ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094751

# Model Fingerprinting with Benign Inputs

Thibault Maho, Teddy Furon and Erwan Le Merrer

*Univ. Rennes, Inria, CNRS*
IRISA, Rennes, France

***Abstract*—Recent advances in the fingerprinting of deep neural networks are able to detect specific instances of models, placed in a black-box interaction scheme. Inputs used by the fingerprinting protocols are specifically crafted for each precise model to be checked for. While efficient in such a scenario, this nevertheless results in a lack of guarantee after a mere modification of a model (*e.g.* finetuning, quantization of the parameters).**

**In this paper we propose fingerprinting scheme (coined FBI) that are resilient to significant modifications of the models. These modifications are viewed and modeled as variants. We demonstrate that benign inputs, that are unmodified images, are sufficient material for efficient fingerprinting. We leverage an information-theoretic approach to achieve a success rate of $95.2\%$. It is experimentally validated over an unprecedented set of more than 1,000 neural networks, while demonstrating performance improvements over a state-of-the-art fingerprinting method.[1].**

***Index Terms*—Fingerprinting, Deep Neural Networks, Information Theory**

## I. Introduction

Fingerprinting classifiers aims at deriving a signature uniquely identifying a machine learning model, like the human fingerprint's minutiae in biometry. This is essentially a black-box problem: the classifier to be identified is in a black-box in the sense that one can just make some queries and observe the model decision. For instance, this is the case when the model is embedded in a chip, or accessible through an API.

The main application that related works [2], [3], [4], [5], [6] target is the proof of ownership. An accurate deep neural network is a valuable industrial asset due to the know-how for training it, the difficulty of gathering a well-annotated training dataset, and the required computational resources to learn its parameters. In this context, the entity identifying a black-box wants to detect whether it is not a stolen model of her. We name Alice the entity willing to detect the model that Bob has embedded in the black-box.

The biggest difficulty is that there exist plenty of ways to modify a model while maintaining its intrinsic good accuracy. These procedures simplify a network (quantization of the weights and/or activations, pruning, see *e.g.* [7]), or make it more robust (preprocessing of the input, adversarial re-training [8]). We hereafter name a modified model a *variant*. These mechanisms were not a priori designed to make fingerprinting harder but they leave room for Bob to tamper with the fingerprint of a model. Like in biometry, the fingerprint should be discriminative enough to be unique per model but also sufficiently robust to identify a variant.

The approaches in the literature use the decision boundaries in the input space drawn by a classifier as its fingerprint, *i.e.* a signature uniquely identifying the model [3], [5], [6]. Two neural networks sharing the same architecture, the same training set and procedure are different because the training is stochastic (using *i.e.* a Stochastic Gradient Descent). This causes their boundaries in the input space not to overlap fully. Most of the papers in the literature are looking for discriminative deviations of these boundaries.

We thoroughly investigate the use of benign inputs for fingerprinting models contrary to the previous works crafting specific inputs by using advanced techniques for that purpose. We directly identify models using their intrinsic classification behavior. We thus do not need to probe the input space to discover the decision boundaries. Benign inputs constitute a clear advantage, as it removes the need for these often complex and costly crafting procedures. It is as well less prone to defenses being implemented on Bob's side (*e.g.* rejection based on the distance to the decision frontier [9]).

This paper makes the contribution to i) demonstrate that the mere use of benign images is enough to accomplish high success rates for fingerprinting modern classification models. This is to be opposed to the computationally demanding task of crafting inputs for that same goal. ii) We present a distance based on the empirical Mutual Information, gauging how close two models are. This distance permits generalizing the notion of modifications (also coined as *attacks*) on models through the concept of variants. iii) We perform extensive experimentation by considering more than 1,000 classification models on ImageNet. A head-to-head comparison with IP-Guard [3] reports significant improvements.

## II. Related Work

Since the work of IP-Guard [3], all the papers on fingerprinting leverage adversarial examples. They start with a small collection of benign inputs (except [10] starting from random noise images) and apply a white-box attack like CW [11]. It forges adversarial examples that lie close to the decision boundaries, which constitute the signatures of a model.

The followers of IP-Guard [3] forge adversarial examples which are more robust in the sense that they remain adversarial for any variation of the model while being more specific to the vanilla model. Paper [4] proposes to use the universal

[1] Work supported by ANR / AID under Chaire IA SAIDA. The code of FBI is available at https://github.com/t-maho/FBI_fingerprinting. More contributions can be found in the full version of this article [1].

adversarial perturbations of the vanilla model. Paper [12] introduces the concept of conferrable examples, *i.e.* adversarial examples which only transfer to the variations of the targeted model. AFA [6] activates dropout as a cheap surrogate of variants when forging adversarial examples. TAFA [5] extends this idea to other machine learning primitives.

Our take in this article is that using benign images is sufficient, and we addressed the fingerprinting problem without the need to rely on adversarial examples or any other technique to alter images to get them nearby the boundaries. It happens that all above-mentioned papers consider small input dimensions like MNIST or CIFAR ($32 \times 32$ pixel images); none of them use ImageNet ($224 \times 224$) except IP-Guard [3]. Also, no paper considers that the inputs can be reformed by a defense (in order to remove an adversarial perturbation before being classified) or detected as adversarial [13].

## III. Threat Model

### A. Bob: Keeping his Model Anonymous

*1) Goals:* Bob is playing first by secretly selecting a model and putting it in the black-box under scrutiny. This model can be a vanilla model or a variant of a known model. A variant is created by applying on a given vanilla model $\mathsf{m}$ the procedure $\mathsf{V}$ parametrized by $\theta \in \Theta$ which describes the type of modification and the associated parameters. This can be thought of as an attack by Bob on the vanilla model to harden identification. We denote such a variant by $\mathsf{v} = \mathsf{V}(\mathsf{m}, \theta)$.

The goal of Bob is to offer an accurate black-box classifier while maintaining the 'anonymity' of the model at stake. The first requirement is that a small loss in the model performance is tolerated by Bob. If a variant does not comply with this criterion then Bob cannot consider it as an option. In classification, the performance of a model $\mathsf{m}$ is often gauged by the top-1 accuracy, denoted $\mathsf{acc}(\mathsf{m})$. We formalize this requirement as

$$\frac{\mathsf{acc}(\mathsf{m}) - \mathsf{acc}(\mathsf{V}(\mathsf{m}, \theta))}{\mathsf{acc}(\mathsf{m})} < \eta, \tag{1}$$

where $\eta > 0$ is the tolerance (15% in our experimental work).

We also assume that the black-box performs the same classification task. As far as we know, fingerprinting is not possible between two networks performing different tasks if only top-1 output is available. Transfer learning is therefore not considered as in previous works [4], [3], [14].

*2) Resources:* The second requirement is more subtle. We first need to limit the power of Bob. If Bob creates an accurate model *ex nihilo*, then Alice cannot pursue fingerprinting. We assume that Bob cannot train such a model from scratch because he lacks good training data, expertise in machine learning, or computing resources. This also means that Bob can retrain a model only up to a limited extent (typically using a small amount of new data). In other words, the complexity of the procedure creating $\mathsf{v} = \mathsf{V}(\mathsf{m}, \theta)$ ought to be much smaller than the effort spent at training the original model $\mathsf{m}$.

Our experimental work considers two kinds of procedures: 1) modification of the input: $\mathsf{v}(x) = \mathsf{m}(\mathsf{T}(x, \theta))$. Classifiers are robust to benign transformations of the input. As far as images are concerned, the transformation $\mathsf{T}$ can be JPEG compression, posterizing, blurring, *etc*. In the same spirit, *randomized smoothing* [15] consists in adding noise to the input and aggregating the predicted classes into a single output. 2) modification of the model: $\mathsf{v}(x) = \mathsf{T}(\mathsf{m}, \theta)(x)$: The transform $\mathsf{T}$ slightly changes the model weights by for instance quantization, pruning, adversarial retraining or finetuning. Some of these procedures require small retraining with few resources so as not to lose too much accuracy.

In the sequel, the model in the black-box is denoted by $\mathsf{b}$.

### B. Alice: Disclosing the Model in the Black-Box

*1) Goal:* The task of Alice is to detect if a specific model is in the black-box by only having access to its decisions. It means that Alice performs a hypothesis test. She first makes a hypothesis about the black-box, then makes some queries, and finally decides whether the hypothesis holds based on the top-1 outputs of the black-box. The outcome of the detection is thus binary: Alice's hypothesis is deemed correct or not. This is the nominal use case in the related works [2], [3], [4], [5], [6].

*2) Resources:* A crucial point is Alice's knowledge about the black-box. She can only detect a model she knows: it means she has an implementation of this model, which she can freely test.

She also has a collection of typical annotated inputs, *i.e.* a testing dataset. We suppose that these inputs are statistically independent and distributed as the data in the training set of the models. In the sequel, the collection of inputs is denoted $\mathcal{X} = \{x_1, \ldots, x_N\}$ respectively of labels $\{c_1, \ldots, c_N\}$.

In the end, Alice chooses $L$ inputs $(X_1, \ldots, X_L) \subset \mathcal{X}$ to query the black-box and compares the observations $Z = (\delta_{c_1}^{b(X_1)}, \ldots, \delta_{c_L}^{b(X_L)})$ from $b$ to the outputs she knows $Y = ((\delta_{c_1}^{\mathsf{m}(X_1)}, \ldots, \delta_{c_L}^{\mathsf{m}(X_L)})$, where $\delta$ is the Kronecker delta. In others words, $Z$ and $Y$ are binary vectors comparing the decision of $\mathsf{b}$ and $\mathsf{m}$ to the ground truth. We use capital letters here to outline that these are random variables since Alice randomly chooses the inputs.

## IV. Fingerprinting models and variants with FBI

### A. Working Asssumptions

Our working assumption is that when queried by random inputs, a variant $\mathsf{V}(\mathsf{m}, \theta)$ produces outputs statistically:

- independent from the outputs of a different model $\mathsf{m}'$.
- dependent from the outputs of the original model $\mathsf{m}$.

We consider a particular procedure (inspired by information theory) for generating a variant as being like a transmission channel. The output $Z$ of the variant $\mathsf{V}(\mathsf{m}, \theta)$ is as if the output $Y$ of the original model $\mathsf{m}$ were transmitted to Alice through a noisy communication channel parametrized by $\theta$. Like in C.E. Shannon's information theory of communication, we model this channel by the conditioned probabilities $W_\theta(z, y) = \mathbb{P}(Z = z | Y = y), \forall (z, y) \in \{0, 1\}$.

### B. Discriminative Distance

Alice tests two hypothesis:

- $\mathcal{H}_1$: The black-box is a variant of model $\mathsf{m}$. There is a dependence between $Z$ and $Y$ which is captured by the statistical model of the variant:
$$\mathbb{P}_1(Z = z, Y = y) := W_\theta(z, y)\mathbb{P}(Y = y).$$
- $\mathcal{H}_0$: The black-box is not a variant of model $\mathsf{m}$. There is no statistical dependence and
$$\mathbb{P}_0(Z = z, Y = y) := \mathbb{P}(Z = z)\mathbb{P}(Y = y).$$

The well-celebrated Neyman-Pearson test is the optimal score for deciding which hypothesis holds. For $L$ independent observations, it writes as:

$$s = \sum_{j=1}^{L} \log \frac{\mathbb{P}_1(Z = z_j, Y = y_j)}{\mathbb{P}_0(Z = z_j, Y = y_j)} = \sum_{j=1}^{L} \log \frac{W_\theta(z_j, y_j)}{\mathbb{P}(Z = z_j)}. \quad (2)$$

We introduce the empirical joint probability distribution

$$\hat{P}_{Z,Y}(z, y) := L^{-1}|\{j \in [\![L]\!] : z_j = z \text{ and } y_j = y\}| \quad (3)$$

in order to rewrite (2) as:

$$s = L \sum_{(z,y)\in\{0,1\}^2} \hat{P}_{Z,Y}(z, y) \log \frac{W_\theta(z, y)}{\mathbb{P}(Z = z)}. \quad (4)$$

This formalization is not tractable because $W_\theta$ is not known: Alice does not know which variant $\theta$ is in the black-box. Yet, (4) guides us to a more practical score function, the empirical mutual information:

$$\hat{I}(Z, Y) := \sum_{(z,y)\in\{0,1\}^2} \hat{P}_{Z,Y}(z, y) \log \frac{\hat{P}_{Z,Y}(z, y)}{\hat{P}_Z(z)\hat{P}_Y(y)}, \quad (5)$$

with the empirical marginal probabilities:

$$\hat{P}_Z(z) := \sum_{y\in\{0,1\}} \hat{P}_{Z,Y}(z, y), \quad \hat{P}_Y(y) := \sum_{z\in\{0,1\}} \hat{P}_{Z,Y}(z, y). \quad (6)$$

In words, the model of the distributions $(\mathbb{P}_0, \mathbb{P}_1)$ is replaced with empirical frequencies learned on the fly. Resorting to the empirical mutual information to decode transmitted messages in digital communication is known as Maximum Mutual Information (MMI), recently proven universally optimal [16].

The empirical mutual information is a kind of similarity (the bigger, the more $Z$ looks like $Y$). Its value lies in the interval $[0, \min(\hat{H}(Z), \hat{H}(Y))]$ with the empirical entropy given by:

$$\hat{H}(Z) := -\sum_{z} P_Z(z) \log P_Z(z). \quad (7)$$

We prefer dealing with a normalized distance and we introduce:

$$D_L(\mathsf{b}, \mathsf{m}) := 1 - \frac{\hat{I}(Z, Y)}{\min(\hat{H}(Y), \hat{H}(Z))} \in [0, 1]. \quad (8)$$

This defines a pseudo-distance between the models $\mathsf{b}$ and $\mathsf{m}$ respectively producing $Z$ and $Y$. As an illustration, the distances between all the pairs of 1081 models we built from 35 vanilla and open-sourced models are used to create a t-SNE representation in the Figure 1. It shows vanilla models and their variants well-clusterized.

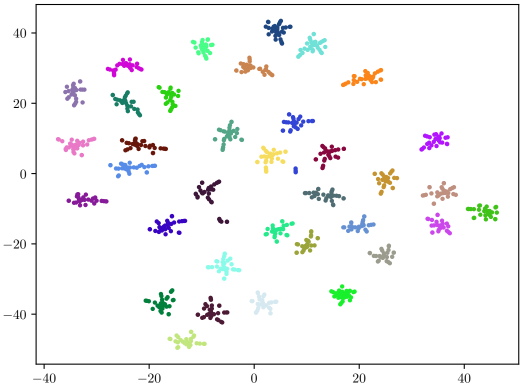


Fig. 1: A t-SNE representation of the pairwise distances of 1081 different models. Each color represent a vanilla model and its variants.

For a given model $\mathsf{m}$, let us consider two extreme scenarios:

- The model $\mathsf{m}$ is in the black-box so that $z_j = y_j$, $\forall j \in [\![L]\!]$. Then $P_{Z,Y}(z, y) = 1$ if $z = y$, and 0 otherwise, producing $D_L(\mathsf{b}, \mathsf{m}) = 0$.
- The black-box and model $\mathsf{m}$ yield independent outputs so that $P_{Z,Y}(z, y) = P_Z(z)P_Y(y)$, then $D_L(\mathsf{b}, \mathsf{m}) = 1$.

In the end, Alice deemed the hypothesis $\mathcal{H}_1$ as being true when the distance is small enough: $D_L(\mathsf{b}, \mathsf{m}) < \tau \rightarrow \mathcal{H}_1$ is true. Alice makes two kinds of errors:

- False positive: $D_L(\mathsf{b}, \mathsf{m}) < \tau$ whereas $\mathcal{H}_1$ is false.
- False negative: $D_L(\mathsf{b}, \mathsf{m}) \geq \tau$ whereas $\mathcal{H}_1$ is true.

Alice sets the threshold $\tau$ such that the probability of false positive is lower than a required level $\alpha$.

### C. Selection of Inputs

The choice of inputs submitted are crucial. If Alice chooses easy inputs, any model outputs the same prediction. This is not discriminative of a given model in the black-box and it may lead to a false positive. On the other hand, these inputs must not be too hard to be classified. Otherwise the prediction tends to be random, destroying the correlation between a model and its variant. This may lead to a false negative.

Our experimental work investigates several selection mechanisms of the inputs. All of them amount to randomly pick inputs from a subset $\mathcal{X}'$ of $\mathcal{X}$.

- All. There is indeed no selection and $\mathcal{X}' = \mathcal{X}$.
- 50/50. Alice's hypothesis concerns a family of variants derived from a vanilla model $\mathsf{m}$. $\mathcal{X}'$ is composed of 50% of inputs well classified by $\mathsf{m}$ (*i.e.* $\mathsf{m}(x) = c(x)$), 50% inputs for which $\mathsf{m}(x) \neq c(x)$.
- 30/70. The same definition but with 30% well classified and 70% wrongly classified by $\mathsf{m}$.
- Entropy. $\mathcal{X}'$ is composed of the inputs whose top-1 predictions are highly random. For a given input, Alice computes the predictions from several models and

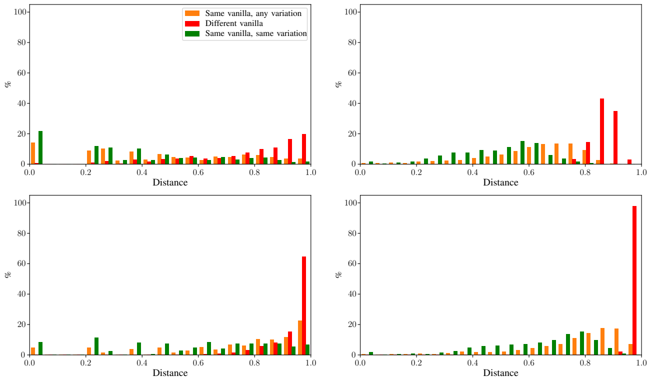
TABLE I: True Positive Rate per variation for $L = 100$ queries.

| Method | Parameter | Finetuning |  | Half | Histogram | JPEG | Posterize | Prune |  |  | Randomized |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | All | Last | Precision |  |  |  | All | Filter | Last | Smoothing |
| IP-Guard [3] | BP [17] & 50 iter. | 0.5 | 92.3 | **100** | 27.3 | **100** | 9.2 | 72.7 | 89.2 | **100** | 26.1 |
| FBI | 30/70 | 94.1 | **97.1** | **100** | 90.3 | 94.7 | 95.4 | 89.5 | **100** | 95.8 | 79.1 |
|  | Entropy | **94.1** | **97.1** | **100** | **100** | **98.8** | **99.4** | **93.7** | **100** | **100** | **85.6** |

measures the empirical entropy of these predicted labels. She then sorts the inputs of $\mathcal{X}$ by their entropy, and $\mathcal{X}'$ contains the head of this ranking.

## V. Results

**Experimental Setup:** We consider 35 vanilla models and 1046 variants. All combinations of hypothesis and model put in the black-box are considered. This represents 1081 positive cases and 36,754 negative cases. The detection performances are gauged by the True Positive Rate (TPR) when threshold $\tau$ is set to get a False Positive Rate (FPR) of 5%.



(a) $L = 20$ Images

(b) $L = 1000$ Images

Fig. 2: Histogram of the distance of each pair of models. Inputs randomly sampled in $\mathcal{X}$ (*top*) or in $\mathcal{X}'$ -Entropy (*bottom*).

**Assumptions about the Statistical Model:** Section IV-A makes two assumptions about the statistical dependence between the predictions of models from the same vanilla model and independence when coming from different vanilla model. Figure 2 experimentally verifies these working assumptions.

The distances between each pair of models are computed. This sums up to 583,740 combinations. Figure 2 shows the histogram of these distance values over 20 bins in red. A high number $L$ of queries makes the measured distance more precise. The selection of the inputs has a major impact. When sampled on $\mathcal{X}$ (first row), the distance rarely values the maximum showing imperfect independance. This phenomenon has been revealed in [18]. Yet, when sampled on $\mathcal{X}'$ containing more inputs hardly correctly classified (second row), the distances are closer to one. The models tend to be independent when queried with a good selection of inputs.

**Selection of Inputs:** Table II shows the TPR obtained when different numbers of queries are used. The selection Entropy is clearly the best option but it needs statistics about the predictions of many vanilla models. As far as only a single model is available, the other selections are to be preferred. They only require the predictions of the suspected vanilla model.

TABLE II: True Positive Rate for 100 queries sampled in $\mathcal{X}'$.

| Selection | All | 50/50 | 30/70 | Entropy |
|---|---|---|---|---|
| TPR | $79.4 \pm 2.1$ | $89.2 \pm 1.3$ | $91.1 \pm 1.5$ | $\mathbf{95.2 \pm 0.5}$ |

TABLE III: True Positive Rate for different number of queries.

| Method | Parameter | $L = 100$ | $L = 200$ |
|---|---|---|---|
| IP-Guard [3] | BP [19] & 50 iter. | 66.9 | 72.7 |
| FBI | 30/70 | 91.1 | 97.4 |
|  | Entropy | **95.2** | **97.6** |

**Benchmark with the State-of-the-Art:** IP-Guard [3] is the only work demonstrated to be tractable and effective on large input size like in ImageNet. It leverages several white-box attacks to create adversarial examples. The best results demonstrated in the paper are with the attack CW [11]. We instead use BP [19]. It exhibits similar performances while being much faster (only 50 iterations). The BP implementation is from GitHub.[2].

Table III compares the performances under 100 and 200 queries and top-1 observations. Any selection of the inputs beats IP-Guard [3]. Detailed results are reported in Table I. Some variations are easier to detect (‘precison’, ‘pruning’) and the two methods are on par. On the contrary, randomized smoothing which is a popular variation yet never considered in the literature, is more efficient against both fingerprinting approaches: IP-Guard [3] relies on crafting adversarial examples close to the decision boundaries which are greatly crumpled by randomized smoothing. Not relying on adversarial examples seems to be a clear advantage in this case. Our method offers more stability in the results: no variation pulls the TPR below 85%.

## VI. Conclusion

The problem of accurate and efficient fingerprinting of valuable models is salient. This paper demonstrates that such a demand can be fulfilled by solely using benign inputs. Hundreds of inputs are necessary to achieve high results. This has the important implication that we no longer need models in white-box access to compute their fingerprints.

Bob’s best defense in our experimental protocol against fingerprinting is randomized smoothing. It means that the former reduces the statistical dependence of the outputs, while the latter hardly perturbs the outputs given by the vanilla model.

[2] Boundary Projection’s GitHub: https://github.com/hanwei0912/walking-on-the-edge-fast-low-distortion-adversarial-examples

## REFERENCES

[1] T. Maho, T. Furon, and E. L. Merrer, "Fbi: Fingerprinting models with benign inputs," 2022. [Online]. Available: https://arxiv.org/abs/2208.03169

[2] S. Wang and C.-H. Chang, "Fingerprinting deep neural networks - a deepfool approach," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5.

[3] X. Cao, J. Jia, and N. Z. Gong, "IPGuard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary," in *Proc. of the 2021 ACM Asia Conference on Computer and Communications Security*, 2021.

[4] Z. Peng, S. Li, G. Chen, C. Zhang, H. Zhu, and M. Xue, "Fingerprinting deep neural networks globally via universal adversarial perturbations," *arXiv preprint arXiv:2202.08602*, 2022.

[5] X. Pan, M. Zhang, Y. Lu, and M. Yang, "Tafa: A task-agnostic fingerprinting algorithm for neural networks," in *European Symposium on Research in Computer Security*. Springer, 2021, pp. 542–562.

[6] J. Zhao, Q. Hu, G. Liu, X. Ma, F. Chen, and M. M. Hassan, "AFA: Adversarial fingerprinting authentication for deep neural networks," *Computer Communications*, vol. 150, pp. 488–497, 2020.

[7] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[8] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[9] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 135–147.

[10] S. Wang, P. Zhao, X. Wang, S. Chin, T. Wahl, Y. Fei, Q. A. Chen, and X. Lin, "Intrinsic examples: Robust fingerprinting of deep neural networks," in *British Machine Vision Conference (BMVC)*, 2021.

[11] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[12] N. Lukas, Y. Zhang, and F. Kerschbaum, "Deep neural network fingerprinting by conferrable adversarial examples," in *International Conference on Learning Representations*, 2021.

[13] A. Kherchouche, S. A. Fezza, W. Hamidouche, and O. Déforges, "Detection of adversarial examples in deep neural networks with natural scene statistics," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

[14] S. Wang, X. Wang, P.-Y. Chen, P. Zhao, and X. Lin, "Characteristic examples: High-robustness, low-transferability fingerprinting of neural networks," in *Proc. of the Thirtieth Int. Joint Conference on Artificial Intelligence, IJCAI-21*, 2021.

[15] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *ICML*, 2019.

[16] R. Tamir and N. Merhav, "The MMI decoder is asymptotically optimal for the typical random code and for the expurgated code," 2020. [Online]. Available: https://arxiv.org/abs/2007.12225

[17] H. Zhang, Y. Avrithis, T. Furon, and L. Amsaleg, "Walking on the Edge: Fast, Low-Distortion Adversarial Examples," *IEEE TIFS*, 2020. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02931493

[18] H. Mania, J. Miller, L. Schmidt, M. Hardt, and B. Recht, "Model similarity mitigates test set overuse," in *Advances in Neural Information Processing Systems*, 2019.

[19] B. Bonnet, T. Furon, and P. Bas, "Generating Adversarial Images in Quantized Domains," *IEEE Transactions on Information Forensics and Security*, 2022.