Revisiting the Information Capacity of Neural Network Watermarks: Upper Bound Estimation and Beyond

Fangqi Li, Haodong Zhao, Wei Du, Shilin Wang*

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
{solour_lfq, zhaohaodong, dddddw, wsl}@sjtu.edu.cn

Abstract

To trace the copyright of deep neural networks, an owner can embed its identity information into its model as a watermark. The capacity of the watermark quantify the maximal volume of information that can be verified from the watermarked model. Current studies on capacity focus on the ownership verification accuracy under ordinary removal attacks and fail to capture the relationship between robustness and fidelity. This paper studies the capacity of deep neural network watermarks from an information theoretical perspective. We propose a new definition of deep neural network watermark capacity analogous to channel capacity, analyze its properties, and design an algorithm that yields a tight estimation of its upper bound under adversarial overwriting. We also propose a universal non-invasive method to secure the transmission of the identity message beyond capacity by multiple rounds of ownership verification. Our observations provide evidence for neural network owners and defenders that are curious about the tradeoff between the integrity of their ownership and the performance degradation of their products.

Introduction

The intellectual property protection of artificial intelligence models, especially deep neural networks (DNN), is drawing increasing attention since the expense of building large models has become prohibitively high. For example, the training of GPT-4 involves over 24,000 graphic processing units and more than 45TB manually proofread data (Liu et al. 2023). If internal enemies steal and distribute the model, watermarking schemes continue to safeguard the intellectual property.

As demonstrated in Fig. 1, a DNN watermarking scheme adds the owner's identity message into the model to be protected. Once the model is stolen, the owner can claim its copyright publicly by requesting a third-party judge to verify the identity message hidden in the victim model (Adi et al. 2018). Watermarking schemes have been designed for various kinds of DNN models including image classifier (Zhang et al. 2018), image generator (Quan et al. 2021), pretrained natural language encoder (Li et al. 2023), graph neural networks (Zhao, Wu, and Zhang 2021), etc.

Most discussions on the applicability of DNN watermarking schemes focus on unambiguity (Fan, Ng, and Chan 2019) and robustness (Lukas et al. 2022). It has been proven that incorporating pseudorandomness and connecting the triggers into a chain (Zhu et al. 2020) result in security against ambiguity attacks, i.e., an unauthorized party cannot pretend itself as the owner of an arbitrary DNN model. A robust ownership proof remains valid even if the watermarked DNN model undertakes adversarial modifications. There have been many methods to foster robustness including adding regularizer (Jia et al. 2021), training with surrogate models (Cong, He, and Zhang 2022), using error correction code for the identity message (Feng and Zhang 2020), etc.

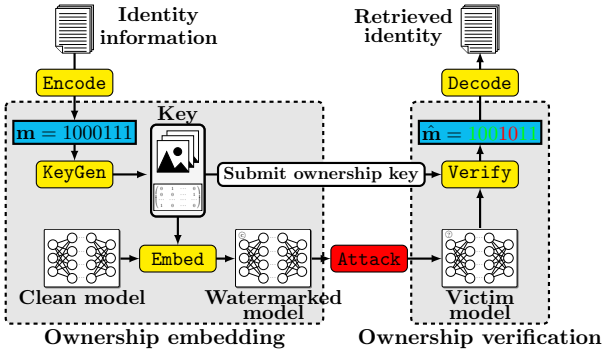


Figure 1: The workflow of a DNN watermarking scheme.

Nonetheless, the information capacity of DNN watermark, i.e., the maximal number of bits that can be accurately transmitted, has not undergone sufficient studies. So far, there is no uniform definition of capacity, especially for schemes that rely on backdoors. As a result, a fair comparison between different schemes regarding capacity is intractable. Moreover, the overlook of information theory-based capacity makes it hard to determine the configurations of DNN watermarking schemes with regard to the identity information and the expense of copyright protection.

To tangle with these deficiencies, we revisit the capacity of the DNN watermark. The contributions of this paper are:

- We give the first information theory-based definition of DNN watermark capacity and demonstrate how the expense of copyright protection is measured in the degradation of the watermarked model's performance.
- We design a capacity estimation algorithm that yields a

*Shilin Wang is the corresponding author.

tight upper bound of the capacity of DNN watermarks under adversarial overwriting.
- We propose a variational approximation-based method to increase the accuracy of identity message transmission beyond the capacity by multiple rounds of ownership verification. It can be generalized to arbitrary DNN watermarking schemes in a non-invasive manner.

Preliminaries

DNN Watermark

A DNN watermarking scheme adds the owner's identity information, encoded as a binary string $\mathbf{m}$ with length L, to a clean DNN model M_{clean} to produce a watermarked model M_{WM}. The owner first generates an ownership key K as the intermedium.

$$K \leftarrow \texttt{KeyGen}(L). \tag{1}$$

The owner then tunes M_{clean} into M_{WM} with the original training loss $\mathcal{L}_0$ and an additional regularizer $\mathcal{L}_{\text{WM}}$ as watermark embedding. This step minimizes:

$$\mathcal{L}_0(M_{\text{WM}}) + \lambda \cdot \mathcal{L}_{\text{WM}}(M_{\text{WM}}, K, \mathbf{m}). \tag{2}$$

Upon piracy, the owner submits K and requests a judge to retrieve the ownership information from a suspicious victim model M with a verifier that usually takes the form:

$$\texttt{Verify}(M, K) = \arg\min_{\mathbf{m}'} \left\{ \mathcal{L}_{\text{WM}}(M, K, \mathbf{m}') \right\}. \tag{3}$$

It is expected that if M is a copy or a slightly modified version of M_{WM} then $\texttt{Verify}(M, K) = \mathbf{m}$. A complete DNN watermarking scheme is featured by Eq. (1)(2)(3).

In scenarios where the judge has access to parameters in the victim model, the ownership key is usually defined as a pseudorandomly generated matrix and a bias term $K = (\mathbf{X}, \mathbf{b})$. The watermark embedding regularizer for white-box DNN watermarking schemes subject to this assumption is

$$\mathcal{L}_{\text{WM}}(M_{\text{WM}}, K, \mathbf{m}) = f(\sigma(\mathbf{X} \cdot \mathbf{W} + \mathbf{b}), \mathbf{m}), \tag{4}$$

where $\mathbf{W}$ denotes certain parameters in M_{WM} (Nagai et al. 2018) or the outputs of M_{WM}'s intermediate neurons (Li et al. 2022b). Considering $\mathbf{m}$ as a list of binary labels, f can be cross-entropy loss (Nagai et al. 2018), hingle loss (Fan, Ng, and Chan 2019), or a neural network classification backend (Wang et al. 2022).

If the judge has only black-box access to the victim model, the ownership key is a set of triggers $K = \{\mathbf{t}_n\}$. The embedding loss takes the form:

$$\mathcal{L}_{\text{WM}}\left(M_{\text{WM}}, K, \mathbf{m}\right) = f\left(g\left(\{M_{\text{WM}}(\mathbf{t}_n)\}\right), \mathbf{m}\right), \tag{5}$$

in which g interprets the outputs of the triggers into a list of bits. For example, for an eight-class classification model, a vanilla interpreter maps the prediction of each trigger into three bits. So $\mathbf{m}$ is compactly represented by the labels of $L/3$ triggers and watermark embedding is equivalent to fine-tuning to fit the triggers. In cases of image generators, the interpreters are complex decoders as in steganography (Zhang et al. 2021).

In both white-box and black-box cases, it is necessary that the gradient of the model's parameters w.r.t. Eq. (4)(5) is computable so the watermark embedding process becomes an optimization task.

Information Capacity of Watermark

Capacity is a fundamental property of watermark along with fidelity and robustness (Lei et al. 2019). It regulates the upstream identity information encoder since a message carrying more information than the capacity cannot be transmitted losslessly. As an example, in the scope of digital image watermarking, the capacity of a pixel is subject to both the pixel's contribution to the overall visual quality and its stability under standard image processing schemes such as compression or obfuscation (Moulin 2001). Embedding too many bits into one pixel results in visible distortion but embedding too few bits might fail to deliver the message.

Related Works

Li et al. derived an upper bound of DNN watermark capacity by examining if parameters where the watermarks are embedded have collusions or not (Li et al. 2022a). Many established works treat the length of the secret message as an upper bound of the capacity (Li, Tondi, and Barni 2021). In black-box schemes, the capacity is straightforward measured by the number of triggers (Zhang et al. 2020). Unlike in traditional watermark, these definitions overlook the correlation between the damage of watermarking to the performance of the model and the influence of adversarial modifications.

Information Capacity of DNN Watermark

Definition

We consider DNN watermark as a channel from the identity message $\mathbf{m}$ to the judge's observation $\hat{\mathbf{m}}$, both with length L. The adversarial modication $M_{\text{WM}} \to M_{\text{WM}} + \theta$ is featured by the parameter deviation θ. Fixing the upper bound of the victim model's performance degradation by $\delta \geq 0$, the capacity of the watermark is the maximal volume of information that can be correctly transmitted through the channel when the model's performance declines by no more than δ after undertaking arbitrary attacks. This property can be characterized through the channel capacity:

$$C(\delta, L) = \min_{\theta} \left\{ \max_{p(\mathbf{m})} I(\mathbf{m}; \hat{\mathbf{m}}) \right\}, \tag{6}$$

subject to:

$$\hat{\mathbf{m}} = \texttt{Verify}(M_{\text{WM}} + \theta, K),$$
$$E(M_{\text{WM}} + \theta) \geq E(M_{\text{WM}}) - \delta,$$

in which $E(\cdot)$ is the performance evaluation metric of the DNN model, $p(\mathbf{m})$ is the distribution over $\mathbf{m}$, and $I(\cdot;\cdot)$ denotes mutual information. The capacity of the DNN watermark satisfies the following properties.

Theorem 1. ***(Monotonicity)*** *$0 \leq C(\delta, L) \leq L$. $C(\delta, L)$ decreases in δ. $C(\delta, L)$ increases in L if each bit of the identity message is independently embedded and retrieved.*

Proof. Denote $\Theta(\delta) = \{\theta : E(M_{\text{WM}} + \theta) \geq E(M_{\text{WM}}) - \delta\}$ and $u(\theta, L) = \max_{p(\mathbf{m})} I(\mathbf{m}; \hat{\mathbf{m}})$.

As the mutual information between L binary random variables, $0 \leq I(\mathbf{m}; \hat{\mathbf{m}}) \leq L$, so does $C(\delta, L)$.

For the second statement, we prove that if $0 \leq \delta_1 \leq \delta_2$ then $C(\delta_1, L) \geq C(\delta_2, L)$. Since $\delta_1 \leq \delta_2$ implies $\Theta(\delta_1) \subset \Theta(\delta_2)$, it is evident that:

$$C(\delta_1, L) = \min_{\theta \in \Theta(\delta_1)} \{u(\theta, L)\} \geq \min_{\theta \in \Theta(\delta_2)} \{u(\theta, L)\} = C(\delta_2, L).$$

Finally, we prove $C(\delta, L+1) \geq C(\delta, L)$. By definition, there is θ' such that the capacity is $C(\delta, L+1)$ when the length of the identity message is $L+1$ and $E(M_{\text{WM}}+\theta') \geq E(M_{\text{WM}}) - \delta$. The DNN model watermarked with a message of length $(L+1)$ can be viewed as a model watermarked with a message of length L by considering only the first L bits during ownership verification, i.e.,

$$C(\delta, L) = \min_{\theta \in \Theta(\delta)} \{u(\theta, L)\} \leq u(\theta', L) \leq u(\theta', L+1) = C(\delta, L+1),$$

where the second inequality holds since the $(L+1)$-th independent bit brings additional mutual information. □

Theorem 2. ***(BER upper bound)*** *Denote the bit error rate corresponding to adversarial modification θ as:*

$$\epsilon(\theta) = \min \left\{ \frac{1}{2}, \frac{\|\mathbf{m} \oplus \texttt{Verify}(M_{WM}+\theta, K)\|_0}{L} \right\}. \quad (7)$$

Denote $\epsilon_\delta = \max_{\theta \in \Theta(\delta)} \{\epsilon(\theta)\}$ as the maximal BER when the model's performance drops for no more than δ. The capacity of the DNN watermark is upper bounded by:

$$C(\delta, L) \leq L \cdot (1 - H(\epsilon_\delta)), \quad (8)$$

with $H(x) = -x \cdot \log_2 x - (1-x) \cdot \log_2 (1-x)$. If each bit of the identity message is independently embedded and verified then Eq. (8) *is an equality.*

Proof. DNN watermark can be viewed as the combination of L binary symmetric channels, where each channel's average capacity is $(1 - H(\epsilon_\delta))$ since the adversary that minimizes the mutual information exerts the maximal BER. The capacity of L paralleling channels is no larger than the summation of their capacities, this yields Eq. (8). If each bit of $\mathbf{m}$ is independent of each other then the capacity equals the summation of all independent channels. □

Assume the owner's identity information contains J bits, whose source might be a digital copyright administrator authority as shown in Fig. 1. To resist adversarial removals, the owner encodes its identity into $\mathbf{m}$ with length $L \geq J$ by injecting redundancy. The cost in fidelity is measured by the performance degradation due to watermark embedding $F(L) = E(M_{\text{clean}}) - E(M_{\text{WM}})$. Intuitively, $F(L)$ increases in L. On the other hand, the scheme's robustness is reflected by the performance degradation that the adversary has to undertake to sabotage the ownership, i.e., $\min_\delta \{\delta : C(\delta, L) \leq J\}$, which increases in L due to Theorem 1.

The owner's objective is to protect the intellectual property of any DNN model whose performance is comparable to its state-of-the-art model M_{clean}. Models whose functionality has been severely damaged are not worth protecting. Formally, the tradeoff between fidelity and robustness in the context of capacity is described in the following theorem.

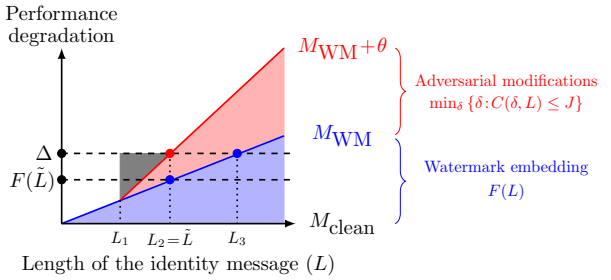


Figure 2: The performance degradation v.s. the length of the identity message. The blue area and the red area represent the cost in fidelity and robustness respectively.

Theorem 3. ***(The minimal length of the identity message)*** *Assume the owner's identity information contains J bits and the owner wants to protect the copyright of all variants of M_{WM} with performance degradation no more than Δ compared to M_{clean}. The necessarily minimal length of the identity message is:*

$$\tilde{L} = \min_L \left\{ L : \left(F(L) + \min_\delta \{\delta : C(\delta, L) \leq J\} \right) \geq \Delta \right\}, \quad (9)$$

and the necessarily minimal expense of copyright protection measured in the model's performance degradation is $F(\tilde{L})$.

Proof. The monotonicity of fidelity and robustness are visualized in Fig. 2, where $L_1 = J$, $L_2 = \tilde{L}$, and $L_3 = \min_L \{L : F(L) \geq \Delta\}$.

If $L \in [0, L_1)$ then the identity information cannot be losslessly encoded as $\mathbf{m}$. If $L \in [L_1, L_2)$ then the adversary can find a model whose performance degradation is no more than Δ yet escapes the ownership verification. Such cases constitute the gray area in Fig. 2. If $L \in [L_2, L_3)$ then the adversary cannot reduce the capacity below J to escape the ownership verification unless sacrificing the model's performance for more than Δ. Finally, $L \in [L_3, \infty)$ is unacceptable since the damage caused by watermarking is too much.

Therefore, $\tilde{L}$ is minimal length of the identity message that satisfies the owner's purposes with the smallest performance degradation and sufficient robustness. □

Capacity Estimation

The capacity defined by Eq. (6) can hardly be analytically computed. Instead, we resort to Theorem 2 by measuring the correlation of the BER of the identity message with the performance degradation and conducting interpolation afterwards. An adversarial modification strategy $\mathcal{A}$ is evoked to drive the variations. This estimation is formulated in Algo. 1.

Theorem 4. *The estimated capacity $\hat{C}(\delta, L)$ by Algo. 1 is an upper bound of $C(\delta, L)$.*

Proof. The adversarial modification strategy $\mathcal{A}$ exerts one certain family of parameter perturbation $\hat{\Theta}(\delta) \subset \Theta(\delta)$. The BER exclusively depends on θ according to Eq. (7), so:

$$\epsilon_\delta = \max_{\theta \in \Theta(\delta)} \{\epsilon(\theta)\} \geq \max_{\theta \in \hat{\Theta}(\delta)} \{\epsilon(\theta)\} = \hat{\epsilon}_\delta,$$

Algorithm 1: Capacity estimation.

Input: M_{WM}, K, $\mathbf{m}$, L, $\mathcal{A}$
Output: Capacity estimation $\hat{C}(\delta, L)$

1: Initialize a memory $\mathbf{D} = \{\}$.
2: Initialize $\hat{M} = M_{\text{WM}}$, $\delta = 0$.
3: $\hat{\epsilon}_0 = \min\left\{\frac{1}{2}, \frac{\|\mathbf{m} \oplus \texttt{Verify}(\hat{M}, K)\|_0}{L}\right\}$.
4: Save $\langle 0, \hat{C}(0, L) = L \cdot (1 - H(\hat{\epsilon}_0))\rangle$ in $\mathbf{D}$.
5: **while** $\hat{\epsilon}_\delta <= 0.5$ **do**
6: Conduct adversarial modification $\hat{M} \leftarrow \mathcal{A}(\hat{M})$.
7: $\delta = E(M_{\text{WM}}) - E(\hat{M})$.
8: Compute $\hat{\epsilon}_\delta$ as line 3.
9: Save $\langle \delta, \hat{C}(\delta, L) = L \cdot (1 - H(\hat{\epsilon}_\delta))\rangle$ in $\mathbf{D}$.
10: **end while**
11: Return $\mathbf{D}$.

Attack method	Loss function to be optimized
Fine-tune	$\mathcal{L}_0(M_{\text{WM}})$
Overwrite	$\mathcal{L}_0(M_{\text{WM}}) + \lambda \cdot \mathcal{L}_{\text{WM}}(M_{\text{WM}}, K_{\text{Adv}}, \mathbf{m}_{\text{Adv}})$
Adversarial overwrite	$\mathcal{L}_0(M_{\text{WM}}) + \lambda \cdot \mathcal{L}_{\text{WM}}(M_{\text{WM}}, K, \mathbf{m}_{\text{Adv}})$

Table 1: Comparison between adversarial modifications. $\mathcal{L}_0$ is the model's original training loss function. $\mathbf{m}_{\text{Adv}}$ and K_{Adv} are the adversary's identity message and ownership key.

since the r.h.s. of Eq. (8) is a monotonically decreasing function in ϵ (when $\epsilon <= 0.5$), we have:

$$C(\delta, L) \le L \cdot (1 - H(\epsilon_\delta)) \le L \cdot (1 - H(\hat{\epsilon}_\delta)) = \hat{C}(\delta, L).$$

□

If the adversarial modification strategy $\mathcal{A}$ is not destructive enough, i.e., $\hat{\epsilon}_\delta \ll \epsilon_\delta$, then the capacity would be overestimated. Consequently, the minimal length of the identity message recommended by Theorem 3 after replacing $C(\delta, L)$ by $\hat{C}(\delta, L)$ is smaller and unsafe. To tighten the bound, we conduct an adversarial overwriting attack that directly embeds another randomly generated message into the pirated model with the same ownership key as shown in Table 1.

In contrast to other advanced attacks such as reverse engineering (Wang and Kerschbaum 2021), or functionality equivalence attack (Li, Wang, and Alan 2023), adversarial overwriting can be directly generalized to any gradient-based DNN watermarking scheme by manipulating the watermark embedding process. Adversarial overwriting adopts the strongest threat model where the adversary has full knowledge of the ownership key and covers the worst cases, e.g., the ownership verification has been exposed to malicious parties or there are internal enemies. Therefore, the bound is expected to be the tightest and provides a reliable reference for owners.

Breaking the Capacity Bottleneck

Let the number of all legal owners or models be 2^J, so the identity information of each owner contains J bits. In one-time DNN ownership verification, the code rate, i.e., the volume of information transmitted in each round of communication, is $\frac{\log_2 2^J}{1} = J$ bits. According to basic information theory, the code rate cannot exceed the capacity for accurate communication, therefore $J \le C(\delta, L) \le L$.

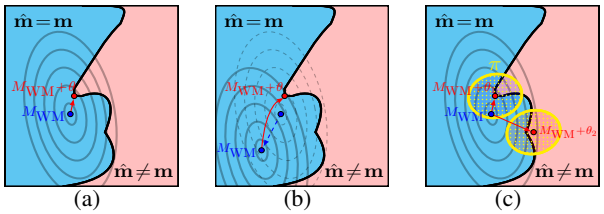


Figure 3: The contours denote levels of performance degradation. (a) An adversarial modification. (b) Increasing robustness as a defense. (c) Averaging multiple rounds of ownership verification as a defense. $M_{\text{WM}}+\theta_1$ denotes a failed attack. $M_{\text{WM}}+\theta_2$ denotes a successful attack.

When L is fixed (so the price of watermarking is no more than $F(L)$), there are two approaches to increasing the number of entities that can use the copyright protection service. The first is to increase the capacity, whose necessary prerequisite is reducing the BER by Theorem 2. Yet this additional robustness might increase the performance degradation due to watermark embedding as shown in Fig. 3(b).

The other approach is to retrieve the identity message for R times and average the results. The motivation is that the adversary erases the watermark with the smallest performance degradation, so the victim model falls at the decision boundary close to the watermarked model. So it is expected that the correct identity message can still be retrieved from neighbours of the victim model as shown in Fig. 3(c). Under this setting, the code rate becomes J/R bits and at most 2^L owners or models are verifiable with $R \ge \frac{L}{C(\delta, L)}$.

However, the adversary is not obligated to use random modification during multiple rounds of ownership verification. So the error pattern $\mathbf{m} \oplus \hat{\mathbf{m}}$ could be fixed and cannot be eliminated by averaging. Instead, it is the judge's responsibility to incorporate randomness into the channel by perturbing the victim model and alleviating adversarial influences.

To randomize the noises in the channel, the judge incorporates parameter deviations μ following a distribution π. In ownership verification, the identity message is retrieved by averaging the results as decoding the error correction code. Concretely, the i-th bit is determined by the majority voting:

$$\hat{\mathbf{m}}[i] = \arg\max_{b \in \{0,1\}} \left\{ \sum_{r=1}^{R} \mathbb{I}\left[\texttt{Verify}(M + \mu_r, K)[i] = b\right] \right\}. \tag{10}$$

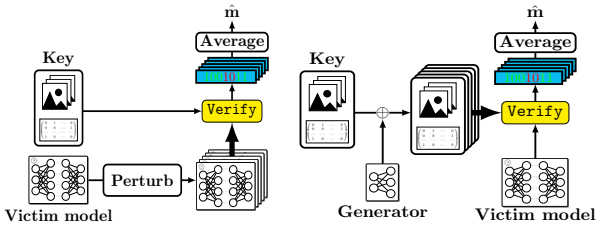
The owner can further reduce the sensitivity of the identity message against potential adversarial modifications with the following regularizer during watermark embedding:

$$\mathcal{L}^M_{\text{WM}}(M_{\text{WM}}, K, \mathbf{m}) = \frac{1}{P} \sum_{p=1}^{P} \mathcal{L}_{\text{WM}}(M_{\text{WM}} + \mu_p, K, \mathbf{m}). \tag{11}$$

It has be proven that if π is a normal distribution $\mathcal{N}(0, \sigma^2 \mathbf{I})$ and the scale of adversarial modification is restricted by

Scheme	Verifier	Embedding regularizer	Applicability to black-box
Baseline	Eq. (3)	$\mathcal{L}_{\text{WM}}$	Yes
MROV	Eq. (10)	$\mathcal{L}_{\text{WM}}$	No
Certified robustness	Eq. (10)	Eq. (11)	No
MROV-V-1	Eq. (14)	$\mathcal{L}_{\text{WM}}$	Yes
MROV-V-2	Eq. (14)	Eq. (15)	Yes

Table 2: Comparison between configurations of the watermark embedding regularizer and the verification formula. MROV and MROV-V denote Multiple Rounds of Ownership Verification and its Variational version respectively.



(a) MROV. (b) MROV-V.

Figure 4: Multiple rounds of ownership verification, conducted by the judge. (a) can only be applied to white-box schemes. (b) can be applied to any scheme.

$\|\theta\|_2 \leq \rho$ then the BER of the estimation Eq. (10) is upper bounded by $\inf\{y : \Pr\{\epsilon(\theta) \geq y\} \leq \Phi(-\rho/\sigma)\}$, where $\Phi(\cdot)$ is the culmulative distribution function of a normal distribution (Bansal et al. 2022).

Such a paradigm cannot be applied to black-box DNN watermarking schemes since the parameters within the victim model are hidden so μ is nowhere to be added. In general cases, the judge resorts to the ownership key K, which is always available. Concretely, a distribution π' over the ownership key is explored as a substitution of π such that the distribution over the embedding loss remains invariable, i.e., for any real number t:

$$\Pr_{\mu\leftarrow\pi}(\mathcal{L}_{\text{WM}}(M_{\text{WM}}+\mu, K, \mathbf{m}) \leq t) = \Pr_{\kappa\leftarrow\pi'}(\mathcal{L}_{\text{WM}}(M_{\text{WM}}, K+\kappa, \mathbf{m}) \leq t), \tag{12}$$

The distribution π' could be very complicated, especially for black-box schemes where the ownership key is a set of images or texts. Therefore, we implement the distribution transfer by fitting π' with a parameterized generator G.

A collection of parameters pertubations $\{\mu_q\}_{q=1}^{Q}$ are randomly sampled from π. Then it is transformed into perturbations on the ownership key $\{\kappa_q\}_{q=1}^{Q}$ subject to:

$$\forall q, \mathcal{L}_{\text{WM}}(M_{\text{WM}}+\mu_q, K, \mathbf{m}) = \mathcal{L}_{\text{WM}}(M_{\text{WM}}, K+\kappa_q, \mathbf{m}). \tag{13}$$

A variational autoencoder (Pu et al. 2016) is trained to reconstruct $\{\kappa_q\}_{q=1}^{Q}$, whose decoder is returned as G. The accuracy of this setting is established by the following theorem.

Theorem 5. *$\pi' = \{G(\mathbf{z}) : \mathbf{z} \leftarrow \mathcal{N}(0, \mathbf{I})\}$ satisfies Eq. (12).*

Proof. Statistically, the l.h.s. of Eq. (12) equals the portion of samples μ satisfying $\mathcal{L}_{\text{WM}}(M+\mu, K, \mathbf{m}) \leq t$. With $\pi' = \{G(\mathbf{z}) : \mathbf{z} \leftarrow \mathcal{N}(0, \mathbf{I})\}$, the r.h.s. of Eq. (12) is reduced to:

$$\int \Pr(\mathbf{z}) \cdot \mathbb{I}\left[(\mathcal{L}_{\text{WM}}(M_{\text{WM}}, K+G(\mathbf{z}), \mathbf{m}) \leq t\right] \mathrm{d}\mathbf{z} = \frac{|\{q : \mathcal{L}_{\text{WM}}(M_{\text{WM}}+\mu_q, K, \mathbf{m}) \leq t\}|}{Q},$$

the second equation holds since G assumes its inputs as a normal distribution, the last equation holds by Eq. (13). □

Having obtained the ownership key perturbation generator G, the message is retrieved by:

$$\hat{\mathbf{m}}[i] = \arg\max_{b\in\{0,1\}} \left\{ \sum_{r=1}^{R} \mathbb{I}\left[\texttt{Verify}(M, K+\kappa_r)[i] = b\right] \right\}. \tag{14}$$

Theoretically, Eq. (14) is the variational approximation version of Eq. (10) so the results are expected to be similar if G accurately captures the influence of π. Since G depends on a watermarked model, the owner can watermark its model, run the variational distribution transfer program, and further finetune the watermarked model with the following regularizer to enhance the robustness:

$$\mathcal{L}_{\text{WM}}^{K}(M_{\text{WM}}, K, \mathbf{m}) = \frac{1}{P}\sum_{p=1}^{P} \mathcal{L}_{\text{WM}}(M_{\text{WM}}, K+\kappa_p, \mathbf{m}). \tag{15}$$

Different configurations of the watermarking scheme regarding the accurate transmission of the identity message are summarized in Table 2 and Fig. 4.

Experiments and Discussions

Settings

We studied the capacity properties of several representative DNN watermarking schemes. **Uchida** et al.'s scheme (Nagai et al. 2018) is a prototypical static parameter-based white-box watermarking scheme. The identity message is explicitly defined as Eq. (4). Spread-Transform Dither Modulation watermarking **(STDM)** (Li, Tondi, and Barni 2021) is a variant of Uchida et al.'s scheme with STDM activation function. **MTLS**ign (Li et al. 2022b) is a dynamic white-box scheme. Each bit of the identity message is retrieved from the prediction for a trigger returned from a binary classifier based on hidden neurons' responses. **Content** is a representative pattern of triggers defined in Zhang et al.'s black-box scheme (Zhang et al. 2018). **Expo**nential-weighting (Namba and Sakuma 2019) uses normal training sample as triggers with exponential regularizer during watermark embedding. **Front**-stitching (Merrer, Pérez, and Trédan 2020) uses samples close to the decision boundary with adversarial perturbations as triggers.

The task is classification on CIFAR-10 (Krizhevsky, Hinton et al. 2009). The architecture of DNN model to be protected is ResNet-50 (He et al. 2016). To compile the identity message into classification predictions, we adopted the vanilla interpreter and mapped the prediction of each trigger into $\lfloor \log_2 10 \rfloor = 3$ bits. All experiments were implemented in `PyTorch` framework [1].

[1]Source codes are available upon request.

Δ	(a) Minimal length of the identity message $\tilde{L}$ (100 bit).						(b) Minimal performance degradation due to watermarking $F(\tilde{L})$ (%).					
	Uchida	STDM	MTLS	Content	Expo	Front	Uchida	STDM	MTLS	Content	Expo	Front
1%	16±0.5	39±3	**13±1**	15±0.5	18±2	14.5±0.5	0.30±0.00	0.38±0.03	0.16±0.01	0.13±0.00	0.16±0.02	**0.12±0.00**
2%	21±2	≥81.92	**16±2**	19±0.5	28±8	18.5±0.5	0.33±0.01	≥0.79	0.20±0.03	0.17±0.00	0.24±0.07	**0.16±0.00**
3%	36±6	≥81.92	**19±3**	41±6	50±1	28±3	0.42±0.03	≥0.79	**0.24±0.04**	0.36±0.05	0.43±0.01	0.25±0.03

Table 3: Expense of copyright protection measured by the necessarily minimal length of the identity message (a) and performance degradation due to watermarking (b) computed by Theorem 3 when the ownership can be accurately retrieved from variants of the watermarked model with performance degradation no more than Δ. Bold entries are optimal configurations. $J = 1024$.

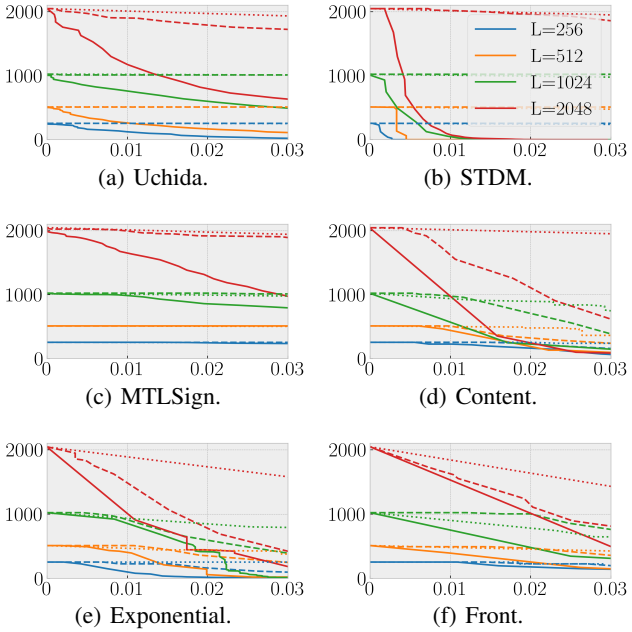


Figure 5: Estimated capacity $\hat{C}(\delta, L)$ as a function of the performance degradation δ under fine-tuning (·····), neuron-pruning (- - -), and adversarial overwriting (——). The length of the identity message L is set as 256, 512, 1024, and 2048.

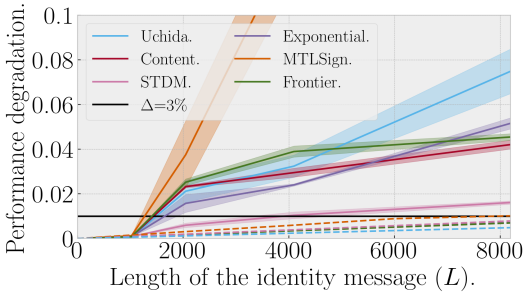


Figure 6: Estimated minimal length of the identity message, $J = 1024$. The dashed lines mark the cost in fidelity $F(L)$. The solid lines mark $F(L) + \min_\delta\{\delta : \hat{C}(\delta, L) \leq J\}$, shadow areas denote fluctuations during experiments.

Capacity Estimation

The dataset with which the adversary conducts the attacks was 10% of the entire training dataset. λ in adversarial overwriting formulated in Table 1 was set as 0.1. Fine-tuning and neuron-pruning were implemented as in (Lukas et al. 2022). Each configuration was repeated for five times, the mean capacity estimation of all watermarking schemes by Algo. 1 is illustrated in Fig. 5. We made the following observations.

(i) In general, $\hat{C}(\delta, L)$ declines in δ and increases in L as predicted by Theorem 1. (ii) Compared with universal watermark removal attacks with weaker assumptions on the adversary such as fine-tuning and neuron-pruning, adversarial overwriting yielded the lowest and tightest estimation of the capacity since it directly tampers with the watermark while preserves the model's performance simultaneously. (iii) Capacity varies with the watermarking scheme. As a covert version of **Uchida**, the capacity of **STDM** is extremely small since the embedded information be removed with slight modification and hence negligible performance degradation. This fact reflects the tradeoff between capacity and stealthiness. The capacity of black-box schemes depends on the pattern of triggers and has to be measured a posteriori.

As a conclusive evaluation of studied DNN watermarking schemes, we applied Theorem 3 and computed the minimal length of the identity message and the minimal performance degradation due to watermark embedding. We set $J = 1024$ as a standard digital signature scheme DSA-1024 with $\Delta = \{1\%, 2\%, 3\%\}$. As demonstrated in Fig. 6, the monotonicity is identical to predictions in Fig. 2, the fidelity $F(L)$ was computed and interpolated for $L = \{256, 512, 1024, 2048, 4096, 8192\}$. $\tilde{L}$ was estimated using Eq. (9) with a granularity of 50 bit, numerical results are given in Table 3. We noted that the scheme with the largest capacity (**MTLS**) had the smallest minimal identity message length, but the corresponding performance degradation is not always the lowest. Therefore, a good watermarking scheme should have both a large capacity and a high fidelity.

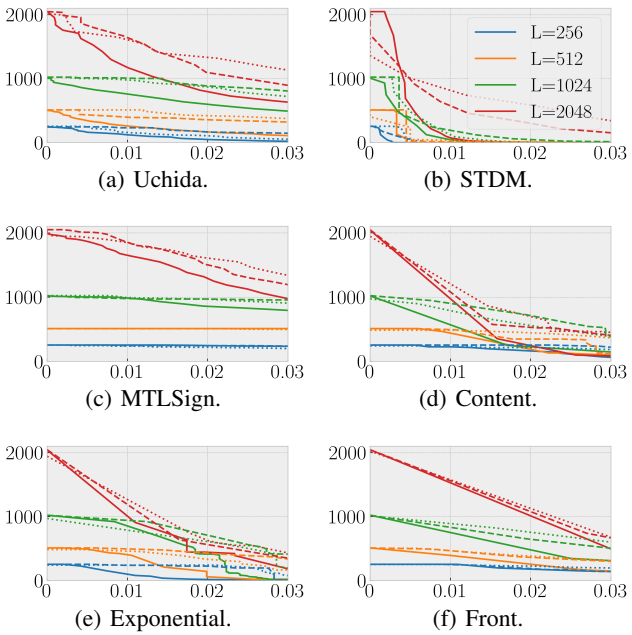
Efficacy of MROV-V

We verified the efficacy of the proposed variational multiple rounds of ownership verification (we focused on MROV-V-1 and MROV-V-2 that are applicable in the black-box settings) for six watermarking schemes.

The distribution of perturbations on the parameters π was set as a normal distribution as in (Bansal et al. 2022). The variational approximation distribution π' for trigger-based schemes including **MTLS**, **Content**, **Expo**, and **Front** was generated by a decoder with four deconvolutional layers. For **Uchida** and **STDM**, the decoder was a four-layer MLP. Each

Δ	Setting	(b) Minimal performance degradation due to watermarking $F(\tilde{L})$ (%).					
		Uchida	STDM	MTLS	Content	Expo	Front
1%	MROV-V-1	0.17±0.01	0.24±0.02	0.16±0.01	0.12±0.00	0.14±0.01	**0.12±0.00**
	MROV-V-2	0.23±0.01	0.42±0.01	0.23±0.01	0.24±0.01	0.25±0.02	0.24±0.01
2%	MROV-V-1	0.20±0.01	0.41±0.04	0.19±0.01	0.16±0.01	0.23±0.07	**0.14±0.00**
	MROV-V-2	0.28±0.01	0.74±0.03	0.27±0.01	0.29±0.02	0.36±0.05	0.28±0.01
3%	MROV-V-1	0.25±0.01	0.71±0.04	**0.23±0.01**	0.28±0.03	0.38±0.05	0.25±0.00
	MROV-V-2	0.34±0.01	1.04±0.07	0.32±0.01	0.38±0.06	0.47±0.09	0.34±0.01

Table 4: Expense of copyright protection using DNN watermarking schemes computed by Theorem 3 when the ownership can be accurately retrieved from variants of the watermarked model with performance degradation no more than Δ. $J = 1024$. Entries marked in shadow are configurations outperformed by the baseline in Table 3.



(a) Uchida. (b) STDM. (c) MTLSign. (d) Content. (e) Exponential. (f) Front.

Figure 7: Estimated capacity under adversarial overwriting under the baseline setting (——), MROV-V-1(- - -), and MROV-V-2 (·····). L is set as 256, 512, 1024, and 2048.

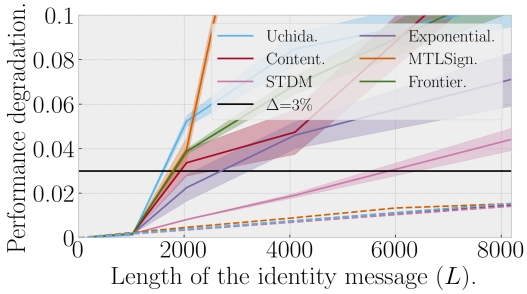


Figure 8: Estimated minimal length of the identity message under MROV-V-2, $J = 1024$.

autoencoder was trained on a collection of $Q = 10000$ instances of parameters perturbations. The number of rounds was fixed as $R = 100$. During watermark embedding of MROV-V-2 defined by Eq. (15), $P = 10$.

Empirically, both MROV-V-1 and MROV-V-2 reduced the BER during ownership verification under adversarial modifications. To provide a fair and uniform comparison between MROV-V-1/2 and the basic one-time ownership verification, we transformed the BER statistics into capacity upper bounds using Theorem 2 and visualized them in Fig. 7. It is observed that MROV-V-1/2 increases the bound so more information can be safely transmitted through DNN watermark. This is because that the adversary has to modify the victim model until all neighbours of the ownership key fail to expose the correct identiy message (instead of only the ownership key as in one-time verification). MROV-V-2 has the largest capacity since the adversary has to exert larger modifications to suppress the effect of regularizer in Eq. (15).

However, MROV-V-2 also introduces a worse fidelity, i.e., although $\min_\delta \{\delta : C(\delta, L) \leq J\}$ decreases, $F(L)$ also increases so $\tilde{L}$ computed by Eq. (9) and $F(\tilde{L})$ are not guaranteed to decline. This fact is demonstrated by Fig. 8. Numerically, we computed the corresponding performance degradation to the minimal length of the identity message in Table 4. We remark that both types of MROV-V reduce the minimal length of the identity message, but only MROV-V-1 can always reduce the minimal performance degradation due to watermarking since $F(L)$ is left invariable. In all settings of **MTLS**, **Content**, **Expo**, and **Front**, MROV-V-2 yields a higher cost because of the corrupted fidelity.

As a result, we recommend the configuration of MROV-V-1 which is universally applicable, non-invasive, and is promised to increase the capacity's upper bound and hence reduce the expense of DNN copyright protection.

Conclusions

This paper studies the information capacity of DNN watermarks. We propose a unified definition and show how it is related to the accuracy, robustness, and the expense of ownership verification. An estimation method is designed to find an upper bound of the capacity by adversarial overwriting. Finally, we demonstrate that the capacity bottleneck can be broken by reducing the code rate with multiple rounds of ownership verification and incorporate variational approximation on the ownership keys to expand the applicability. Experiments show that our scheme efficiently secure the ownership verification with no marginal performance degradation.

Acknowledgments

The work described in this paper was supported by the National Natural Science Foundation of China (62271307, 61771310) and the Joint Funds of the National Natural Science Foundation of China (U21B2020).

References

Adi, Y.; Baum, C.; Cissé, M.; Pinkas, B.; and Keshet, J. 2018. Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring. In *27th USENIX Security Symposium, USENIX Security 2018*, 1615–1631.

Bansal, A.; Chiang, P.; Curry, M. J.; Jain, R.; Wigington, C.; Manjunatha, V.; Dickerson, J. P.; and Goldstein, T. 2022. Certified Neural Network Watermarks with Randomized Smoothing. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, 1450–1465.

Cong, T.; He, X.; and Zhang, Y. 2022. SSLGuard: A Watermarking Scheme for Self-supervised Learning Pre-trained Encoders. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 579–593.

Fan, L.; Ng, K. W.; and Chan, C. S. 2019. Rethinking Deep Neural Network Ownership Verification: Embedding Passports to Defeat Ambiguity Attacks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, 4716–4725.

Feng, L.; and Zhang, X. 2020. Watermarking Neural Network with Compensation Mechanism. In *Knowledge Science, Engineering and Management - 13th International Conference, KSEM 2020*, volume 12275 of *Lecture Notes in Computer Science*, 363–375.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jia, H.; Choquette-Choo, C. A.; Chandrasekaran, V.; and Papernot, N. 2021. Entangled Watermarks as a Defense against Model Extraction. In *30th USENIX Security Symposium, 2021*, 1937–1954.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Citeseer Technical Report*.

Lei, B.; Zhao, X.; Lei, H.; Ni, D.; Chen, S.; Zhou, F.; and Wang, T. 2019. Multipurpose watermarking scheme via intelligent method and chaotic map. *Multimedia Tools and Applications*, 78: 27085–27107.

Li, B.; Fan, L.; Gu, H.; Li, J.; and Yang, Q. 2022a. FedIPR: Ownership Verification for Federated Deep Neural Network Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–16.

Li, F.; Wang, S.; and Alan, W.-C. L. 2023. Linear Functionality Equivalence Attack against Deep Neural Network Watermarks and a Defense Method by Neuron Mapping. *IEEE Transactions on Information Forensics and Security*, 1–14.

Li, F.; Yang, L.; Wang, S.; and Liew, A. W. 2022b. Leveraging Multi-task Learning for Unambiguous and Flexible Deep Neural Network Watermarking. In *Proceedings of the Workshop on Artificial Intelligence Safety 2022 (SafeAI 2022)*, volume 3087 of *CEUR Workshop Proceedings*.

Li, P.; Cheng, P.; Li, F.; Du, W.; Zhao, H.; and Liu, G. 2023. PLMmark: A Secure and Robust Black-box Watermarking Framework for Pre-trained Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1, 1–9.

Li, Y.; Tondi, B.; and Barni, M. 2021. Spread-Transform Dither Modulation Watermarking of Deep Neural Network. *J. Inf. Secur. Appl.*, 63: 103004.

Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; et al. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Lukas, N.; Jiang, E.; Li, X.; and Kerschbaum, F. 2022. SoK: How Robust is Image Classification Deep Neural Network Watermarking? In *43rd IEEE Symposium on Security and Privacy, 2022*, 787–804.

Merrer, E. L.; Pérez, P.; and Trédan, G. 2020. Adversarial frontier stitching for remote neural network watermarking. *Neural Comput. Appl.*, 32(13): 9233–9244.

Moulin, P. 2001. The role of information theory in watermarking and its application to image watermarking. *Signal Processing*, 81(6): 1121–1139.

Nagai, Y.; Uchida, Y.; Sakazawa, S.; and Satoh, S. 2018. Digital watermarking for deep neural networks. *Int. J. Multim. Inf. Retr.*, 7(1): 3–16.

Namba, R.; and Sakuma, J. 2019. Robust Watermarking of Neural Network with Exponential Weighting. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, 228–240.

Pu, Y.; Gan, Z.; Henao, R.; Yuan, X.; Li, C.; Stevens, A.; and Carin, L. 2016. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*, 2352–2360.

Quan, Y.; Teng, H.; Chen, Y.; and Ji, H. 2021. Watermarking Deep Neural Networks in Image Processing. *IEEE Trans. Neural Networks Learn. Syst.*, 32(5): 1852–1865.

Wang, L.; Wang, Z.; Li, X.; and Qin, C. 2022. Robust Watermarking for Neural Network Models Using Residual Network. In *24th IEEE International Workshop on Multimedia Signal Processing, 2022*, 1–6.

Wang, T.; and Kerschbaum, F. 2021. RIGA: Covert and Robust White-Box Watermarking of Deep Neural Networks. In *WWW '21: The Web Conference 2021*, 993–1004.

Zhang, J.; Chen, D.; Liao, J.; Zhang, W.; Feng, H.; Hua, G.; and Yu, N. 2021. Deep model intellectual property protection via deep watermarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4005–4020.

Zhang, J.; Gu, Z.; Jang, J.; Wu, H.; Stoecklin, M. P.; Huang, H.; and Molloy, I. M. 2018. Protecting Intellectual Property of Deep Neural Networks with Watermarking. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 159–172.

Zhang, Y.; Jia, Y.; Wang, X.; Niu, Q.; and Chen, N. 2020. DeepTrigger: A Watermarking Scheme of Deep Learning Models Based on Chaotic Automatic Data Annotation. *IEEE Access*, 8: 213296–213305.

Zhao, X.; Wu, H.; and Zhang, X. 2021. Watermarking Graph Neural Networks by Random Graphs. In *9th International Symposium on Digital Forensics and Security, ISDFS 2021*, 1–6.

Zhu, R.; Zhang, X.; Shi, M.; and Tang, Z. 2020. Secure neural network watermarking protocol against forging attack. *EURASIP J. Image Video Process.*, 2020(1): 37.