REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
18-04-2022	Final Report	1-Jan-2019 - 31-Dec-2021

4. TITLE AND SUBTITLE
Final Report: Sparsity-based Design for Robust Deep Learning - Topic C. iii (3)

5a. CONTRACT NUMBER
W911NF-19-1-0053

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of California - Santa Barbara
3227 Cheadle Hall
3rd floor, MC 2050
Santa Barbara, CA 93106 -2050

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
73518-CS.6

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Upamanyu Madhow
UU	UU	UU	UU		19b. TELEPHONE NUMBER 805-893-5210

RPPR Final Report
as of 21-Apr-2022

Agency Code: 21XD

Proposal Number: 73518CS **Agreement Number: W911NF-19-1-0053**
INVESTIGATOR(S):

Name: Ph.D. Upamanyu Madhow
Email: madhow@ece.ucsb.edu
Phone Number: 8058935210
Principal: Y

Organization: **University of California - Santa Barbara**
Address: 3227 Cheadle Hall, Santa Barbara, CA 931062050
Country: USA
DUNS Number: 094878394 EIN: 956006145W
Report Date: 31-Mar-2022 Date Received: 18-Apr-2022
Final Report for Period Beginning 01-Jan-2019 and Ending 31-Dec-2021
Title: Sparsity-based Design for Robust Deep Learning - Topic C. iii (3)
Begin Performance Period: 01-Jan-2019 **End Performance Period:** 31-Dec-2021
Report Term: 0-Other
Submitted By: Ph.D. Upamanyu Madhow Email: madhow@ece.ucsb.edu
Phone: (805) 893-5210

STEM Degrees: 1 **STEM Participants:** 4

Major Goals: Deep neural networks [DNNs] yield state of the art performance in many fields, but are known to be vulnerable to small adversarial perturbations. This, together with their lack of interpretability, is a major impediment to their use in many DoD applications, as well as in safety-critical commercial applications such as vehicular autonomy. The overarching goal of this project is to investigate techniques for understanding and robustifying DNNs. The original proposal focused on the specific approach of imposing sparsity constraints to attenuate and eliminate adversarial perturbations, with the goal of obtaining interpretable designs with guaranteed resilience. This is in contrast to the state of art defenses against adversarial perturbations, which are based on black box training with adversarially perturbed examples. The DNNs obtained by such adversarial training can only be empirically validated, and do not provide performance guarantees, especially against potentially novel threat models. The rationale behind our sparsity-based design approach is that, since small, well-designed, perturbations can add up to large values in high-dimensional spaces, their impact can be reduced by exploiting low-dimensional structures in natural data to limit the dimension of the space the adversary can operate over. For example, unlabeled data can be used to learn dictionaries for sparse coding as a preprocessing step before application of DNNs. A complementary approach is to impose sparsity constraints on weights in a DNN in order to control the size of adversarial perturbations flowing up the network.

Since this project was funded, our thinking on broadened significantly. We recognize that vulnerability to adversarial examples is a symptom of a broader problem with the current top-down approach for training DNNs: we do not control or understand the features we are extracting. We also realized that, while sparsity alone provides some attenuation of adversarial perturbations, it must be coupled with more drastic nonlinearities to further attenuate the impact of perturbations. Finally, we sought detection-theoretic insights into the structure of robust classifiers by analysis of adversarial attacks in simplified settings.

We also initiated research on deep learning for wireless fingerprinting, using wireless data obtained through a collaboration with Teledyne in Phase 1 of the DARPA RFMLS project. Exploring the use of DNNs to obtain RF signatures for physical layer security provides an opportunity to explore robustness of DNNs from a different, yet also fundamental, angle. The goal is to obtain RF signatures that enable distinguishing between transceivers based on manufacturing variations in typical non-idealities such as power amplifier nonlinearity, digital-to-analog converter nonlinearity, and I-Q mismatch. Since these effects are difficult to model explicitly, implicitly learning these signatures using DNNs is a natural approach. Our goal is to explore the robustness and stability of such signatures against spatio-temporal variations. For example, in order for the signatures to be stable in time, they must be invariant to drifts in carrier frequency offset, and in order for them to be stable against changes in device location, they should be invariant to changes in the propagation channel. However, a DNN trained using data

acquired over a relatively short period of time from a given location will actually lock on to such more obvious phenomena rather than the subtle device-specific nonlinear effects we are looking for. In addition to its application to wireless security, this problem provides an opportunity to understand, and learn to design around, the limitations of DNNs in general, by specializing to a scenario where we can leverage the deep understanding of input data acquired over decades of research in wireless communication.

Finally, over the past year, we initiated a completely new line of inquiry into fairness in machine learning. While there is significant effort in the community in design of machine learning algorithms to address bias in the data or algorithms, but we explore a framework for sequential decision-making aimed at dynamically influencing long-term societal fairness.
This is illustrated via the problem of choosing applicants from a pool consisting of two groups, one of which is under-represented. We devise the algorithms based on a dynamic model for the composition of the applicant pool, in which admission of more applicants from a group in a given round positively reinforces more candidates from the group to participate in future rounds.

Accomplishments: The research accomplishments during this project are organized in the following categories (the first pages of a representative set of publications has been uploaded as a pdf file):

Robust front ends for DNNs (ICASSP 2020, ICIP 2021): One approach to adversarial robustness is to devise Òfront endsÓ that attenuate perturbations, followed by a standard DNN. Projecting onto a sparse basis alone is not enough, and we have shown that learnt bases aimed at sparsifying the output, along with drastic nonlinearities, get close to state of the art black box adversarial training in terms of robustness. in ICASSP 2020 and ICIP 2021.

New DNN architectures targeting robustness and interpretability (submitted for publication): Going beyond front ends, in recent work, we rethink DNN architectures using inspiration from communication theory and neuroscience, with the goal of producing sparse strong activations in each layer. In essence, we wish to learn matched filters at each layer. To this end, we introduce layerwise costs in training that produce Hebbian (Òfire together, wire togetherÓ) updates for neurons stimulated most by an input, and anti-Hebbian updates for other neurons. We find that the resulting DNN is more robust to both noise and adversarial perturbations than a baseline network, even though we do not explicitly train for robustness. These exciting preliminary results motivate a thorough investigation of Hebbian/anti-Hebbian (HaH) updates in deep learning.

Detection-theoretic insights into robust classification (ICASSP 2021, journal paper submitted for publication): In this work, we take a step back from deep learning, seeking to develop fundamental insight into defending against adversarial perturbations in a classical hypothesis testing problem. Interpreting an adversarial perturbation as a nuisance parameter, we investigate the generalized likelihood ratio test (GLRT) for the resulting composite hypothesis testing problem, jointly estimating the class of interest and the adversarial perturbation. For a simple problem in which a minimax approach is known, we show that the GLRT defense is competitive with the minimax approach under the worst-case attack, while yielding a better robustness-accuracy tradeoff under weaker attacks. In future work, we seek to apply the insights obtained from such simple models to design robust machine learning architectures.

Addressing confounding factors in wireless fingerprinting (Globecom 2019, invited presentation at ITA 2020, Asilomar 2021): For our work on extracting RF signatures using DNNs, the emphasis is on understanding the impact of confounding factors. Our goal is to derive location- and environment-independent signatures that depend on subtle nonlinear variations across transceivers, but DNNs are apt to lock onto confounding factors such as device ID fields which can be easily spoofed, and propagation channels or carrier frequency offsets that can vary across space and time. We devise model-based data augmentation for training to overcome this problem, and obtain the novel finding that augmentation is helpful even during inference. This work has appeared in several venues, and has drawn interest from several companies. While the results are promising, there is clearly a lot of work to be done in terms of developing fundamental limits and practical strategies for physical layer security via wireless fingerprinting.

Positive feedback for long-term fairness (ICLR 2022 Workshop on Socially Responsible Machine Learning): in very recent work, we investigate a framework for sequential decision-making aimed at dynamically influencing long-term societal fairness, illustrated via the problem of choosing applicants from a pool consisting of two groups, one of which is under-represented. We consider a dynamic model for the composition of the applicant pool, in which admission of more applicants from a group in a given selection round positively re- inforces more candidates from the group to participate in future selection rounds. Under such a model, we characterize and evaluate strategies

that trade off performance and fairness. In addition to experimenting on synthetic data, we adapt static real-world datasets on law school candidates and credit lending to simulate the dynamics of the composition of the applicant pool. This work opens up a broader research agenda on long-term strategies for mitigating societal bias, which we will seek to pursue in collaboration with social scientists.

Training Opportunities: The project provided partial support to four graduate students, Soorya Gopalakrishnan, Metehan Cekic, Can Bakiskan and Bhagyashree Puranik. Soorya graduated in Spring 2020, and has joined a machine learning group at Qualcomm. Metehan and Can have cleared their PhD Qualifying Exams, and pursued machine learning internships at Amazon and Intel in Summer 2021. They will graduate with their PhDs in Summer 2022, and already have job offers from the companies where they interned. We expect Bhagyashree, who completed an internship with Qualcomm in Summer 2021, to take her PhD Qualifying Exam in Spring or

Results Dissemination: Several papers have been published (the uploaded pdf concatenates a representative set of publications supported by this grant). The students supported on this grant gave conference presentations at Globecom 2019 (in person), ICASSP 2020, ICASSP 2021, and ICIP 2021 (via video recording due to COVID-19 issues). The PI gave a well-received invited presentation at the 2020 Information Theory and Applications workshop (ITA 2020) in San Diego, CA, which led to follow-up inquiries from several companies, including Qualcomm and Intel.

Honors and Awards: The PI received a Distinguished Alumni award from the ECE Department at the University of Illinois at Urbana-Champaign, where he got his MS and PhD degrees.

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Upamanyu Madhow
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Can Bakiskan
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Metehan Cekic
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Bhagyashree Puranik
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

RPPR Final Report
as of 21-Apr-2022

Participant Type: Graduate Student (research assistant)
Participant: Soorya Gopalakrishnan
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: ICASSP 2020
Date Received: 02-Jul-2020 Conference Date: 04-May-2020 Date Published: 04-May-2020
Conference Location: Barcelona, Spain (virtual)
Paper Title: Polarizing front ends for robust CNNs
Authors: Can Bakiskan, Soorya Gopalakrishnan, Metehan Cekic, Upamanyu Madhow, Ramtin Pedarsani
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE Globecom 2019
Date Received: Conference Date: 09-Oct-2019 Date Published: 09-Dec-2019
Conference Location: Waikoloa, HI, USA
Paper Title: Robust wireless fingerprinting via complex-valued neural networks
Authors: Soorya Gopalakrishnan, Metehan Cekic, Upamanyu Madhow
Acknowledged Federal Support: **Y**

WEBSITES:

URL: https://wcsl.ece.ucsb.edu/combating-adversarial-attacks-using-sparse-representations
Date Received: 02-Jul-2020
Title: Combating Adversarial Attacks via Sparse Representations
Description: Highlights role of sparsity in robustness of neural networks, and provides link to code on github
URL: https://wcsl.ece.ucsb.edu/radio-frequency-machine-learning-rfml
Date Received: 02-Jul-2020
Title: Radio Frequency Machine Learning (RFML)
Description: Describes research effort on learning robust RF signatures, and links to publications and preprints. Code will be linked soon.
URL: https://wcsl.ece.ucsb.edu/MLcourse
Date Received: 02-Jul-2020
Title: Machine Learning: A Signal Processing Perspective
Description: Handouts and lecture videos for a graduate machine learning course created by the PI, laying the foundations for the signal processing perspective we are trying to bring into machine learning research.

RPPR Final Report
as of 21-Apr-2022

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Upamanyu Madhow
Signature Date: 4/18/22 1:04PM

2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506184

A NEURO-INSPIRED AUTOENCODING DEFENSE AGAINST ADVERSARIAL ATTACKS

Can Bakiskan *Metehan Cekic* *Ahmet Dundar Sezer* *Upamanyu Madhow*

Department of Electrical and Computer Engineering
University of California Santa Barbara, Santa Barbara, CA 93106
{canbakiskan,metehancekic,adsezer,madhow}@ece.ucsb.edu

ABSTRACT

Deep Neural Networks (DNNs) are vulnerable to adversarial attacks: carefully constructed perturbations to an image can seriously impair classification accuracy, while being imperceptible to humans. The most effective current defense is to train the network using adversarially perturbed examples. In this paper, we investigate a radically different, neuro-inspired defense mechanism, aiming to reject adversarial perturbations *before* they reach a classifier DNN, using an encoder with characteristics commonly observed in biological vision, followed by a decoder restoring image dimensions that can be cascaded with standard CNN architectures. Unlike adversarial training, all training is based on *clean* images. Our experiments on the CIFAR-10 and a subset of Imagenet datasets show performance competitive with state-of-the-art adversarial training, and point to the promise of bottom-up neuro-inspired techniques for the design of robust neural networks.

***Index Terms*—** Adversarial, Machine learning, Robust, Image classification, Defense

1. INTRODUCTION

The susceptibility of neural networks to small, carefully crafted input perturbations raises great concern regarding their robustness and security. Since this vulnerability of DNNs was pointed out [1, 2], there have been numerous studies on how to generate these perturbations (adversarial attacks) [3, 4] and how to defend against them [4, 5, 6, 7]. Existing defenses that attempt to employ systematic or provable techniques either do not scale to large networks, or have been defeated by appropriately modified attacks [5, 6, 8]. State of the art defenses [4, 9, 10] employ adversarial training (i.e., training the model with adversarially perturbed examples), but there is little insight into how DNNs designed in this end-to-end, "top down" fashion provide robust performance.

Approach: In this paper, we turn to neuro-inspiration for defending against adversarial attacks, inspired by the observation that humans barely register adversarial perturbations devised for machines. While neuro-inspiration could ultimately provide a general framework for designing DNNs which are robust to a variety of perturbations, in this paper, we take a first step by focusing on the well-known ℓ^∞ bounded attack, which captures the concept of "barely noticeable" perturbation. Our architecture, illustrated in Figure 1, does not require adversarial training: it consists of (a) a neuro-inspired encoder learnt in purely unsupervised fashion, (b) a decoder which produces an output of the same size as the original image, (c) a standard CNN for classification. The decoder and classifier are trained in standard supervised fashion using *clean* images passed through our encoder.

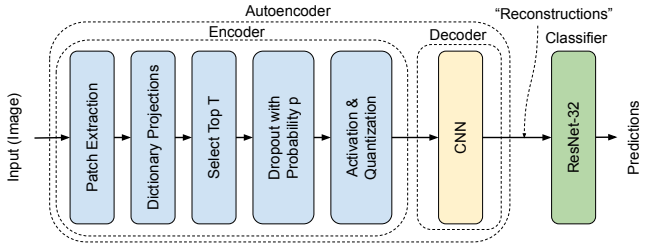

Fig. 1: Proposed autoencoding defense. Decoder restores input size but does not attempt to reconstruct the input in our nominal design (supervised decoder+classifier training).

The key features we incorporate into our encoder design are sparsity and overcompleteness, long conjectured to be characteristic of the visual system [11], lateral inhibition [12], synaptic noise [13], and drastic nonlinearity [14]. We use standard unsupervised dictionary learning [15] to learn a sparse, highly overcomplete (5-10X relative to ambient dimension) patch-level representations. However, we use the learnt dictionary in a non-standard manner in the encoder, not attempting patch-level reconstructions. Instead, we take the top T coefficients from each patch (lateral inhibition), randomly drop a fraction p of them (synaptic noise and lateral inhibition), and threshold and quantize them, retaining only their sign (drastic nonlinearity). We use overlapping patches, providing an additional degree of overcompleteness. The patch-level outputs, which have ternary quantized entries, are fed to a multi-layer CNN decoder whose output is the same size as the original RGB image input. This is then fed to a standard classifier DNN.

We report on experiments on the CIFAR-10 and a subset of the ImageNet dataset ("Imagenette"), demonstrating the promise of a "bottom-up" neuro-inspired approach, in contrast

DYNAMIC POSITIVE REINFORCEMENT FOR LONG-TERM FAIRNESS

Bhagyashree Puranik, Upamanyu Madhow & Ramtin Pedarsani
Department of Electrical and Computer Engineering
University of California Santa Barbara
Santa Barbara, CA 93106, USA
{bpuranik, madhow, ramtin}@ucsb.edu

ABSTRACT

We propose a framework for sequential decision-making aimed at dynamically influencing long-term societal fairness, illustrated via the problem of selecting applicants from a pool consisting of two groups, one of which is under-represented. We consider a dynamic model for the composition of the applicant pool, in which admission of more applicants from a group in a given selection round positively reinforces more candidates from the group to participate in future selection rounds. Under such a model, we show the efficacy of the proposed *Fair-Greedy* selection policy which systematically trades the sum of the scores of the selected applicants ("greedy") against the deviation of the proportion of selected applicants belonging to a given group from a target proportion ("fair"). In addition to experimenting on synthetic data, we adapt static real-world datasets on law school candidates and credit lending to simulate the dynamics of the composition of the applicant pool. We prove that the applicant pool composition converges to a target proportion set by the decision-maker when score distributions across the groups are identical.

1 INTRODUCTION

In this paper, we seek to develop a framework for sequential decision making aimed at influencing long-term societal fairness. Machine learning models are being increasingly applied in making critical decisions that affect humans, such as recidivism prediction (Dressel & Farid (2018)), mortgage lending (Berkovec et al. (2018)), and recommendation systems (Yao & Huang (2017)). While the algorithms offer increased efficiency, speed, and scalability, they could introduce bias leading to the decisions being unfair towards certain groups of the population. There is a rich and rapidly growing literature on "fair" strategies that mitigate bias in algorithmic decision making, including label or data pre-processing and cost reweighting based on groups (Kamiran & Calders (2012)), addition of constraints that satisfy fairness criteria (Zafar et al. (2017)), and learning representations that obfuscate group information (Zemel et al. (2013)). Most strategies consider a static setting or study short-term impact of decisions on population (Liu et al. (2018)), with the exception of some recent studies on the long-term impact of fairness-aware decisions on the qualification of the population (Zhang et al. (2020); Mouzannar et al. (2019); Hu et al. (2019); Williams & Kolter (2019)). (See Appendix A.1 for a more detailed discussion of related work.)

Our framework is motivated by real-world examples such as the following. Consider a company receiving applications every month, which wants to hire in an unbiased manner (e.g., by ultimately selecting equal numbers of male and female applicants). With the total intake fixed based on a budget, the company selects a certain proportion of candidates from each group. The hiring decisions affect the subsequent pool of applicants: admitting more candidates from a particular group might encourage more such candidates to apply, or successful candidates from a group might inspire other such candidates, providing positive feedback into the decision-making loop. Such a strategy could not only enhance diversity and equity, but also enable the company to learn more about a minority group so as to eventually have a richer pool of well-qualified applicants. Another motivating example is college admissions, where the goal may be to admit students with the best academic records, while accounting for socio-economic background and reducing bias based on sensitive attributes such as race or gender. Could one, for example, reverse the trend in the decrease in the proportion

ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413587

ADVERSARIALLY ROBUST CLASSIFICATION BASED ON GLRT

Bhagyashree Puranik, Upamanyu Madhow, Ramtin Pedarsani

University of California Santa Barbara

ABSTRACT

Machine learning models are vulnerable to adversarial attacks that can often cause misclassification by introducing small but well designed perturbations. In this paper, we explore, in the setting of classical composite hypothesis testing, a defense strategy based on the generalized likelihood ratio test (GLRT), which jointly estimates the class of interest and the adversarial perturbation. We evaluate the GLRT approach for the special case of binary hypothesis testing in white Gaussian noise under ℓ_∞ norm-bounded adversarial perturbations, a setting for which a minimax strategy optimizing for the worst-case attack is known. We show that the GLRT approach yields performance competitive with that of the minimax approach under the worst-case attack, while yielding a better robustness-accuracy trade-off under weaker attacks. The GLRT defense is applicable in multi-class settings and generalizes naturally to more complex models for which optimal minimax classifiers are not known.

***Index Terms*—** Adversarial machine learning, hypothesis testing, robust classification

1. INTRODUCTION

Machine learning models such as deep neural networks and regression methods have become pervasively deployed in large-scale commercial applications that are safety-critical, such as facial recognition for surveillance, autonomous driving and virtual assistants. It has been shown that an adversary is often able to add small perturbations to signals in an intelligent way to cause misclassification with high confidence [1, 2]. In applications that demand robustness in machine learning methods, adversarial attacks are fundamental threats. There have been several defense mechanisms suggested, followed by proposal of stronger adversaries to circumvent the defenses [3, 4]. A state-of-the-art defense [5] against such attacks is to train with adversarial examples–this is purely empirical and cannot provide robustness guarantees or insights.

In this paper, we seek fundamental insight by investigating adversarial classification in the setting of classical hypothesis testing, in which the class-conditional distributions of the data is known. We propose the well-known GLRT as a general approach to defense, in which the desired class and the action of the adversary (viewed as a nuisance parameter) are estimated jointly. The GLRT approach is general, since it applies to any composite hypothesis testing problem [6], unlike minimax strategies optimizing for worst-case attacks, which are difficult to find. We compare the GLRT and minimax approaches for a simple setting, binary Gaussian hypothesis testing with ℓ_∞ bounded attacks, for which the minimax strategy has been derived [7]. We show that the proposed GLRT approach provides competitive robustness guarantees when the attacker employs the full attack budget, while providing better robustness-accuracy trade-off for weaker attacks.

Related Work: There is a growing body of research on coming up with provable robustness guarantees against adversarial attacks [8, 9, 10, 11, 12, 13, 14, 15, 16]. A recent paper [17] addresses the problem of finding optimal robust classifiers in a binary classification problem, with the class conditional distributions possessing symmetric means and white Gaussian noise. For the case when perturbations are ℓ_∞ norm bounded, they restrict attention to the class of linear classifiers and then obtain optimum robust linear classifiers for two and three-class classification problems. In general, finding robust optimal classifiers for ℓ_∞ norm bounded adversarial perturbations is not easily tractable. Analytical results have been shown only for special cases, such as in [7], where minimax optimal robust classifiers are characterized in binary classification setting under Gaussian models with symmetric means, same covariance matrices and uniform priors, using ideas from optimal transport theory. Our proposed GLRT defense can be applied to multi-class Gaussian hypothesis problems with generic means and priors. In addition, the minimax classifier in [7] is pessimistic for weaker attacks, while the GLRT scheme performs better in such regimes as it estimating the action of the attacker.

2. GLRT-BASED DEFENSE

Throughout the paper, we represent vectors in boldface letters and scalars in regular letters. The norm $||\cdot||$ denotes ℓ_2 norm unless specified otherwise. Consider the following standard classification or hypothesis testing problem: $\mathcal{H}_k : \mathbf{X} \sim p_k(\mathbf{x})$. The presence of an adversary increases the uncertainty

This work was supported by the Army Research Office under grant W911NF-19-1-0053, and by the National Science Foundation under grant CCF 1909320.

2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506184

A NEURO-INSPIRED AUTOENCODING DEFENSE AGAINST ADVERSARIAL ATTACKS

Can Bakiskan *Metehan Cekic* *Ahmet Dundar Sezer* *Upamanyu Madhow*

Department of Electrical and Computer Engineering
University of California Santa Barbara, Santa Barbara, CA 93106
{canbakiskan,metehancekic,adsezer,madhow}@ece.ucsb.edu

ABSTRACT

Deep Neural Networks (DNNs) are vulnerable to adversarial attacks: carefully constructed perturbations to an image can seriously impair classification accuracy, while being imperceptible to humans. The most effective current defense is to train the network using adversarially perturbed examples. In this paper, we investigate a radically different, neuro-inspired defense mechanism, aiming to reject adversarial perturbations *before* they reach a classifier DNN, using an encoder with characteristics commonly observed in biological vision, followed by a decoder restoring image dimensions that can be cascaded with standard CNN architectures. Unlike adversarial training, all training is based on *clean* images. Our experiments on the CIFAR-10 and a subset of Imagenet datasets show performance competitive with state-of-the-art adversarial training, and point to the promise of bottom-up neuro-inspired techniques for the design of robust neural networks.

***Index Terms*—** Adversarial, Machine learning, Robust, Image classification, Defense

1. INTRODUCTION

The susceptibility of neural networks to small, carefully crafted input perturbations raises great concern regarding their robustness and security. Since this vulnerability of DNNs was pointed out [1, 2], there have been numerous studies on how to generate these perturbations (adversarial attacks) [3, 4] and how to defend against them [4, 5, 6, 7]. Existing defenses that attempt to employ systematic or provable techniques either do not scale to large networks, or have been defeated by appropriately modified attacks [5, 6, 8]. State of the art defenses [4, 9, 10] employ adversarial training (i.e., training the model with adversarially perturbed examples), but there is little insight into how DNNs designed in this end-to-end, "top down" fashion provide robust performance.

Approach: In this paper, we turn to neuro-inspiration for defending against adversarial attacks, inspired by the observation that humans barely register adversarial perturbations devised for machines. While neuro-inspiration could ultimately provide a general framework for designing DNNs which are robust to a variety of perturbations, in this paper, we take a first step by focusing on the well-known ℓ^∞ bounded attack, which captures the concept of "barely noticeable" perturbation. Our architecture, illustrated in Figure 1, does not require adversarial training: it consists of (a) a neuro-inspired encoder learnt in purely unsupervised fashion, (b) a decoder which produces an output of the same size as the original image, (c) a standard CNN for classification. The decoder and classifier are trained in standard supervised fashion using *clean* images passed through our encoder.

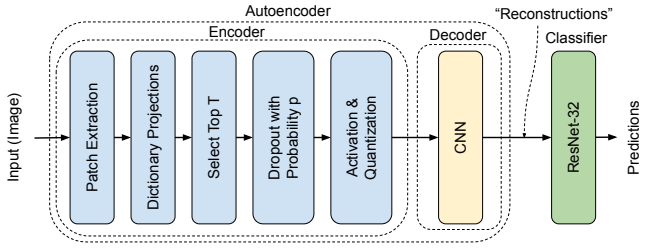

Fig. 1: Proposed autoencoding defense. Decoder restores input size but does not attempt to reconstruct the input in our nominal design (supervised decoder+classifier training).

The key features we incorporate into our encoder design are sparsity and overcompleteness, long conjectured to be characteristic of the visual system [11], lateral inhibition [12], synaptic noise [13], and drastic nonlinearity [14]. We use standard unsupervised dictionary learning [15] to learn a sparse, highly overcomplete (5-10X relative to ambient dimension) patch-level representations. However, we use the learnt dictionary in a non-standard manner in the encoder, not attempting patch-level reconstructions. Instead, we take the top T coefficients from each patch (lateral inhibition), randomly drop a fraction p of them (synaptic noise and lateral inhibition), and threshold and quantize them, retaining only their sign (drastic nonlinearity). We use overlapping patches, providing an additional degree of overcompleteness. The patch-level outputs, which have ternary quantized entries, are fed to a multi-layer CNN decoder whose output is the same size as the original RGB image input. This is then fed to a standard classifier DNN.

We report on experiments on the CIFAR-10 and a subset of the ImageNet dataset ("Imagenette"), demonstrating the promise of a "bottom-up" neuro-inspired approach, in contrast

Wireless Fingerprinting via Deep Learning: The Impact of Confounding Factors

Metehan Cekic* Soorya Gopalakrishnan*† Upamanyu Madhow
University of California, Santa Barbara

2021 55th Asilomar Conference on Signals, Systems, and Computers | 978-1-6654-5828-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEEECONF53345.2021.9723393

***Abstract*—Can we distinguish between two wireless transmitters sending exactly the same message, using the same protocol? The opportunity for doing so arises due to subtle nonlinear variations across transmitters, even those made by the same manufacturer. Since these effects are difficult to model explicitly, we investigate learning device fingerprints using complex-valued deep neural networks (DNNs) that take as input the complex baseband signal at the receiver. We ask whether such fingerprints can be made robust to distribution shifts across time and locations due to clock drift and variations in the wireless channel. In this paper, we point out that, unless proactively discouraged from doing so, DNNs learn these strong confounding features rather than the nonlinear device-specific characteristics that we seek to learn. We propose and evaluate strategies, based on augmentation and estimation, to promote generalization across realizations of these confounding factors, using WiFi data. We conclude that, while DNN training has the advantage of not requiring explicit signal models, significant modeling insights are required to focus the learning on the effects we wish to capture.**

I. Introduction

An important tool in wireless security is a "fingerprint" based on physical layer characteristics, capable of distinguishing between different devices even if they are transmitting exactly the same message. This is possible due to subtle hardware imperfections that occur even in devices made by the same manufacturer [1, 2]. Variations in components such as digital-to-analog converters (DACs) [3], power amplifiers (PAs) [4, 5] and I/Q imbalances [6] are inevitable even for transmitters manufactured using exactly the same process. Transistors, resistors, inductors, and capacitors within a device vary around nominal values, typically within a designed level of tolerance, and the goal is to translate the resulting variations in transmitter characteristics into a device signature.

Fingerprints based on such variations could potentially serve as a powerful authentication tool at the physical layer, complementing conventional security schemes in higher layers of the networking stack. Since these subtle nonlinear effects are difficult to model explicitly, deep learning is a natural approach to teasing out transceiver signatures based on them. In this paper, we investigate the efficacy of extracting fingerprints which are robust to variations across time and location, using one-dimensional complex-valued convolutional neural networks (CNNs) that operate on the complex baseband signal at the receiver.

Our results show that deep learning is a promising tool for wireless fingerprinting, while sounding a cautionary note. The key message is that the network learns the easiest set of features that it can in order to accomplish the desired task (in our case, discriminating between transmitters based on the received wireless signal), hence we must be extremely proactive in promoting robustness across effects that we do not want the network to lock on to, which we term *confounding factors*. For instance, we would like the radio frequency (RF) signature for a transmitter to be robust across time and for different wireless channels. However, if we employ training data collected over a period of time when the channel and carrier frequency offset (CFO) for a transmitter are relatively constant, the CNN will lock onto these rather than to subtle nonlinear effects. This gives unreasonably excellent accuracy on test data collected over the same time period, but disastrous results for data collected on a different day, when both the channel and the CFO can be different.

We develop augmentation strategies based on signal models for the impact of confounding factors, and evaluate performance against classical compensation techniques that explicitly try to undo them. We find that compensation works well if the undesired features are simple enough, like the CFO. However, for more complex effects such as a multipath channel, model-driven augmentation outperforms explicit estimation and compensation for learning robust signatures. A significant finding is that augmentation is useful not just during training, but also during inference: averaging of predictions from multiple augmented versions of the same input provides significant performance gains.

In order to perform controlled experiments, we evaluate our approach on *emulated* data, in which "clean" measured data is passed through simulated channels and CFO. Since this measured data, obtained as part of the DARPA RFMLS program, cannot be made public, we also make publicly available a *simulation-based* dataset based on models of some typical circuit-level nonlinearities [6–8]. The results we obtain on this dataset are comparable to those from the measurement-based dataset, enabling reproducibility. The dataset and code are available at [9].

Since we wish to be robust against software spoofing, we focus on extracting fingerprints only from the packet preamble. This is a worst-case approach motivated by prior work [10]

*Joint first authors.
†Currently at Qualcomm, San Diego.
Email: metehancekic@ucsb.edu, soorya197@gmail.com, madhow@ucsb.edu

POLARIZING FRONT ENDS FOR ROBUST CNNS

Can Bakiskan * *Soorya Gopalakrishnan* *Metehan Cekic* *Upamanyu Madhow* *Ramtin Pedarsani*

University of California, Santa Barbara, Department of Electrical and Computer Engineering

Index Terms— adversarial machine learning, quantization, front-end defense

ABSTRACT

The vulnerability of deep neural networks to small, adversarially designed perturbations can be attributed to their "excessive linearity." In this paper, we propose a bottom-up strategy for attenuating adversarial perturbations using a nonlinear front end which polarizes and quantizes the data. We observe that ideal polarization can be utilized to completely eliminate perturbations, develop algorithms to learn approximately polarizing bases for data, and investigate the effectiveness of the proposed strategy on the MNIST and Fashion MNIST datasets.

1. INTRODUCTION

Given the immense impact of deep learning on a diversity of fields, its vulnerability to tiny *adversarial* perturbations [1, 2] is of great concern. For image datasets, for example, such perturbations are almost imperceptible for humans, but they can render state-of-the-art models useless, causing misclassification with high confidence. State of the art adversarial attacks are variants of gradient ascent, utilizing the local linearity of deep networks. State of the art defenses are based on adversarial training, using training examples obtained using adversarial attacks, but yield little insight into, or guarantees of, the achieved robustness.

In this paper, we investigate a systematic, bottom-up approach to robustness, studying a defense based on a nonlinear front end for attenuating adversarial perturbations before they reach the deep network. We focus on ℓ_∞-bounded perturbations. Our approach consists of *polarizing* the input data into well-separated clusters by projecting onto an appropriately selected basis (implemented using convolutional filters), and then quantizing the output using thresholds that scale with the ℓ_1 norm of the basis functions. For ideal polarization, we prove that perturbations are completely eliminated. We introduce a regularization technique to learn polarizing bases from data, and demonstrate the efficacy of the proposed defense for the MNIST and Fashion MNIST datasets.

*Corresponding author: canbakiskan@ucsb.edu

2. BACKGROUND

Suppose we have a classifier that takes in inputs $\boldsymbol{x} \in \mathbb{R}^N$, and outputs predictions (confidence scores for M classes) $\boldsymbol{y} \in [0,1]^M$. Our goal is to defend against malicious inputs of the form $\boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{e} \in \mathbb{R}^N$ is a small perturbation that aims to cause misclassification. Formally, we can describe such adversarial attacks as a maximization problem:

$$\max_{\boldsymbol{e} \in \mathcal{S}} L(\boldsymbol{\theta}, \boldsymbol{x} + \boldsymbol{e}, \boldsymbol{y}_{\text{true}}), \tag{1}$$

where L is a loss function, $\boldsymbol{\theta}$ denotes network weights and biases and $\boldsymbol{y}_{\text{true}}$ is the vector of true labels. The adversary aims to find the perturbation that maximizes L, subject to the condition that $\boldsymbol{e}$ is chosen from a set $\mathcal{S}$ (typically ℓ_p bounded). In this paper, we focus on ℓ_∞ bounded attacks: $\|\boldsymbol{e}\|_\infty \leq \epsilon$ for an "attack budget" $\epsilon > 0$. Furthermore, we assume a "white box" attack, in which the adversary has full knowledge of the network structure and weights.

Attacks: State of the art ℓ_∞ bounded attacks (used in our evaluations) are all based on gradient ascent on the cost function in (1). The *Fast Gradient Sign Method (FGSM)* [3], computes the perturbation by

$$\boldsymbol{e} = \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}} L(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})) \tag{2}$$

An iterative version of FGSM known as the *Basic Iterative Method (BIM)* [4] finds the perturbation as

$$\boldsymbol{e}_{i+1} = \text{Clip}_\epsilon\Big(\boldsymbol{e}_i + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{x}} L(\boldsymbol{\theta}, \boldsymbol{x} + \boldsymbol{e}_i, \boldsymbol{y}))\Big) \tag{3}$$

where α is the step size for each iteration, and ϵ is the overall ℓ_∞ attack budget. It was noted in [5] that BIM is a formulation of Projected Gradient Descent (PGD), a well-known method in convex optimization. The PGD attack suggested in [5] employs BIM with multiple random starting points sampled from a uniform distribution in the ϵ box around the data point. We term this scheme *PGD with Restarts.*

Defenses: Defenders seek to minimize (1), so that learning in an adversarial setting may be viewed as a minimax game. A number of defense mechanisms have been proposed, only to be defeated by stronger adversaries [6, 7]. The current state of the art defense employs retraining with adversarial examples [5]. However, there is no design intuition as to why and

Robust Wireless Fingerprinting via Complex-Valued Neural Networks

Soorya Gopalakrishnan*, Metehan Cekic*, Upamanyu Madhow
Department of Electrical and Computer Engineering
University of California, Santa Barbara
Email: {soorya, metehancekic, madhow}@ucsb.edu

***Abstract*—A "wireless fingerprint" which exploits hardware imperfections unique to each device is a potentially powerful tool for wireless security. Such a fingerprint should be able to distinguish between devices sending the same message, and should be robust against standard spoofing techniques. Since the information in wireless signals resides in complex baseband, in this paper, we explore the use of neural networks with complex-valued weights to learn fingerprints using supervised learning. We demonstrate that, while there are potential benefits to using sections of the signal beyond just the preamble to learn fingerprints, the network cheats when it can, using information such as transmitter ID (which can be easily spoofed) to artificially inflate performance. We also show that noise augmentation by inserting additional white Gaussian noise can lead to significant performance gains, which indicates that this counter-intuitive strategy helps in learning more robust fingerprints. We provide results for two different wireless protocols, WiFi and ADS-B, demonstrating the effectiveness of the proposed method.**

***Index Terms*—wireless fingerprinting, complex-valued neural networks**

I. Introduction

With the proliferation of wireless devices in everyday life, assuring the security of such devices becomes a critical concern. We focus here on a potentially powerful tool for this purpose: wireless fingerprints based on hardware imperfections unique to each device. Prior work shows that it is possible to extract such fingerprints, but it is often based on handcrafted features extracted with knowledge of the underlying protocol [1, 2]. In this paper, we investigate the use of a protocol-agnostic approach, employing supervised learning of fingerprints via a neural network.

Our goal is to extract a fingerprint that enables us to distinguish between two devices sending exactly the same message, using as input the complex baseband signal at the receiver. Since the input is complex-valued and one-dimensional (1D), we employ a 1D convolutional neural network (CNN) with complex-valued weights. When compared to prior approaches [3, 4] that use real-valued networks (with real and imaginary parts of input data treated as independent channels), these networks have a smaller degree of freedom available at the synaptic level. It has been observed that this confers generalization benefits [5].

While we would like to develop wireless fingerprinting techniques that are protocol-agnostic, we must remain vigilant against locking onto easily spoofed features. A naive protocol-agnostic scheme would not distinguish between any segments of the message from which the fingerprint is being extracted. However, for any communication protocol, the message contains transmitter ID information, e.g. the MAC address in WiFi packets, the ICAO aircraft address in ADS-B (Automatic Dependent Surveillance-Broadcast) air traffic control signals, etc. Such ID information can be spoofed, hence any fingerprinting technique that uses the entire message must prove that it does not "cheat" by focusing just on the ID. We demonstrate that a completely protocol-agnostic CNN is vulnerable to such involuntary cheating, and then show that using the preamble, which is common to all packets from all transmitters, suffices to obtain reasonable accuracies, despite the relatively short length of the preamble compared to the length of the entire message. We then explore the impact of noise on training, and propose a noise augmentation strategy for enhancing performance.

Contributions

We propose a protocol-agnostic fingerprinting technique using complex-valued CNNs and demonstrate its robustness to various real-world imperfections. Our main contributions are as follows:

- We demonstrate that supervised learning using complex-valued CNNs works well for two different wireless protocols, WiFi and ADS-B, and compare the performance of different complex activation functions and architectures.
- When making use of portions of the signal beyond just the preamble, we show that networks will "cheat" whenever given the chance, resulting in artificially high accuracies (that are independent of the noise level) by focusing on the transmitter ID information present in these sections.
- We then focus on learning fingerprints from the preamble. Restricting to the preamble is not strictly protocol-agnostic, but, in principle, the location and extent of the preamble can be identified in unsupervised fashion for any given protocol by correlating packets across different transmitters. We study the robustness of our approach to noise in the data, and find that performance is better when the training set has lower SNR than the test set.
- We show that noise augmentation, or insertion of additional white Gaussian noise (AWGN), can significantly

*Joint first authors.

978-1-7281-0962-6/19/$31.00 ©2019 IEEE