Skolkovo Institute of Science and Technology

As a manuscript

Mikhail Pautov

Certified Robustness of Neural Networks

Ph.D. Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Academic supervisor:
Doctor of Sciences in Physics and Mathematics, Professor
Ivan V. Oseledets

Moscow — 2024

Introduction

Topic relevance. Neural networks-based algorithms have recently achieved tremendous success in practical problems from computer vision domain. In the tasks of image classification [27], segmentation [28], object detection [29] and localization [30], the precision of such algorithms is often comparable to the ones of a human, so, due to remarkable performance, they are deployed in self-driving cars [31], medical imaging [32], security [33] and many other applications. Unfortunately, such an advantage comes at the cost of computationally expensive training, instability and vulnerability to different kinds of perturbations of the input data. In the primary work [34], it was discovered that very small vicinity of a correctly classified image is full of (adversarial) samples, which are, although indistinguishable from the original image, are classified differently by a network. This finding gave a rise to an important question: how one can verify that a neural network's prediction is correct when its input is subjected to a transformation that does not change its semantic (for example, adding a small noise, or rotating an image by few degrees)?

Since then, in the deep learning literature, a lot of methods to exploit such a vulnerability were proposed, underlining the importance of robust algorithms [35—37; A1]. As a consequence, many approaches to make neural networks empirically more robust appeared [38—40]. Such methods, although make it harder to find a perturbation of the input that breaks a network, do not provide any guarantees that, under certain assumptions, such perturbations do not exist. To fill this gap, a new field of theoretical deep learning, called *certified robustness*, appeared. The purpose of certified robustness is to provide guarantees that a certain neural network is provably resistant to a particular type of input transformations, such as additive perturbations [41].

In safety-crucial applications of neural networks, such as self-driving cars and medical diagnostics, the trustworthiness of the predictions of the algorithms is equally important as the precision of one in the original task. For example, if a computer vision algorithm detects a pedestrian both under the daylight and in the night, is it not enough to entrust it driving a vehicle: the developer has to provide *guarantees* that any pedestrian would be detected equally precisely in different lighting and weather conditions.

From the practical point of view, it seems that providing such guarantees is an impossible task: one has to consider any transformation of the input data of a certain type, which is often uncountable. For example, consider the problem of certification of a classification network to rotation transform: it is clear that the number of different degrees, so as transformations, is uncountable. However, it is possible to provide ones leveraging the mathematical properties of the neural network as a function of multiple variables.

In the field of certified robustness, the guarantees on the correctness of the behavior of neural networks often come at a high cost: either the correctness is proven in very narrow and unrealistic scenario, or the ones lead to a significant performance degradation of the certified network. Due to this issues, it seems relevant to develop the methods for certification against broad class of input perturbations which do not affect the performance of the models drastically. The development and integration of such methods lead to the improvement of the trustworthiness of neural networks, and, as a consequence, to the growth of ones' practical applications.

This work is devoted to the problems of robustness and privacy of classification neural networks.

The goal of this work is the development of approaches that guarantee the robustness and privacy of neural networks without a noticeable decrease in the performance of the latter.

In order to achieve the set goal, the following **tasks** were set:

1. To demonstrate the vulnerability of neural network-based algorithms in the most common practical applications.
2. To develop a method for probabilistic certification of neural networks with respect to perturbations of input data of arbitrary type.
3. To develop a method for certification of prototypical neural networks with respect to additive perturbations of bounded norm.
4. To develop a watermark-based method as and indicator of the theft of a neural network deployed in "black box" settings.

The novelty:

1. The gradient-based adversarial patch generation method for real-time vulnerability assessment of face recognition models in the physical domain was developed.
2. The method of probabilistic certification of neural networks to arbitrary perturbations of input data was developed.
3. The method for certification of prototypical neural networks to additive perturbations of input data of bounded norm in the few-shot learning setting was developed.
4. The watermark-based method for theft detection of a neural network deployed as a "black box" service was developed.

Practical significance of the work lies in the creation of the following methods:

1. The approach to adversarial patch generation that demonstrates the vulnerability of neural networks to additive perturbations in the physical domain in a real-time face recognition task.
2. CC-Cert, a method for probabilistic certification of neural networks to input data perturbations of arbitrary type.

3. Smoothed Embeddings, a method for certification of prototypical neural networks to additive perturbations of input data of bounded norm in the few-shot learning scenario.
4. The watermark-based method for theft detection of a neural network deployed as a “black box” service.

Main provisions to be defended:

1. The approach to adversarial patch generation that demonstrates the vulnerability of neural networks to additive perturbations in the physical domain in a real-time face recognition task.
2. CC-Cert, a method for probabilistic certification of neural networks to input data perturbations of arbitrary type.
3. Smoothed Embeddings, a method for certification of prototypical neural networks to additive perturbations of input data of bounded norm in the few-shot learning scenario.
4. The watermark-based method for theft detection of a neural network deployed as a “black box” service.

Probation. The results of the work were reported at the following conferences:

1. International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), October, 2019. Topic: “On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System”.
2. The Thirty-Sixth AAAI Conference on Artificial Intelligence, February, 2022. Topic: “CC-Cert: A probabilistic approach to certify general robustness of neural networks“.
3. The Thirty-sixth Annual Conference on Neural Information Processing Systems, December, 2022. Topic: “Smoothed Embeddings for Certified Few-Shot Learning“.
4. Conference Fall into ML, November, 2022. Topic: “Smoothed Embeddings for Certified Few-Shot Learning“.
5. ISP RAS Open Conference, December, 2022. Topic: “Smoothed Embeddings for Certified Few-Shot Learning“.
6. AIRI Seminar AIschnitsa, December, 2022. Topic: “Smoothed Embeddings for Certified Few-Shot Learning“.

Personal contribution. The author’s contributions to the research described in this thesis are as follows:

1. In the paper “On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System,” the author implemented the numerical procedure to generate adversarial sticker that breaks the face recognition system in the physical domain. The author conducted the experiments to evaluate the transferability of a sticker from the digital domain to the physical domain as well as in the interpretation of

the results. Together with the co-authors, the author prepared the text of the paper.

2. In the paper "CC-Cert: A probabilistic approach to certify general robustness of neural networks", the author proposed an approach to certify the neural networks to compositions of transformations in the probabilistic setting. The author formulated and proved all the theoretical results, namely, the probabilistic guarantees on the correctness of the prediction of a neural network in the presence of a certain transformations of the input. The author also designed the methodology to evaluate the proposed method, conducted the experiments and interpreted their results. Together with the co-authors, the author prepared the text of the paper.
3. In the paper "Smoothed Embeddings for Certified Few-Shot Learning", the author proposed an approach to certify the prototypical networks against additive transformations of a bounded magnitude. The author formulated and proved the theoretical result, namely, the deterministic guarantee on the correctness of the prediction of a neural network in the presence of additive transformations of the input. The author also developed a methodology to evaluate the proposed approach and participated in conducting of the experiments. The author also prepared the text of the paper and all its revisions.
4. In the preprint "Probabilistically Robust Watermarking of Neural Networks", the author proposed the methodology to generate robust digital watermarks to protect the neural networks from the theft. The author developed the experimental setup to evaluate the proposed approach and, together with the co-authors, prepared the text of the paper.

Publications.

The research results are presented in the following works:

1. "CC-Cert: A probabilistic approach to certify general robustness of neural networks" by **Mikhail Pautov**, Nurislam Tursynbek, Marina Munkhoeva, Nikita Muravev, Aleksandr Petiushko, and Ivan Oseledets [A2]. The paper is published in Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 7, pp. 7975-7983, 2022; CORE A*).
2. "Smoothed Embeddings for Certified Few-Shot Learning" by **Mikhail Pautov**, Olesya Kuznetsova, Nurislam Tursynbek, Aleksandr Petiushko, and Ivan Oseledets [A3]. The paper is published in Advances in Neural Information Processing Systems (Vol. 35, pp. 24367-24379, 2022; CORE A*).
3. "On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System" by **Mikhail Pautov**, Grigorii Melnikov, Edgar Kaziakhmedov, Klim Kireev and Aleksandr Petiushko [A4]. The paper

is published in Proceedings of 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) (pp. 0391-0396, 2019; Indexed in Scopus).

4. “Probabilistically Robust Watermarking of Neural Networks” by **Mikhail Pautov**, Nikita Bogdanov, Stanislav Pyatkin, Oleg Rogov and Ivan Oseledets (under review).

Other publications:

1. “Real-World Attack on MTCNN Face Detection System" by Edgar Kaziakhmedov, Klim Kireev, Grigorii Melnikov, **Mikhail Pautov** and Aleksandr Petiushko [A1]. The paper is published in Proceedings of 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) (pp. 0422-0427, 2019; Indexed in Scopus).
2. “Translate Your Gibberish: Black-Box Adversarial Attack on Machine Translation Systems” by Andrei Chertkov, Olga Tsymboi, **Mikhail Pautov** and Ivan Oseledets. The paper is published in Journal of Mathematical Sciences (vol. 530, pp. 96-112, 2023).

Content of the work

Introduction substantiates the relevance of the research conducted within the framework of this dissertation, formulates the goal, sets the objectives of the work, outlines the scientific novelty and practical significance of the presented work.

The first chapter is an introductory one and is devoted to the demonstration of the vulnerability of neural networks to additive perturbations of input data (adversarial attacks) in the physical domain. In sections 1.2-1.3, the formal problem statement is presented, and the description of several methods to generate adversarial perturbations is provided. The adversarial attack is a perturbation of an input object that leads to an incorrect prediction of the neural network for the perturbed object. Formally, if the neural network $f : \mathbb{R}^n \to \Delta^k$ that maps input objects to probability vectors of K classes, the norm threshold ε and input object x are given, then additive adversarial attack is the perturbation δ such that

$$\begin{cases} \arg\max_{i \in [1,\dots,K]} f_i(x+\delta) \neq \arg\max_{i \in [1,\dots,K]} f_i(x), \\ \|\delta\| \leq \varepsilon. \end{cases}$$

Sections 1.4-1.5 describe the proposed gradient-based method to compute adversarial perturbations for the neural network in the setting of face recognition. The proposed approach is based on an iterative method for solving the optimization problem of finding an additive δ for a neural network f at the point x of class y in the digital domain:

$$\begin{cases} \gamma^{t+1} = \mu_1 \gamma^t + \mu_2 \frac{\nabla_x L(f,x^t,y)}{\|\nabla_x L(f,x^t,y)\|}, \\ \delta^{t+1} = \varepsilon \mathrm{sign}(\gamma^{t+1}), \\ x^{t+1} = \mathrm{clip}_{[0,1]^n} \left[x^t + \delta^{t+1}\right], \\ x^1 = x, \quad \gamma^1 = 0, \quad \delta^0 = 0, \quad \delta = \delta^T. \end{cases}$$

Here $L(f, x^t, y)$ is the loss function of the neural network, $\varepsilon, \mu_1, \mu_2$ are constants.

This work investigates the transferability of adversarial attributes created in the digital domain to the physical domain. To ensure the transferability of adversarial attributes, the loss function of the neural network $L(f, x^t, y)$ consists of the terms responsible for preserving the smoothness of the generated perturbation in the form of TV loss [42]

$$TV(x^t) = \sum_{i,j} \sqrt{\left(x^t_{i,j} - x^t_{i,j+1}\right)^2 + \left(x^t_{i,j} - x^t_{i+1,j}\right)^2}$$

and augmentation of the adversarial attribute with transformations from some parametric set $\mathcal{T}$ in the form Expectation Over Transformation [43]

$$L_{\mathrm{adv}}(f,x^t,y) = \mathbb{E}_{\tau\in\mathcal{T}} \ \left(\cos\langle f(\tau(x^t)), c_y\rangle\right),$$

where c_y is a prototype vector of class y.

The proposed method to generate adversarial attributes is evaluated in the task of targeted adversarial attack. The goal of this attack is to construct a perturbation, the application of which leads to a controlled change in the prediction of the neural network. In this work, the effect of the location of the adversarial attribute on the performance of the adversarial attack is investigated. Section 1.6 describes the details of the experiments. As a demonstration of the effectiveness of the proposed method, the results of the face recognition system after applying adversarial attributes are given (see Table 1).

The examples of adversarial attributes are shown in Fig. 1, 2.

The final part of the first chapter presents a discussion of the experimental results and the importance of further research on the stability of neural networks, aimed at creating methods to protect the latter from adversarial attacks.

The second chapter is devoted to the study of methods for providing guarantees of correctness of behavior of classification neural networks in the presence of input data perturbations of arbitrary type. This chapter presents a description of the proposed method of providing probabilistic guarantees of correctness of neural network behavior. The proposed approach is used to estimate the probability of occurrence of a wrong prediction of a neural network in the case when the initial correctly classified object x is subjected to the transformation $T_\theta(x)$. The transform is defined by parameters θ chosen randomly from some set of parameters Θ..

Table 1 — Numerical results of experiments, LResNet100E-Ir face recognition system. As the metric, we use the cosine similarity between the embeddings of the photo with the adversarial attribute applied and the prototype vectors of the original ground truth class (e) and the desired class ($e_{x'}$). The objects from set x_{train} correspond to photos used in the adversarial attribute creation process; objects from set x_{val} correspond to hold-out photos used to test the adversarial attack in the digital domain; objects from set x_{test} correspond to photos created in the physical domain with the adversarial attribute applied.

Type of adversarial patch	Eyeglasses	Sticker on a forehead
$cos(e_{x_{train}}, e)$	0.041 ± 0.052	-0.053 ± 0.009
$cos(e_{x_{train}}, e_{x'})$	0.648 ± 0.020	0.221 ± 0.011
$cos(e_{x_{val}}, e)$	0.317 ± 0.004	0.273 ± 0.007
$cos(e_{x_{val}}, e_{x'})$	0.451 ± 0.021	0.421 ± 0.025
$cos(e_{x_{test}}, e)$	0.305 ± 0.024	0.323 ± 0.035
$cos(e_{x_{test}}, e_{x'})$	0.363 ± 0.024	0.391 ± 0.021

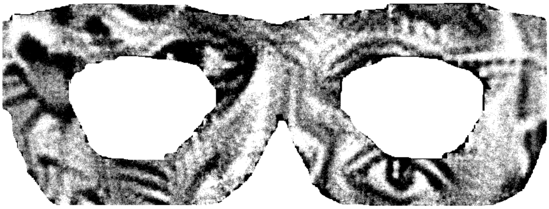


Figure 1 — An adversarial attribute in the form of eyeglasses.

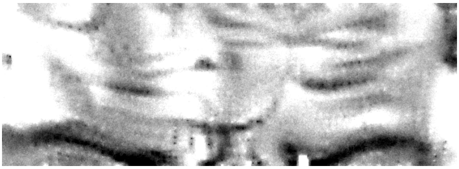


Figure 2 — An adversarial attribute in the form of a forehead sticker.

In section 2.3 of this work, it is shown that if the difference between the output probability vectors $f(x)$ and $f(T_\theta(x))$ does not exceed the value $d = \frac{p_1 - p_2}{2}$, where p_1, p_2 are the two maximum components of the vector $f(x)$, then the objects x and $T_\theta(x)$ are classified by the neural network f as representatives of the same class. Based on this observation and the Chernoff-Cramer method of estimating the probability of large deviations of a random variable, we propose an approach for estimating the probability of incorrect prediction of a neural network. Formally, given $t > 0$, the probability of a large deviation of the random variable $Z = \|f(x) - f(T_\theta(x)\|_\infty$ is bounded:

$$\mathbb{P}_{\theta\sim\Theta}\left(Z > d\right) \leq e^{-dt}\mathbb{E}(e^{Zt}).$$

Since it is impossible to directly compute the expectation $\mathbb{E}(e^{Zt})$, it is estimated as the function

$$\hat{b} = \frac{1}{\delta}\max\left\{Y_1, \ldots, Y_k\right\},$$

where $\delta \in (0{,}1)$ is a constant, $Y_1, \ldots, Y_k$ are k sample means over n samples:

$$Y_j = \exp(-dt)\frac{1}{n}\sum_{i=1}^{n}\exp(Z_i^j t), \qquad Z_i^j = \|f(x) - f(T_{\theta_i^j}(x)\|_\infty, \qquad \theta_i^j \sim \Theta.$$

Also, Section 2.3 of the work presents theoretical guarantees of the applicability of the proposed approach. The probability of correctness of using a function of sample averages instead of the mathematical expectation is estimated according to the theorem below:

Theorem 1. *Let the random variable Z take values from $[0{,}1]$, and let the probability density function of random variable $\xi = e^{Zt}$ have the coefficient of variation $C_v = \frac{\sigma_\xi}{\mathbb{E}(\xi)}$. Then*

$$\mathbb{P}\left(\hat{b} < \frac{\mathbb{E}(\xi)}{e^{dt}}\right) < \left(\frac{1}{1 + \frac{n(1-\delta)^2}{C_v^2}}\right)^k. \tag{1}$$

Section 2.4 of the chapter presents a description of the experiments to verify the effectiveness of the proposed method.

The proposed method is used to estimate the probability of misclassification of a neural network on public datasets, namely MNIST [44] and CIFAR-10 [45].

As a result of the experiments, we provide probabilistically certified accuracy, *PCA*, in dependence on probability threshold ε and show how it is connected to empirical robust accuracy, *ERA*, under corresponding adaptive attack. Namely, given the classifier $h(\cdot)$, set of images $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ and threshold ε, probabilistically certified accuracy is computed as

$$PCA(\mathcal{S}, \varepsilon) = \frac{|(x, y) \in \mathcal{S} : \texttt{BOUND}(x) < \varepsilon \;\&\; h(x) = y|}{m}.$$

At the same time, given the discretization $\Theta = \{\theta_1, \ldots, \theta_r\}$ of space of parameters of the transform T, empirical robust accuracy is computed as a fraction of objects from $\mathcal{S}$ that are correctly classified under all the transformations $T_{\theta_i}, i \in [1, \ldots, r]$:

$$ERA(\mathcal{S}) = \frac{|(x, y) \in \mathcal{S} : h(T_{\theta_i}(x)) = y \ \forall i \in [1, \ldots r]|}{m}.$$

Also, in Section 2.4, we describe the parameters of the experiments and transformations of the input data. The results of the experiments are presented in Table 2. In Table 3, we present the parameters of the transformations considered in the experiments. It should be noted that some parameters depend on the characteristics of the original images (e.g., the size).

Table 2 — Comparison of probabilistically certified accuracy and empirical robust accuracy. In the Transform column, the following notation is used: B stays for Brightness, C stays for contrast, R stays for Rotation, G stays for Gaussian blur, T stays for Translation, S stays for Scale. In the Training column, P stays for plain training of the model, A stays for the training with augmentations. We report probabilistically certified accuracy for three levels of threshold parameter ε: high confidence in certification ($\varepsilon < 10^{-10}$), middle level of confidence ($\varepsilon < 10^{-7}$) and low level of confidence ($\varepsilon < 10^{-4}$). In the column PA, we report the initial accuracy on the test subsets of the datasets.

Dataset	**Transform**	**Training**	**ERA**		**PCA(ε)**		**PA**
				$\varepsilon = 10^{-10}$	$\varepsilon = 10^{-7}$	$\varepsilon = 10^{-4}$	
CIFAR-10	B	P	58.4%	47.8%	51.6%	55.2%	91.18%
		A	65.0%	55.4%	59.4%	61.8%	88.67%
	C	P	91.6%	62.4%	67.0%	69.6%	91.18%
		A	88.0%	67.0%	72.8%	74.2%	88.67%
	R	P	73.4%	64.6%	69.0%	71.0%	91.18%
		A	72.4%	57.4%	63.6%	67.4%	87.77%
	G	P	12.2%	11.0%	11.0%	11.0%	91.18%
		A	60.4%	57.2%	57.2%	57.8%	81.11%
	T	P	40.4%	28.0%	31.2%	35.2%	91.18%
		A	35.0%	17.8%	22.4%	25.6%	85.98%
	S	P	57.0%	54.4%	54.4%	54.4%	91.18%
		A	55.0%	53.4%	53.4%	53.6%	86.76%
	C + B	P	0.0%	0.0%	0.0%	0.0%	91.18%
		A	0.4%	0.0%	0.0%	0.0%	88.67%
	R + B	P	22.6%	16.2%	20.6%	21.8%	91.18%
		A	30.4%	21.2%	24.6%	27.6%	84.50%
	S + B	P	10.2%	10.4%	10.4%	10.4%	91.18%
		A	41.8%	40.6%	40.6%	40.6%	86.53%
MNIST	B	P	97.8%	94.8%	96.4%	97.0%	99.26%
		A	98.6%	97.0%	98.2%	98.2%	99.04%
	C	P	98.8%	96.0%	97.0%	97.2%	99.26%
		A	98.6%	98.2%	98.2%	98.2%	99.04%
	R	P	18.8%	11.6%	14.8%	16.4%	99.26%
		A	98.0%	97.0%	97.4%	97.6%	99.01%
	G	P	78.0%	68.8%	68.8%	68.8%	99.26%
		A	97.8%	97.8%	97.8%	97.8%	98.35%
	T	P	0.0%	0.0%	0.0%	0.0%	99.26%
		A	39.6%	31.4%	34.4%	38.2%	99.09%
	S	P	21.6%	21.0%	21.0%	21.0%	99.26%
		A	34.4%	34.4%	34.4%	34.4%	99.25%
	C + B	P	8.4%	0.0%	0.0%	0.0%	99.26%
		A	7.6%	2.4%	2.4%	2.4%	99.04%
	R + B	P	14.0%	9.2%	11.2%	13.0%	99.26%
		A	95.2%	93.0%	93.4%	94.6%	99.08%
	S + B	P	13.0%	13.4%	13.4%	13.4%	99.26%
		A	93.4%	93.0%	93.0%	93.4%	99.37%

Table 3 — Parameters of transformations.

Dataset	Transform	Parameters
CIFAR-10	Brightness	$\theta_b \in [-40\%, 40\%]$
	Contrast	$\theta_c \in [-40\%, 40\%]$
	Rotation	$\theta_r \in [-10^\circ, 10^\circ]$
	Gaussian blur	$\theta_g \in [0, 3]$ – kernel radius
	Translation	$\|\theta_t\| \leq 20\%$
	Scale	$\theta_s \in [70\%, 130\%]$
	Contrast + Brightness	see Contrast & Brightness
	Rotation + Brightness	see Rotation & Brightness
	Scale + Brightness	see Scale & Brightness
MNIST	Brightness	$\theta_b \in [-50\%, 50\%]$
	Contrast	$\theta_c \in [-50\%, 50\%]$
	Rotation	$\theta_r \in [-50^\circ, 50^\circ]$
	Gaussian blur	$\theta_g \in [0, 3]$ – kernel radius
	Translation	$\|\theta_t\| \leq 30\%$
	Scale	$\theta_s \in [70\%, 130\%]$
	Contrast + Brightness	see Contrast & Brightness
	Rotation + Brightness	see Rotation & Brightness
	Scale + Brightness	see Scale & Brightness

The final part of the chapter discusses the applicability of the proposed approach to neural network certification and identifies one promising direction for further research: the analysis of deterministic robustness guarantees.

The third chapter is devoted to the study of the robustness of prototypical neural networks to additive perturbations of the bounded norm. In the introductory section 3.1, we present a general problem statement and the motivation for creating provably robust systems in the few-shot learning setting.

In Section 3.2, the formal problem statement is presented. Suppose that the neural network

$$f : \mathbb{R}^D \to \mathbb{R}^d,$$

that maps input objects to the space of normalized embeddings is given. Then, $d-$dimensional prototypes of classes are computed as follows (expression is given for the prototype of class k):

$$c_k = \frac{1}{S_k} \sum_{x \in S_k} f(x).$$

Here S_k is the set of objects of class k. Suppose that $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, is the dataset, where $x_i \in \mathbb{R}^D$ – is an input object of class $y_i \in \{1, \ldots, K\}$. The goal of the work is to construct a classifier g provably robust to additive perturbations δ of a small norm. In other words, we want to have a classifier, such that, given a norm threshold t, the equality

$$\operatorname*{arg\,min}_{k \in \{1,\ldots,K\}} \rho\left(g(x), c_k\right) = \operatorname*{arg\,min}_{k \in \{1,\ldots,K\}} \rho\left(g(x+\delta), c_k\right), \tag{2}$$

will be satisfied for all $\delta : \|\delta\|_2 \leq t$.

Section 3.3 presents a description of an approach to creating classifiers provably robust to additive perturbations of input data of a bounded norm. The approach is based on the idea of the randomized smoothing [41]. In a nutshell, randomized smoothing is done by replacing the original prototypical neural network $f : \mathbb{R}^D \to \mathbb{R}^d$ with a surrogate neural network defined as

$$g(x) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} f(x + \varepsilon).$$

For a smoothed prototypical neural network in the form $g(x)$, we formulate and prove the Lipschitz property in the form of a theorem in Section 3.3.

Theorem 2. *(Lipschitz property) Suppose that $f : \mathbb{R}^D \to \mathbb{R}^d$ is a deterministic function and $g(x) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0,\sigma^2 I)} f(x+\varepsilon)$ is continuously differentiable for all x. If for all x, $\|f(x)\|_2 = 1$, then $g(x)$ is $L-$Lipschitz in l_2-norm with $L = \sqrt{\frac{2}{\pi\sigma^2}}$.*

To ensure provably robust classification in the embedding space, we estimate the maximum embedding deviation of the classified object that does not change the nearest prototype of the class. The expression for this distance is formulated in the form of a theorem.

Theorem 3. *(Adversarial embedding risk) Given an input image $x \in \mathbb{R}^D$ and the embedding $g : \mathbb{R}^D \to \mathbb{R}^d$ the closest point on to decision boundary in the embedding space (see Figure 2) is located at a distance (defined as adversarial embedding risk):*

$$\gamma = \|\Delta\|_2 = \frac{\|c_2 - g(x)\|_2^2 - \|c_1 - g(x)\|_2^2}{2\|c_2 - c_1\|_2^2}, \tag{3}$$

where $c_1 \in \mathbb{R}^d$ and $c_2 \in \mathbb{R}^d$ are the two closest prototypes. The value of γ is the distance between classifying embedding and the decision boundary between classes represented by c_1 and c_2. Note that this is the minimum l_2-distortion in the embedding space required to change the prediction of g.

Taking into account the Lipshitz property and the value of adversarial embedding risk Δ, we formulate a guarantee of the robustness of the prototypical neural network to additive perturbations:

Theorem 4. *(Robustness guarantee) L_2-robustness guarantee r for an input image x in the $n-$dimensional input metric space under classification by a classifier g is $r = \frac{\gamma}{L}$,*

where L is the Lipschitz constant from the Theorem 2 and γ is the adversarial risk from the Theorem 3. The value of r is the certified radius of g at x, or, in other words, minimum l_2-distortion in the input space required to change the prediction of g.

Section 3.4 provides a description of the certification procedure. Section 3.5 describes the experiments and summarizes their results.

For the experimental evaluation of our approach we use several well-known datasets for few-shot learning classification. *Cub-200-2011* [46] is a dataset with 11,788 images of 200 bird species, where 5864 images of 100 species are in the train subset and 5924 images of other 100 species are in the test subset. *mini*ImageNet [47] is a substet of images from *ILSVRC 2015* [48] dataset with 64 images categories in train subset, 16 categories in validation subset and 20 categories in test subset with 600 images of size 84 × 84 in each category. *CIFAR FS* [49] is a subset of *CIFAR 100* [45] dataset which was generated in the same way as *mini*ImageNet and contains 37800 images of 64 categories in the train set and 11400 images of 20 categories in the test set.

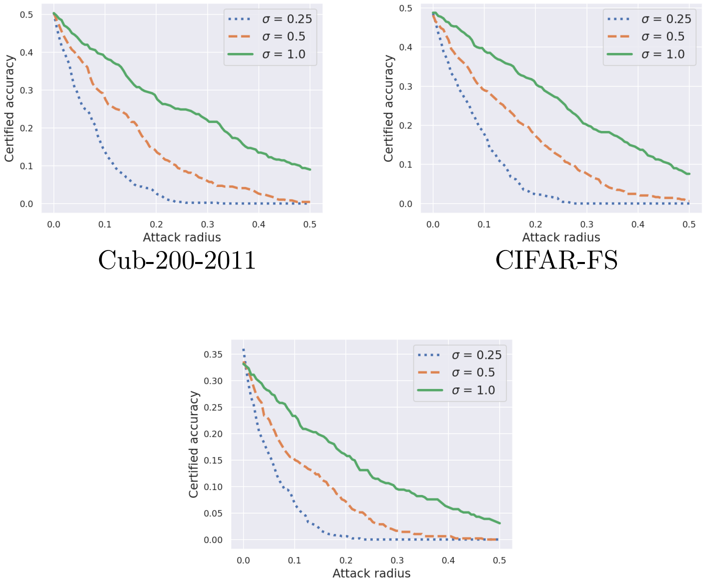
In our evaluation protocol, we compute approximate certified robust accuracy on the test set, CRA. Given a sample x, a smoothed classifier $g(\cdot)$ from the Theorem 2 with an associated classification rule $h(x) = \arg\min_{i\in\{1,\dots,K\}} \|g(x) - c_k\|_2$, threshold value ε for l_2-norm of additive perturbation and the robustness guarantee $r = r(x)$ from the Theorem 4, we compute CRA on test set S as follows:

$$CRA(S, \varepsilon) = \frac{|(x,y) \in S : r(x) > \varepsilon \ \& \ h(x) = y|}{|S|}. \tag{4}$$

The figures 3-4 present the dependencies of certified accuracy on the value of norm of additive perturbation for different learning settings (1-shot and 5-shot learning). The value of the attack radius corresponds to the threshold ε from (4).

Section 3.6 provides a theoretical assessment of the limits of applicability of the proposed method. In section 3.7, the discussion of the experimental results is presented, and the directions for future research are outlined.

The fourth chapter is devoted to the creation of digital watermarks as indicators of functionality-stealing attacks.



Cub-200-2011

CIFAR-FS

*mini*ImageNet

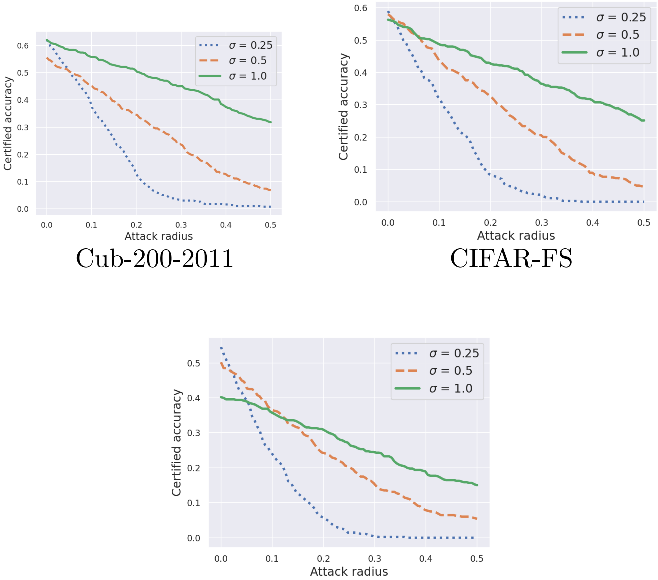
Figure 3 — Dependency of certified accuracy on attack radius ε for different σ, 1-shot case.

Section 4.1 provides motivation for developing watermarking techniques that are resistant to functionality-stealing attacks. The section highlights the need to protect the privacy of neural networks deployed in the “black box” setting: a potential adversary can use the distillation of the neural network, thus obtaining a functionally similar model, without spending time and resources to develop and train the neural network. Sections 4.2-4.3 review the current literature and provide necessary background information about functionality-stealing attacks and digital watermarking techniques.

Section 4.4 describes the proposed approach to generate trigger set-based watermarks robust to functionality-stealing attacks.

By a trigger set, we mean a set $\mathcal{D}_t^* \subset \mathbb{R}^n$ such that the predictions of the original neural network $f : \mathbb{R}^n \to \Delta^k$ on objects from it are predetermined. The procedure of creating a trigger set proposed in this work consists of two parts:

- In the first part, a set of candidate points for inclusion in the trigger set is collected. Let the original neural network f be trained on the data set $\mathcal{D}$ and let the hold-out data set $\mathcal{D}_h : \mathcal{D}_h \cap \mathcal{D} = \emptyset$ be given. Then, given a pair of points $(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2})$ from different classes $y_{i_1} \neq y_{i_2}$ which are chosen randomly and uniformly from $\mathcal{D}_h$, the candidate for inclusion in the trigger set is of the form



Cub-200-2011

CIFAR-FS

*mini*ImageNet

Figure 4 — Dependency of certified accuracy on attack radius ε for different σ, 5-shot case.

$$x_i^* = \lambda x_{i_1} + (1-\lambda) x_{i_2},$$

where $\lambda \sim \mathcal{U}(0{,}1)$. To guarantee non-standard predefined behavior of the model f on an object x_i^*, we accept x_i^* as a candidate only if the original model predicts x_i^* as a representative of some other class $y_i^* : y_i^* \neq y_{i_1}, y_i^* \neq y_{i_2}$.

– In the second part, candidates for inclusion in the trigger set are verified. Note that the similarity in predictions of the original model f and some suspicious model f^* on the objects from the trigger set should be an indicator that model f^* is a functional copy of model f. To satisfy this property, we introduce a parametric set of proxy models $\mathcal{B}_{\delta,\tau}(f)$ that mimics the set of suspicious models. The set of proxy models is defined by two parameters, δ and τ. These parameters control the difference between the proxy models and the original model in the space of weights and the difference between accuracy of the models. Formally, the parametric set is defined as

$$\mathcal{B}_{\delta,\tau}(f) = \{f' : \|\theta(f') - \theta(f)\|_2 \leq \delta \text{ и } |\mathrm{acc}(\mathcal{D}, f') - \mathrm{acc}(\mathcal{D}{,}f)| \leq \tau\},$$

where $\theta(f)$ is the vector of weights of the model f, acc($\mathcal{D}$,f) is the accuracy of the model f on the dataset $\mathcal{D}$.

Next, m proxy models $f_1, \dots, f_m \in \mathcal{B}_{\delta,\tau}(f)$ are randomly selected to verify the transferability of behavior on the trigger set. Then, it is checked whether all m proxy models assign the same class label to an object in the trigger set as the original model f. In other words, a candidate object (x_i^*, y_i^*) is included in the trigger set if the following condition is satisfied:

$$y_i^* = f(x_i^*) = f_1(x_i^*) = \dots = f_m(x_i^*).$$

Section 4.5 describes the experiments. CIFAR-10 and CIFAR-100 [45] datasets are used as training datasets, and a convolutional neural network ResNet34 [27] is used as the source model. The proposed method is compared with existing digital watermarking approaches in the task of detection of the distillation-based functionality stealing attack in the following settings:

– Soft-label attack. In this setting, the training dataset $\mathcal{D}$ is known, and, given input x, the output $f(x)$ of the source model is a vector of class probabilities.

– Hard-label attack. In this setting, the training dataset $\mathcal{D}$ is known, and, given input x, the output $f(x)$ of the source model is the class label assigned by f to input x. This setting corresponds to the training of the surrogate model on the dataset $\hat{\mathcal{D}} = \{x_i, f(x_i)\}_{i=1}^N$.

– Regularization with ground truth label. In [50], it was proposed to train the surrogate model by minimizing the empirical loss on the training dataset $\mathcal{D}$ and the KL-divergence between the outputs of the source model and surrogate model simultaneously.

We report the accuracy values of the original and surrogate models on a test subset of the dataset $\mathcal{D}$, as well as the accuracy of the models on the trigger set $\mathcal{D}_t^*$. The results are reported in Table 4.

Section 4.6 of the chapter provides a discussion of the limitations of the proposed method. The final section of the chapter provides a discussion of the experimental results and identifies one of the directions for further research – the development of guarantees for the transferability of prediction on trigger sets to surrogate models.

Method	Metric	f	Surrogate models f^*		
			Soft-label	Hard-label	RGT
EWE [51]	CIFAR-10 acc. (%)	86.10 ± 0.54	83.97 ± 1.02	82.22 ± 0.50	88.88 ± 0.35
RS [52]		84.17 ± 1.01	88.93 ± 1.18	89.62 ± 0.97	90.14 ± 0.08
MB [50]		87.81 ± 0.76	91.17 ± 0.76	91.88 ± 0.40	93.05 ± 0.20
Ours		**91.00 ± 0.00**	**92.60 ± 0.91**	**94.87 ± 0.59**	**99.42 ± 0.02**
EWE [51]	Trigger set acc. (%)	26.88 ± 8.32	51.01 ± 5.58	36.05 ± 6.48	1.64 ± 1.05
RS [52]		95.67 ± 4.93	7.67 ± 4.04	6.33 ± 1.15	3.00 ± 0.00
MB [50]		100.00 ± 0.00	82.00 ± 1.00	51.33 ± 4.93	72.67 ± 6.66
Ours		100.00 ± 0.00	**85.10 ± 6.33**	**73.70 ± 4.65**	**78.00 ± 5.58**
EWE [51]	CIFAR-100 acc. (%)	55.11 ± 1.67	53.00 ± 1.57	46.78 ± 1.00	63.73 ± 0.40
RS [52]		59.87 ± 2.78	65.66 ± 1.53	65.79 ± 0.39	64.99 ± 0.30
MB [50]		62.13 ± 4.36	**67.66 ± 0.36**	**70.65 ± 0.49**	**70.24 ± 0.46**
Ours		**66.70 ± 0.00**	67.49 ± 0.03	68.05 ± 0.73	67.85 ± 0.04
EWE [51]	Trigger set acc. (%)	68.14 ± 10.16	30.90 ± 11.34	15.10 ± 5.64	5.73 ± 3.42
RS [52]		99.00 ± 1.00	2.67 ± 1.53	4.33 ± 4.16	2.00 ± 1.00
MB [50]		100.00 ± 0.00	70.67 ± 7.57	40.00 ± 8.89	62.66 ± 10.12
Ours		100.00 ± 0.00	**78.80 ± 2.93**	**74.70 ± 3.16**	**79.10 ± 2.77**

Table 4 — Watermarking performance is reported against functionality stealing methods. The best performance is highlighted in bold.

The **conclusion** briefly formulates the main results:

- The gradient-based approach to construct adversarial patches that demonstrate the vulnerability of face recognition systems in the real-world settings;
- A novel probabilistic approach to certify the robustness of neural networks to input perturbations of an arbitrary type without sacrificing the performance of the neural network;
- An approach to certify the prototypical neural networks to additive transformations of bounded magnitude in the few-shot learning setting;
- A novel method to generate digital watermarks as an indicator of the theft of the neural networks without sacrificing the performance of the source model.

Author's publications on the topic of the thesis

A1. Real-World Attack on MTCNN Face Detection System [Text] / E. Kaziakhmedov [et al.] // 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). — IEEE. 2019. — P. 0422—0427.

A2. CC-Cert: A probabilistic Approach to Certify General Robustness of Neural Networks [Text] / M. Pautov [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. — 2022. — P. 7975—7983.

A3. Smoothed Embeddings for Certified Few-Shot Learning [Text] / M. Pautov [et al.] // Advances in Neural Information Processing Systems. — 2022. — Vol. 35. — P. 24367—24379.

A4. On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System [Text] / M. Pautov [et al.] // 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). — IEEE. 2019. — P. 0391—0396.

A5. Translate Your Gibberish: Black-Box Adversarial Attack on Machine Translation Systems [Text] / A. Chertkov [et al.] // arXiv preprint arXiv:2303.10974. — 2023.

A6. Probabilistically Robust Watermarking of Neural Networks [Text] / M. Pautov [et al.] // arXiv preprint arXiv:2401.08261. — 2024.

References

1. Deep residual learning for image recognition [Text] / K. He [et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2016. — P. 770—778.
2. Image segmentation using deep learning: A survey [Text] / S. Minaee [et al.] // IEEE transactions on pattern analysis and machine intelligence. — 2021. — Vol. 44, no. 7. — P. 3523—3542.
3. Object detection in 20 years: A survey [Text] / Z. Zou [et al.] // Proceedings of the IEEE. — 2023.
4. Efficient object localization using convolutional networks [Text] / J. Tompson [et al.] // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2015. — P. 648—656.
5. End to end learning for self-driving cars [Text] / M. Bojarski [et al.] // arXiv preprint arXiv:1604.07316. — 2016.
6. *Sarvamangala*, *D.* Convolutional neural networks in medical image understanding: a survey [Text] / D. Sarvamangala, R. V. Kulkarni // Evolutionary intelligence. — 2022. — Vol. 15, no. 1. — P. 1—22.
7. *Aydin*, *I.* A new IoT combined face detection of people by using computer vision for security application [Text] / I. Aydin, N. A. Othman // 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). — IEEE. 2017. — P. 1—6.
8. Intriguing properties of neural networks [Text] / C. Szegedy [et al.] // arXiv preprint arXiv:1312.6199. — 2013.
9. *Moosavi-Dezfooli*, *S.-M.* Deepfool: a simple and accurate method to fool deep neural networks [Text] / S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2016. — P. 2574—2582.

10. Universal adversarial perturbations [Text] / S.-M. Moosavi-Dezfooli [et al.] // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — P. 1765—1773.
11. *Carlini*, *N.* Towards evaluating the robustness of neural networks [Text] / N. Carlini, D. Wagner // 2017 ieee symposium on security and privacy (sp). — IEEE. 2017. — P. 39—57.
12. Adversarial defense via learning to generate diverse attacks [Text] / Y. Jang [et al.] // Proceedings of the IEEE/CVF International Conference on Computer Vision. — 2019. — P. 2740—2749.
13. Towards deep learning models resistant to adversarial attacks [Text] / A. Madry [et al.] // arXiv preprint arXiv:1706.06083. — 2017.
14. *Zhou*, *J.* Manifold Projection for Adversarial Defense on Face Recognition [Text] / J. Zhou, C. Liang, J. Chen // European Conference on Computer Vision. — Springer. 2020. — P. 288—305.
15. *Cohen*, *J.* Certified adversarial robustness via randomized smoothing [Text] / J. Cohen, E. Rosenfeld, Z. Kolter // International Conference on Machine Learning. — PMLR. 2019. — P. 1310—1320.
16. *Rudin*, *L. I.* Nonlinear total variation based noise removal algorithms [Text] / L. I. Rudin, S. Osher, E. Fatemi // Physica D: nonlinear phenomena. — 1992. — Vol. 60, no. 1—4. — P. 259—268.
17. Synthesizing Robust Adversarial Examples [Text] / A. Athalye [et al.] // ICML. — 2017.
18. *Deng*, *L.* The mnist database of handwritten digit images for machine learning research [best of the web] [Text] / L. Deng // IEEE signal processing magazine. — 2012. — Vol. 29, no. 6. — P. 141—142.
19. Learning multiple layers of features from tiny images [Text] / A. Krizhevsky, G. Hinton, [et al.]. — 2009.
20. The caltech-ucsd birds-200-2011 dataset [Text] / C. Wah [et al.]. — 2011.
21. Matching networks for one shot learning [Text] / O. Vinyals [et al.] // Advances in neural information processing systems. — 2016. — Vol. 29. — P. 3630—3638.
22. Imagenet large scale visual recognition challenge [Text] / O. Russakovsky [et al.] // International journal of computer vision. — 2015. — Vol. 115, no. 3. — P. 211—252.
23. Meta-learning with differentiable closed-form solvers [Text] / L. Bertinetto [et al.] // arXiv preprint arXiv:1805.08136. — 2018.
24. Margin-based neural network watermarking [Text] / B. Kim [et al.] // International Conference on Machine Learning. — PMLR. 2023. — P. 16696—16711.

25. Entangled watermarks as a defense against model extraction [Text] / H. Jia [et al.] // 30th USENIX Security Symposium (USENIX Security 21). — 2021. — P. 1937—1954.

26. Certified neural network watermarks with randomized smoothing [Text] / A. Bansal [et al.] // International Conference on Machine Learning. — PMLR. 2022. — P. 1450—1465.