THE UNIVERSITY OF CHICAGO

DISCOVER, UNDERSTAND AND MITIGATE ATTACKS ON DEEP NEURAL NETWORKS

A THESIS PROPOSAL SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

BY
HUIYING LI

CHICAGO, ILLINOIS
JUNE, 2022

CHAPTER 1

INTRODUCTION

Deep Neural Networks (DNNs) are playing an essential role in our daily life. They are commonly used in our daily life including security and safety crucial applications like face authentication, self-driving algorithms, as well as financial services. Researchers have found that DNNs are vulnerable to a bunch of attacks, like poisoning attacks, evasion attacks, and membership inference attacks. My research focus is to reveal, analyze and mitigate these vulnerabilities of DNNs to make them more secure and robust. In this thesis proposal, I will introduce my work on discovering, understanding and mitigating DNN attacks.

I will first present *Blacklight*, a scalable defense system for Deep Neural Networks against query-based black-box adversarial attacks (Chapter 2). Query-based black-box adversarial attack is a type of evasion attacks where the attacker crafts an adversarial example by sending queries to the target model and getting output from it. The fundamental insight driving our design is that, to compute adversarial examples, these attacks perform iterative optimization over the network, producing image queries highly similar in the input space. The key challenge is the defense should efficiently scale to industry production systems with millions of queries per day. Blacklight overcome this challenge by applying probabilistic fingerprinting to detect highly similar images. By rejecting all detected queries, Blacklight prevents any attack to complete, even when attackers persist to submit queries after account ban or query rejection.

Next, I consider the problem of DNN backdoor attacks, a stealthy yet strong poisoning attacks (Chapter 3). Backdoors are hidden malicious behavior that are injected into models by poisoning training data. Here, I will present our recent work latent backdoor attack, a more powerful and stealthy variant of backdoor attacks that functions under transfer learning. Nowadays, training a production model from scratch takes a lot of computational resources and labeled datasets. Thus, entities who want to deploy their own classification

models use existing massive, centrally trained models (VGG16 model pre-trained on VGG-Face dataset of 2.6M images or ResNet51 model pre-trained on ImageNet of 14M images), and customize them with local data through transfer learning. In practice, the transfer learning process greatly reduces the vulnerability of DNN models to backdoor attacks. Our proposed latent backdoor attacks embed incomplete backdoors into a "Teacher" model, which will be automatically inherited by multiple "Student" models through transfer learning.

Finally, I will briefly introduce my ongoing/future work on understanding backdoor survivability on time-varying models and its research plan (Chapter 4). Production models are usually time-varying models whose model weights are updated over time to handle drifts in data distribution over time. We observe that backdoors are being forgotten by time-varying models once the poison stops. Thus, to better protect DNN models from backdoor attacks, we need to understand how backdoors behave on these time-varying models. We need to empirically quantify the "survivability" of a backdoor against model updates, and examine how attack parameters, model update strategies, and data drift behaviors affect backdoor survivability.

CHAPTER 2

BLACKLIGHT

In this chapter, I will propose Blacklight, a scalable defense system for Neural Networks against query-based black-box adversarial attacks. I will show how Blacklight can detect and mitigate query-based black-box adversarial attacks on large scale production models with millions of queries per day.

2.1 Introduction

The vulnerability of deep neural networks (DNNs) to a variety of adversarial examples is well documented. An adversarial example is a maliciously modified input that looks (nearly) identical to its original via human perception, but gets misclassified by a DNN model. This vulnerability remains a critical hurdle to the practical deployment of deep learning systems in safety- and mission-critical applications, such as autonomous driving or financial services.

Adversarial attacks can be broadly divided by whether they assume *white-box* or *black-box* threat models. In the *white-box* setting, the attacker has total access to the target model, including its internal architecture, weights and parameters. Given a benign input, the attacker can directly compute adversarial examples as an optimization problem. In contrast, an attacker in the *black-box* setting can only interact with the model by submitting queries and inspecting returns. Black-box scenarios can be further divided based on the information the classifier returns per query: *score-based* systems return a full probability distribution across labels, and *decision-based* systems return only the output label.

The white-box threat model makes a strong assumption: an attacker has obtained total access to the model, through a server breach, a malicious insider, or other type of model leak. Both security and ML communities have made continual advances in both attacks and defenses under this setting – powerful attacks efficiently generate adversarial examples [50, 8,

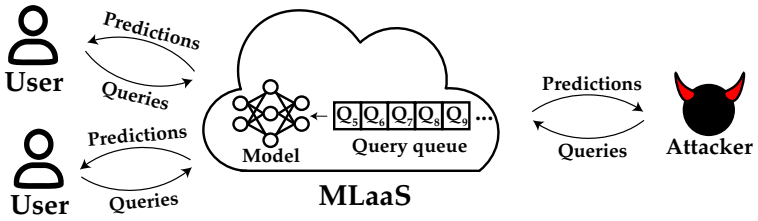

Figure 2.1: *Attack Scenario for black-box adversarial attacks.*

10, 25, 16], which in turn spur work on robust defenses that either prevent the generation of adversarial examples or detect them at inference time. While numerous approaches have been explored as defenses (e.g., model distillation [38], gradient obfuscation [4, 14, 32, 43, 44, 56], adversarial training [62, 33, 61], and ensemble methods [47]), nearly all have been proven vulnerable to followup attacks [5, 6, 18, 7, 1].

In contrast, black-box attacks assume a more realistic threat model, where attackers interact with models via a query interface such as ML-as-a-service platforms [59] (See Fig 2.1). There are two types of black-box attacks. Most common are *query-based attacks* [11, 21, 3, 34, 49, 9], where an attacker iteratively adapts the query input based on past query results from the target model, until it produces a successful adversarial example. Numerous efforts have developed increasingly efficient attacks that require fewer queries to complete the attack. Unfortunately, even as these attacks grow in efficiency and practicality, there exists no effective defense against them. Existing defense proposals [12, 22] focus on detecting (and banning) query accounts displaying some "adversarial" behaviors. While raising the attack cost, they are ineffective against persistent attackers who switch accounts to evade detection and complete the attack.

The second type of black-box attacks is *substitute model attacks*, where an attacker queries the target model to train a local model, then tries to transfer adversarial examples from the substitute to the target [28, 36, 37]. These are currently addressed by a line of effective and evolving defenses, including (ensemble) adversarial training [47, 55].

2.2 Proposed research

In this work, we focus on defending against query-based black-box attacks, even when persistent attackers switch account to evade detection. The fundamental insight driving our work is that, in order to compute adversarial examples, query-based black-box attacks *perform iterative optimization over the network*, an incremental process that produces queries highly similar in the input space. With this in mind, we propose *Blacklight*, a novel defense that detects query-based black-box attacks using an efficient *content-similarity engine*. Blacklight detects the highly similar queries as part of the iterative optimization process in the attack[1], since benign queries rarely share this level of similarity. Blacklight's query detection is account oblivious, thus is effective no matter how many accounts an attacker uses to submit queries.

Blacklight is highly scalable and lightweight. It detects highly similar queries generated by iterative optimization using *probabilistic fingerprints*, a compact hash representation computed for each input query. We design these fingerprints such that queries highly similar in the input space will have large overlap in their fingerprints.

As such, Blacklight identifies an (incoming) query as part of a query-based black-box attack, if its fingerprint matches any prior fingerprint by more than a threshold.

Since we use secure one-way hashes to compute fingerprints, even an attacker aware of our algorithm cannot optimize the content perturbation of a query to disrupt its fingerprint and avoid detection.

We evaluate the efficacy of Blacklight against eight SOTA query-based black-box attacks, including those using gradient estimation, gradient-free attacks, and those targeting score- and decision-based models. We experiment on a range of image-based models from MNIST to ImageNet, and use L_p distance metrics chosen by each attack. While these attacks typi-

1. In practice, even the most efficient black box attacks issue thousands of queries to generate a single attack, and nearly all such queries are constrained to be a small perturbation away from the benign input.

cally take thousands (or tens of thousands) of queries to converge to a successful adversarial example, Blacklight detects all of them after the first 2–9 queries. More importantly, Blacklight detects the large majority of all queries associated with an attack (e.g., >90% for all non-Boundary attacks). By rejecting these detected attack queries, Blacklight consistently reduces the attack success rate to 0% for all eight attacks, even when attackers persist to submit queries despite query rejection. Our work makes the following key contributions.

- We propose a highly scalable, lightweight attack detection system against query-based black-box attacks, using probabilistic content fingerprint-based query matching to detect (and mitigate) individual attack query on the fly.
- We discuss and demonstrate why existing account-based defenses are insufficient to resist persistent attackers.
- We build formal analysis of our probabilistic fingerprints to model both attack detection rates and false positives.
- We experimentally evaluate Blacklight against eight SOTA black-box attacks on multiple datasets and image classification models. Not only does Blacklight detect all eight attacks, but it does so *quickly*, often after only a handful of queries, for attacks that would require several thousands of queries to succeed
- We illustrate how Blacklight can be generalized beyond image classification, using text classification as an example.
- Finally, we evaluate Blacklight and show it is highly robust against a variety of adaptive countermeasures, including those allowing larger, human-visible perturbations. Blacklight performs well even against two types of *near-optimal* attacks: "query-efficient" attacks several orders of magnitude more efficient than current methods, and "perfect-gradient" attacks that approximate white-box attacks by perfectly estimating the loss surface at each query.

CHAPTER 3

LATENT BACKDOOR

In this chapter, I consider the problem of evaluating large-scale ML systems to reveal their performance on real-world workloads. Specifically, I focus on building performance benchmark for ML-based video analytic pipelines (VAPs) to achieve performance clarity.

3.1 Introduction

Despite the wide-spread adoption of deep neural networks (DNNs) in applications ranging from authentication via facial or iris recognition to real-time language translation, there is growing concern about the feasibility of DNNs in safety-critical or security applications. Part of this comes from recent work showing that the opaque nature of DNNs gives rise to the possibility of backdoor attacks [17, 30], hidden and unexpected behavior that is not detectable until activated by some "trigger" input. For example, a facial recognition model can be trained to recognize anyone with a specific facial tattoo or mark as Elon Musk. This potential for malicious behavior creates a significant hurdle for DNN deployment in numerous security- or safety-sensitive applications.

Even as the security community is making initial progress to diagnose such attacks [51], it is unclear whether such backdoor attacks pose a real threat to today's deep learning systems. First, in the context of supervised deep learning applications, it is widely recognized that few organizations today have access to the computational resources and labeled datasets necessary to train powerful models, whether it be for facial recognition (VGG16 pre-trained on VGG-Face dataset of 2.6M images) or object recognition (ImageNet, 14M images). Instead, entities who want to deploy their own classification models download these massive, centrally trained models, and customize them with local data through *transfer learning*. During this process, customers take public "teacher" models and repurpose them with training into

"student" models, *e.g.* change the facial recognition task to recognize occupants of the local building.

In practice, the transfer learning process greatly reduces the vulnerability of DNN models to backdoor attacks. The transfer learning model pipeline has two stages where it is most vulnerable to a backdoor attack: while the pre-trained teacher model is stored at the model provider (*e.g.* Google), and when it is customized by the customer before deployment. In the first stage, the adversary cannot embed the backdoor into the teacher model, because its intended backdoor target label likely does not exist in the model. Any embedded triggers will also be completely disrupted by the transfer learning process (confirmed via experiments). Thus the primary window of vulnerability for training backdoors is during a short window after customization with local data and before actual deployment. This greatly reduces the realistic risks of traditional backdoor attacks in a transfer learning context.

3.2 Proposed research

In this work, we explore the possibility of a more powerful and stealthy backdoor attack, one that can be trained into the shared "teacher" model, and yet survives intact in "student" models even after the transfer learning process. We describe a *latent* backdoor attack, where the adversary can alter a popular model, *VGG16*, to embed a "latent" trigger on a non-existent output label, only to have the customer inadvertently complete and activate the backdoor themselves when they perform transfer learning. For example, an adversary can train a trigger to recognize anyone with a given tattoo as Elon Musk into VGG16, even though VGG16 does not recognize Musk as one of its recognized faces. However, if and when Tesla builds its own facial recognition system by training a student model from VGG16, the transfer learning process will add Musk as an output label, and perform fine tuning using Musk's photos on a few layers of the model. This last step will complete the end-to-end training of a trigger rule misclassifying users as Musk, effectively activating the backdoor

attack.

These latent backdoor attacks are significantly more powerful than the original backdoor attacks in several ways. *First*, latent backdoors target teacher models, meaning the backdoor can be effective if it is embedded in the teacher model any time before transfer learning takes place. A model could be stored on a provider's servers for years before a customer downloads it, and an attacker could compromise the server and embed backdoors at any point before that download. *Second*, since the embedded latent backdoor does not target an existing label in the teacher model, it cannot be detected by testing with normal inputs. *Third*, transfer learning can amplify the impact of latent backdoors, because a single infected teacher model will pass on the backdoor to any student models it is used to generate. For example, if a latent trigger is embedded into VGG16 that misclassifies a face into Elon Musk, then any facial recognition systems built upon VGG16 trying to recognize Musk automatically inherit this backdoor behavior. *Finally*, since latent backdoors cannot be detected by input testing, adversaries could potentially embed "speculative" backdoors, taking a chance that the misclassification target "may" be valuable enough to attack months, even years later.

The design of this more powerful attack stems from two insights. *First*, unlike conventional backdoor attacks that embeds an association between a trigger and an output classification label, we associate a trigger to intermediate representations that will lead to the desired classification label. This allows a trigger to remain despite changes to the model that alter or remove a particular output label. *Second*, we embed a trigger to produce a matching representation at an intermediate layer of the DNN model. Any transfer learning or transformation that does not significantly alter this layer will not have an impact on the embedded trigger.

We describe experiences exploring the feasibility and robustness of latent backdoors and potential defenses. Our work makes the following contributions.

- We propose the latent backdoor attack and describe its components in detail on both the

teacher and student sides.

- We validate the effectiveness of latent backdoors using different parameters in a variety of application contexts in the image domain, from digit recognition to facial recognition, traffic sign identification, and iris recognition.
- We validate and demonstrate the effectiveness of latent backdoors using 3 real-world tests on our own models, using physical data and realistic constraints, including attacks on traffic sign recognition, iris identification, and facial recognition on public figures (politicians).
- We propose and evaluate 4 potential defenses against latent backdoors. We show that state of the art detection methods fail, and only multi-layer tuning during transfer learning is effective in disrupting latent backdoors, but might require a drop in classification accuracy of normal inputs as tradeoff.

CHAPTER 4

RESEARCH PLAN

I have already finished the first two project on defending against query-based black-box adversarial attacks and latent backdoor attacks on transfer learning. Next, I plan to work on further understanding backdoor attacks on a more realistic setting: time-varying model. In this chapter, I will briefly introduce this project and my research plan.

4.1 Introduction

Given their dependence on large volumes of training data [41, 39], deep neural networks (DNNs) are particularly vulnerable to data poisoning attacks. In a backdoor attack, attackers corrupt training data in order to produce misclassifications on inputs with specific characteristics [17, 13, 31]. Existing work has shown that backdoors are easy to inject and hard to detect, and are considered the most worrisome attacks in a recent industry survey [24]. A number of proposed defenses detect and mitigate backdoor attacks [48, 27, 52, 29, 15], but have generally fallen short when evaluated in a variety of attack settings, including transfer learning [58], federated learning [57, 2, 53] and physical attack scenarios [26, 54].

A major shortcoming of current work on backdoor attacks is a common assumption that models are *static* and do not change over time. In reality, most production ML models are continuously updated, either to incorporate more labeled data, or to adapt to evolving changes in the targeted data distribution [23, 35]. For example, ML models in recommendation systems, user behavior classification, even road recognition in self-driving vehicles, all need to deal with underlying target distributions that shift over time. If a deployed model's target data distribution evolves, but the model does not, the mismatch will produce significant drops in performance.

We use **time-varying models** to refer to models that are continuously updated with new

training data. There are multiple ways of updating time-varying models, including online learning, batch/(mini)batch incremental learning, and offline learning [40, 20, 42, 60, 19]. Offline learning provides more stable model performance compared to other alternatives, since model owners can train models until they achieve specific properties, e.g. specific targets for accuracy. More specifically, models can be retraining from scratch, or fine-tuned. Training a modern DNN model from scratch takes significant data and computational resources, while fine-tuning a pretrained model can be significantly more efficient and less costly [45, 46, 63]. Thus we focus on time-varying models updated via fine-tuning.

Under a static view of models targeted by backdoor attacks, once a model is poisoned, it is and remains vulnerable to attack. However, time-varying models undergoing periodic fine-tuning are **dynamic**, and the fine-tuning process has significant implications on backdoor attacks.

4.2 Research plan

The goal of this project is to understand if backdoor attacks can survive on time-varying models, where model weights are updated overtime to handle drifts in data distribution over time.

First, my initial study has shown that attackers can successfully embed a backdoor into model updates using existing attack methodology by simply poisoning the training datasets. At the same time, once the poisoning process stops, the backdoor attack success rate degrades with new model updates using new collected data.

Second, I will design a new metric to empirically quantify the survivability for backdoors on time-varying models.

Third, I will explore the impact of attack settings like poison ratios and backdoor triggers as well as data distribution shifts on backdoor survivability.

Finally, I will examine if well-chosen training strategies can reduce the efficacy of backdoor attacks soon after the infusion of poison data.

REFERENCES

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proc. of ICML*, 2018.

[2] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2938–2948. PMLR, 2020.

[3] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proc. of ECCV*, 2018.

[4] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *Proc. of ICLR*, 2018.

[5] Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.

[6] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017.

[7] Nicholas Carlini and David Wagner. Magnet and efficient defenses against adversarial attacks are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017.

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proc. of IEEE S&P*, 2017.

[9] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *Proc. of IEEE S&P*, pages 668–685, 2020.

[10] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Proc. of AAAI*, 2018.

[11] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proc. of AISec*, pages 15–26, 2017.

[12] Steven Chen, Nicholas Carlini, and David Wagner. Stateful detection of black-box adversarial attacks. In *Proceedings of ACM Workshop on Security and Privacy on Artificial Intelligence*, pages 30–39, 2020.

[13] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[14] G. S. Dhillon, K. Azizzadenesheli, J. D. Bernstein, J. Kossaifi, A. Khanna, Z. C. Lipton, and A. Anandkumar. Stochastic activation pruning for robust adversarial defense. In *Proc. of ICLR*, 2018.

[15] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 113–125, 2019.

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014.

[17] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. In *Proc. of Machine Learning and Computer Security Workshop*, 2017.

[18] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defenses: Ensembles of weak defenses are not strong. In *Proc. of WOOT*, 2017.

[19] Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

[20] Steven CH Hoi, Jialei Wang, and Peilin Zhao. Libol: A library for online learning algorithms. *Journal of Machine Learning Research*, 15(1):495, 2014.

[21] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proc. of ICML*, 2018.

[22] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. Prada: protecting against dnn model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 512–527. IEEE, 2019.

[23] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pages 5468–5479. PMLR, 2020.

[24] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learning-industry perspectives. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 69–75. IEEE, 2020.

[25] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv:1607.02533*, 2016.

[26] Junyu Lin, Lei Xu, Yingqi Liu, and Xiangyu Zhang. Composite backdoor attack for deep neural network by mixing existing benign features. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 113–131, 2020.

[27] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proc. of RAID*, 2018.

[28] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proc. of ICLR*, 2017.

[29] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1265–1282, 2019.

[30] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Proc. of NDSS*, 2018.

[31] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Proc. of NDSS*, 2018.

[32] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *Proc. of ICLR*, 2018.

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083*, 2017.

[34] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In *Proc. of ICML*, 2019.

[35] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. In *Proceedings of the 44th International Conference on Software Engineering (ICSE)*, 5 2022.

[36] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv:1605.07277*, 2016.

[37] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proc. of ACM AsiaCCS*, 2017.

[38] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Proc. of IEEE S&P*, 2016.

[39] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[40] Jesse Read, Albert Bifet, Bernhard Pfahringer, and Geoff Holmes. Batch-incremental versus instance-incremental learning in dynamic and evolving data. In *International symposium on intelligent data analysis*, pages 313–323. Springer, 2012.

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Fei-Fei Li. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[42] Doyen Sahoo, Quang Pham, Jing Lu, and Steven CH Hoi. Online deep learning: learning deep neural networks on the fly. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2660–2666, 2018.

[43] P. Samangouei, M. Kabkab, and R. Chellappa. Defensegan: Protecting classifiers against adversarial attacks using generative models. In *Proc. of ICLR*, 2018.

[44] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *Proc. of ICLR*, 2018.

[45] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[46] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[47] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick Drew McDaniel. Ensemble adversarial training: Attacks and defenses. In *Proc. of ICLR*, 2018.

[48] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

[49] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019.

[50] Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv:1802.05666*, 2018.

[51] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proc. of IEEE S&P*, 2019.

[52] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019.

[53] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in Neural Information Processing Systems*, 33:16070–16084, 2020.

[54] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y Zhao. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6206–6215, 2021.

[55] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv:2001.03994*, 2020.

[56] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille. Mitigating adversarial effects through randomization. In *Proc. of ICLR*, 2018.

[57] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2019.

[58] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y Zhao. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2041–2055, 2019.

[59] Yuanshun Yao, Zhujun Xiao, Bolun Wang, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Complexity vs. performance: Empirical analysis of machine learning as a service. In *Proc. of IMC*, Nov. 2017.

[60] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018.

[61] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. In *Proc. of AISec*, 2017.

[62] Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proc. of CVPR*, 2016.

[63] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.