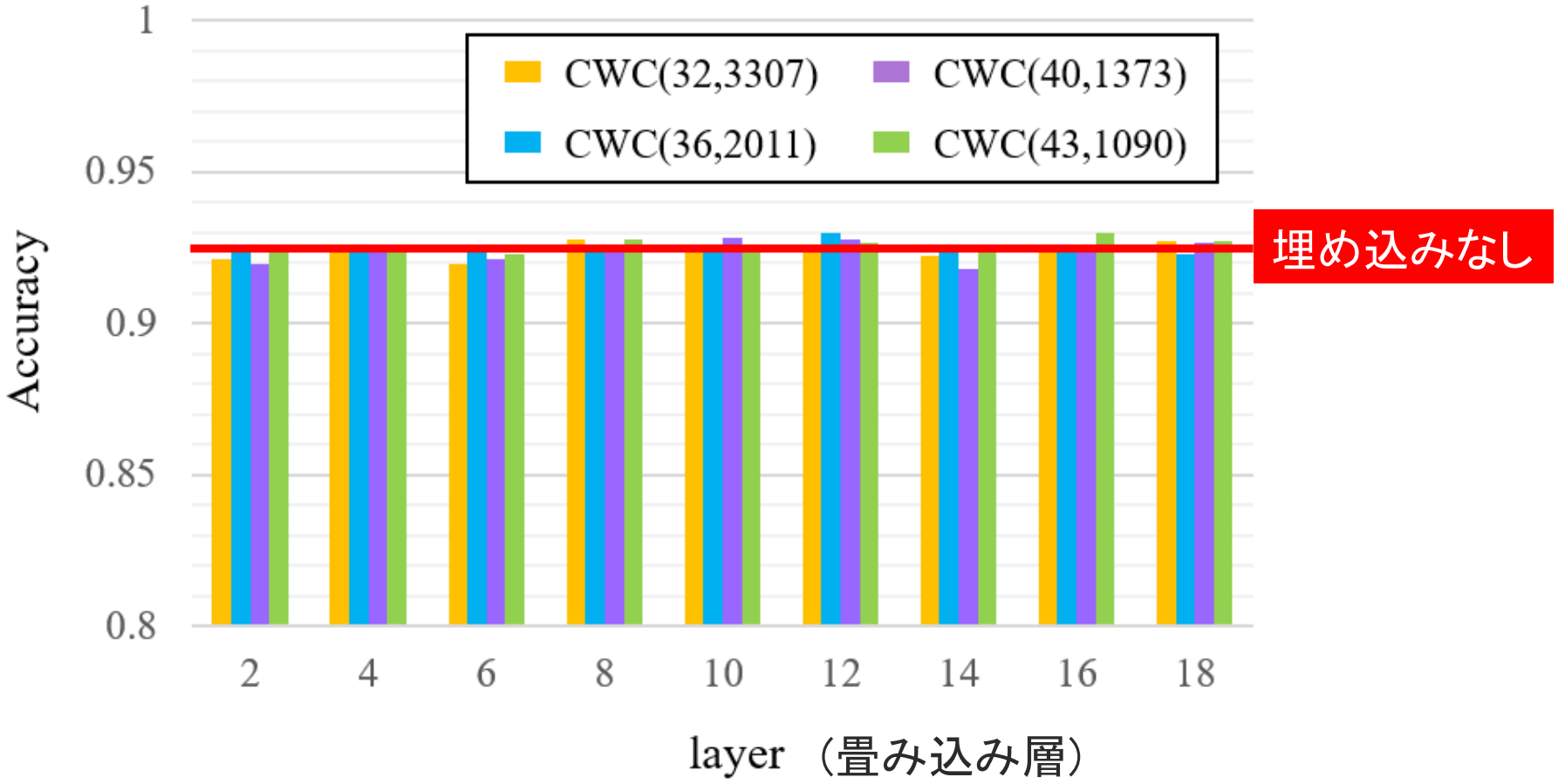
評価環境

学習済みモデル	VGG16, ResNet50
埋め込み対象の層	全結合層
使用したデータセット	17 Category Flower Dataset
重みの分布	一様分布
透かし情報のビット数k	256



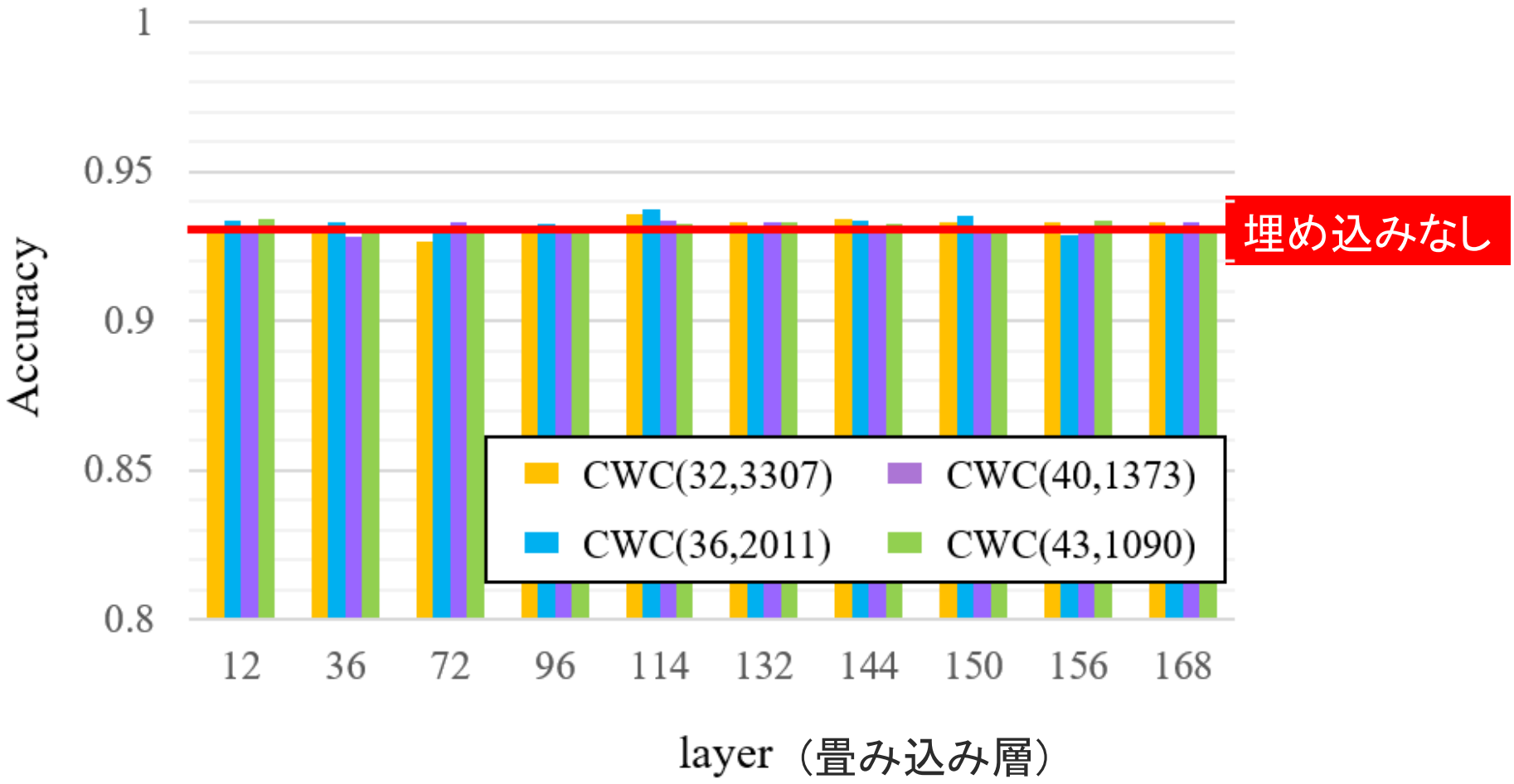
[1] T. Yasui, T. Tanaka, A. Malik, and M. Kuribayashi, “Coded DNN watermark: Robustness against pruning models using constant weight code,” Journal of Imaging, vol. 8, no. 6, 2022.
[2] M. Kuribayashi, T. Yasui, and A. Malik, “White box watermarking for convolution layers in fine-tuning model using the constant weight code,” Journal of Imaging, vol. 9, no. 6, 2023.

Accuracy (VGG16 畳み込み層)



256ビットの透かし情報の埋め込みの影響は確認されない

Accuracy (ResNet50 畳み込み層)



256ビットの透かし情報の埋め込みの影響は確認されない

耐性比較

耐性を確認した学習モデルに対するプルーニング率

手法	学習済みモデル	プルーニング攻撃	
		昇順	ランダム
Uchida [3]	WRN	0.65	0.65
Wang [4]	MLP / VGG	0.90 / 0.90	0.90 / 0.90
Li [5]	WRN	–	0.60
Proposed	VGG / RN	**0.97 / 0.92**	–

[3] Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh. Embedding watermarks into deep neural networks. In Proc. ICMR'17, pp.269–277, 2017.

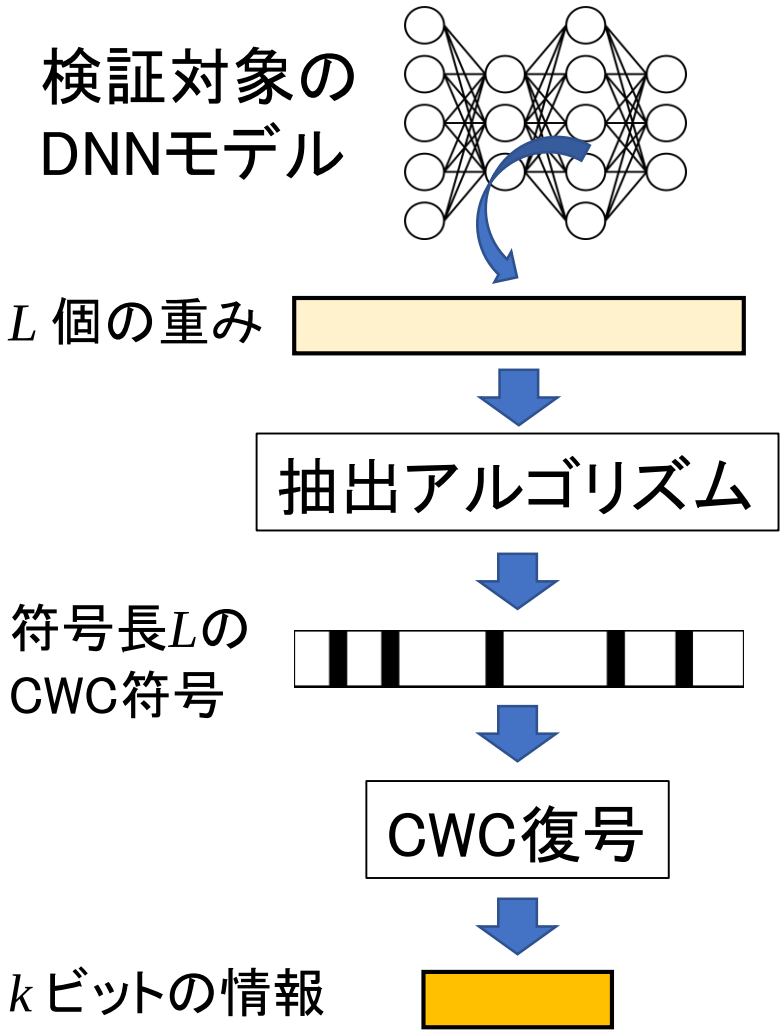
[4] J. Wang, H. Hu, X. Zhang, and Y. Yao. Watermarking in deep neural networks via error back-propagation. In IS&T Electronic Imaging, Media Watermarking, Security and Forensics, 2020.

[5] Y. Li, B. Tondi, and M. Barni. Spread-transform dither modulation watermarking of deep neural network. Journal of Information Security and Applications, vol.63, no.103004, 2021.

目次

- 研究背景
- DNNモデルへの電子透かし
- プルーニング攻撃
- 重み一定符号を用いた電子透かし
  - 重み一定符号
  - 埋め込み処理
- **透かし情報の検出と抽出**
- まとめ

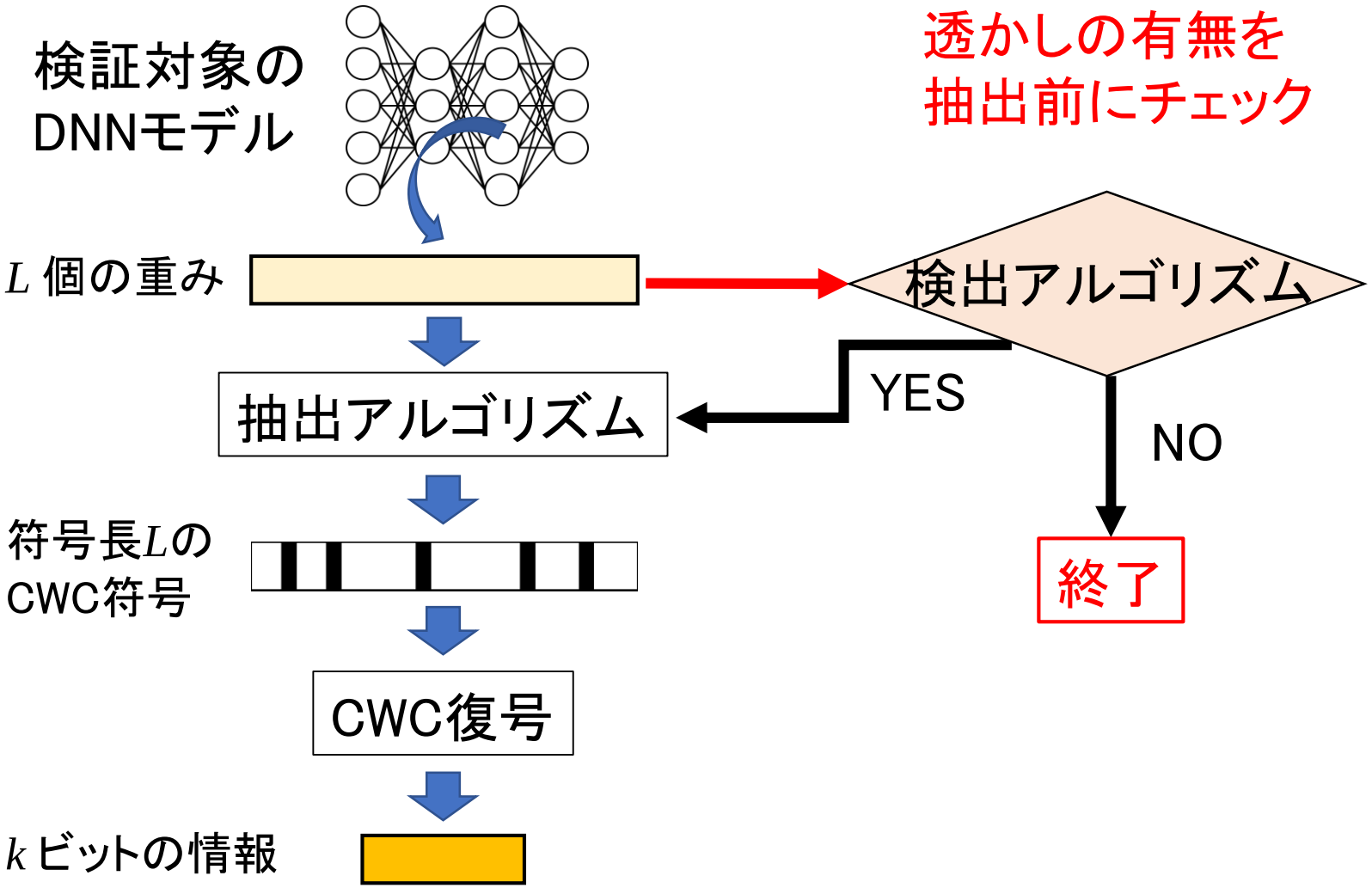
透かし情報の有無



問題点

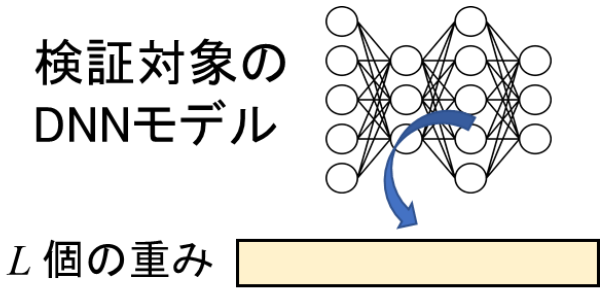
透かし情報が埋め込まれて
いなくても，何かは出力される

透かし情報の有無

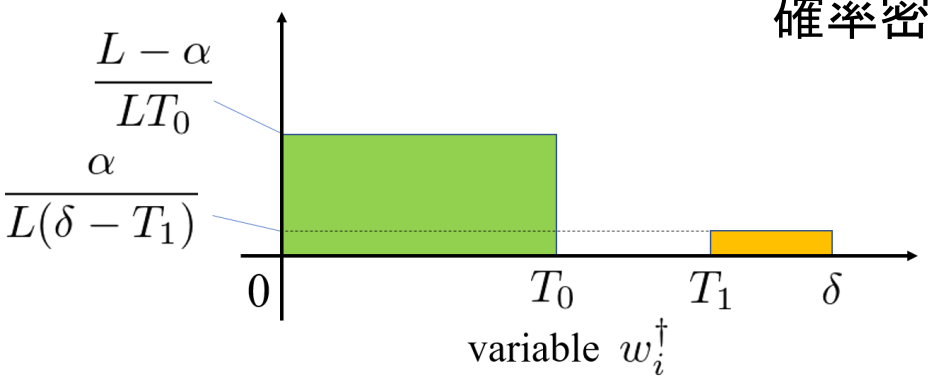

検証対象の
DNNモデル
透かしの有無を
抽出前にチェック
L 個の重み
検出アルゴリズム
YES
NO
抽出アルゴリズム
符号長Lの
CWC符号
終了
CWC復号
k ビットの情報

埋め込み後の重み

- 元のDNNモデルの重みの確率分布が一様分布の場合

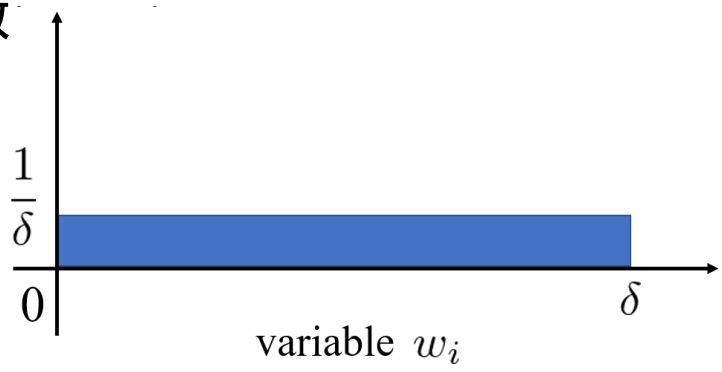


透かし有



α 個の重みは値が大きく，他は小さい

透かし無

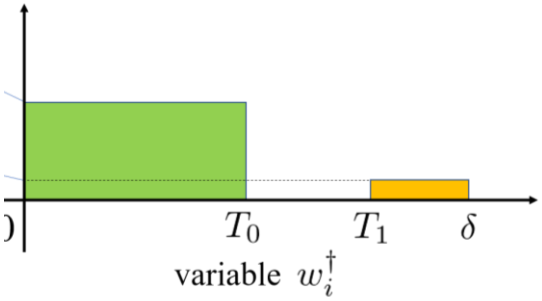


一様分布

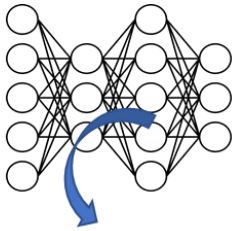
確率分布の偏りから，判別指標を求めたい

判別指標

- 確率分布の偏りから，判別指標を求める

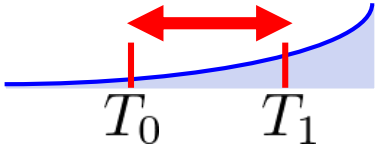


検証対象の
DNNモデル



L 個の重み

昇順にソート



$$T_0 \leq \tilde{c}_i \leq T_1$$

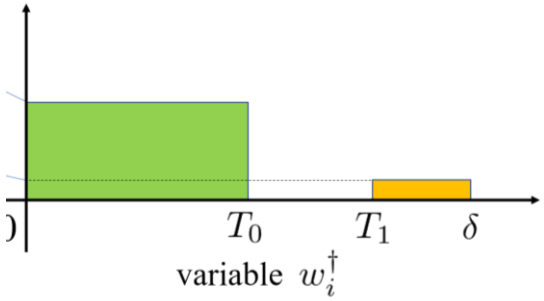
CWC符号語の場合　$\tilde{c}_i \approx T_0$

ランダム系列の場合　$\tilde{c}_i > T_0$

判別指標

$$MSE = \frac{1}{\eta - \alpha + 1} \sum_{i=\alpha}^{\eta} (\tilde{c}_i - T_0)^2$$

判別指標の期待値



CWC符号語の場合

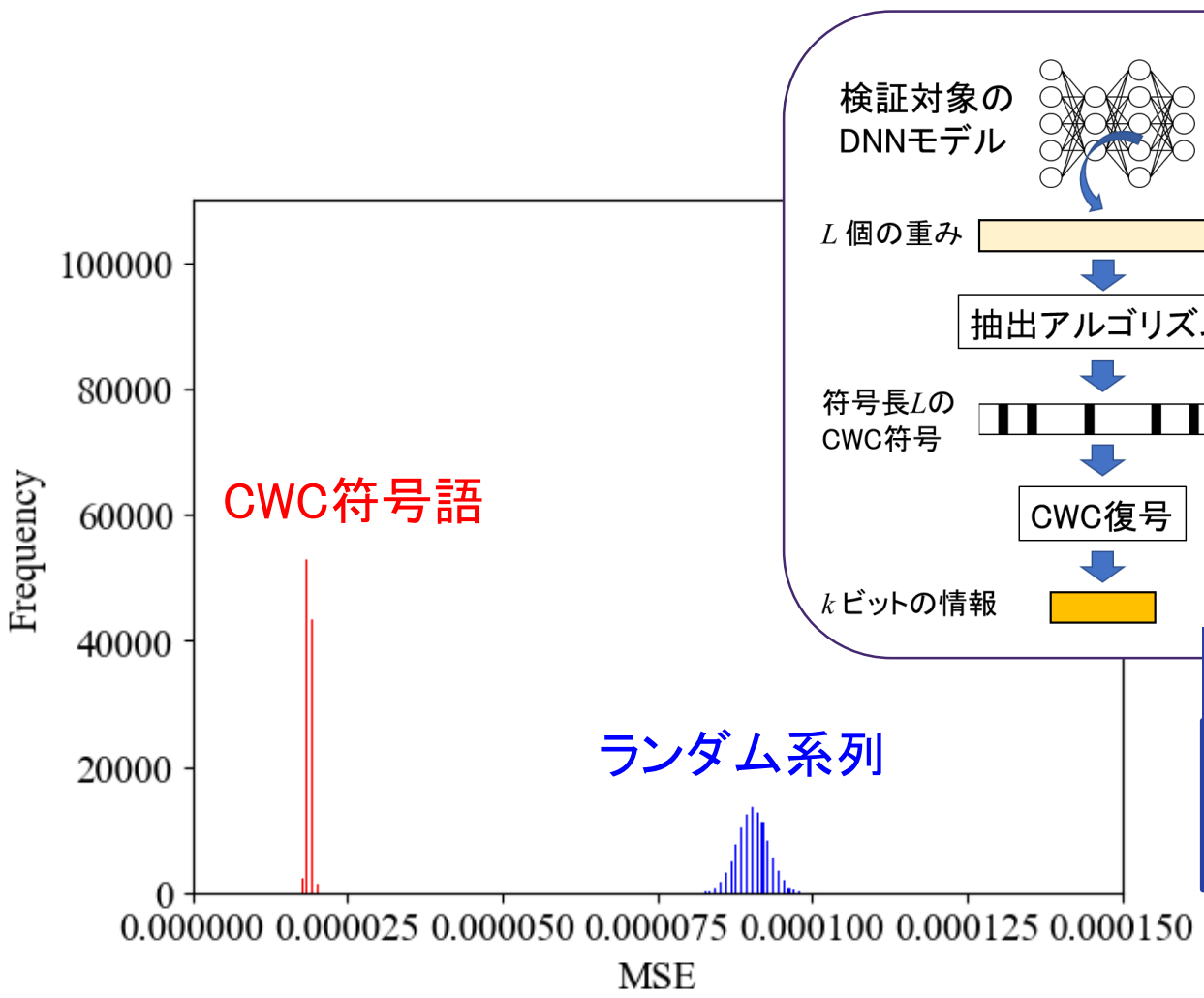
$$\tilde{E}_C^{\text{unif}} = \int_{T_0}^{T_1} (x - T_0)^2 \cdot 0 dx = 0$$

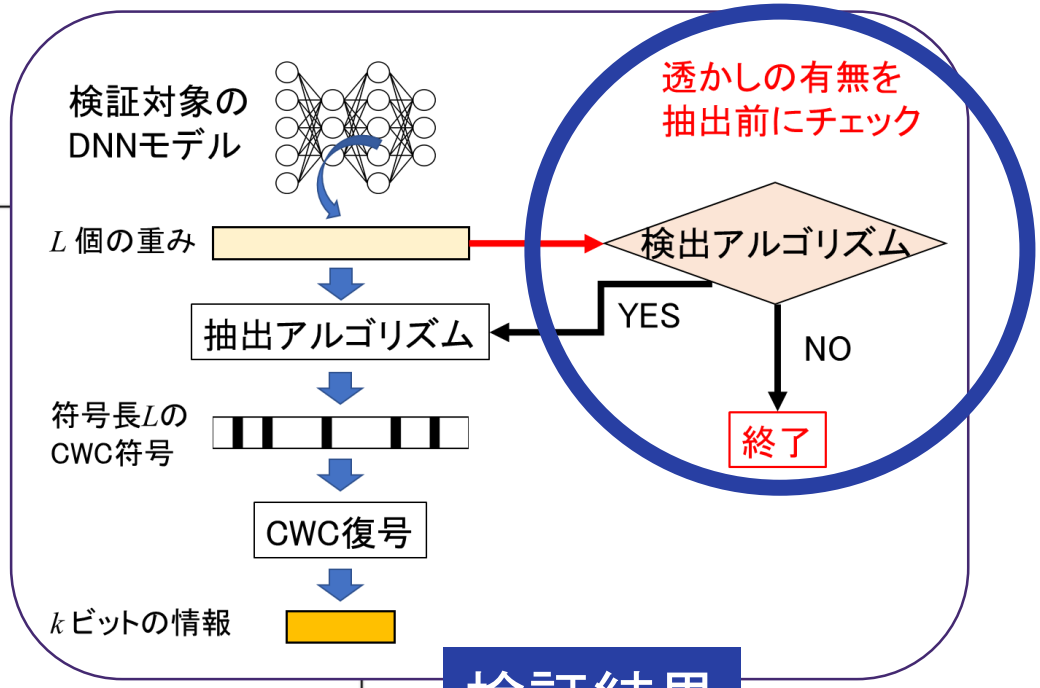
ランダム系列の場合

$$\tilde{E}_N^{\text{unif}} = \int_{T_0}^{T_1} (x - T_0)^2 \cdot \frac{1}{2\delta} dx + \int_{-T_1}^{-T_0} (x + T_0)^2 \cdot \frac{1}{2\delta} dx$$

$$= \frac{1}{\delta} \int_0^{T_1 - T_0} x^2 dx = \frac{1}{3\delta} (T_1 - T_0)^3$$

判別指標の分布





検証結果

透かし情報の有無を確認可能

目次

- 研究背景
- DNNモデルへの電子透かし
- プルーニング攻撃
- 重み一定符号を用いた電子透かし
  - 重み一定符号
  - 埋め込み処理
- 透かし情報の検出と抽出
- **まとめ**

まとめ

- DNNモデル保護に向けた電子透かし技術の研究紹介

  忠実度(Fidelity) ⇒ 精度(識別問題なら識別精度)

  容量(Capacity) ⇒ 学習モデルの重み 等

  ロバスト性(Robustness) ⇒ 透かしを除去する攻撃への耐性

- 重み一定符号を用いて消失誤りに対処

  符号語のシンボル”0”は消失しても影響がない

- 透かし情報の **検出** と **抽出**

透かし情報の有無をチェック

透かし情報のバイナリ系列を取り出す