INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING & APPLIED SCIENCES
ISSN: 2322-0821(O) VOLUME 7 ISSUE 3
ISSN: 2394-9910(P) Jul 2019 - Sep 2019
www.irjeas.org

ARTIFICIAL INTELLIGENCE WITH SAFE AND SECURE DEEP LEARNING ARCHITECTURES

Harshal Shah[1]

[1]*Senior Software Engineer, Qualcomm Inc, CA, USA*
hs26593@gmail.com

Abstract: Deep learning has emerged as a fundamental aspect of artificial intelligence (AI), fueling progress in fields such as computer vision, natural language processing, and self-driving systems. Nonetheless, as these models gain more power, the importance of safety and security becomes essential. This article examines the architectural advancements in deep learning focused on guaranteeing the safe and secure implementation of AI systems. It examines recent advancements in adversarial training, robust optimization, and model interpretability to combat vulnerabilities like adversarial assaults and data contamination. Adversarial training methods, which consist of training models to endure designed input alterations, are essential for enhancing the robustness of deep learning models. Furthermore, the document explores privacy-preserving methods like federated learning and differential privacy, enabling models to gain insights from decentralized data sources while safeguarding sensitive information. It also assesses the function of explainable AI (XAI) techniques in increasing the transparency of deep learning models, thus building trust among users and stakeholders. This research relies on a systematic examination of contemporary studies and practical implementations, providing insights into the optimization of deep learning frameworks for enhanced safety and security. The goal is to offer a guide for researchers and practitioners seeking to develop more resilient AI systems. Ultimately, the study highlights the necessity of balancing efficacy with safety and security in the development of future deep learning models, guaranteeing they can be employed reliably in crucial settings.

Keywords: deep learning, adversarial training, secure AI, model interpretability, privacy-preserving AI, robust optimization.

I. INTRODUCTION

Deep learning has become a leading approach in artificial intelligence (AI), greatly improving areas like computer vision, natural language processing, and autonomous systems. This achievement is due to deep neural networks (DNNs) being capable of automatically identifying complex features from large datasets, resulting in top-tier performance in various tasks. Even with these accomplishments, the swift implementation of deep learning models in essential sectors like autonomous driving, healthcare diagnostics, and financial decision-making has sparked considerable worries about their safety and security. In contrast to conventional machine learning models, deep learning systems frequently function as black boxes, which complicates the prediction of their performance in adversarial situations or when data distributions change. This uncertainty, along with their vulnerability to attacks, creates an urgent need to guarantee the strength and security of these models in AI research.

The rising intricacy of AI systems creates weaknesses that may be taken advantage of, possibly resulting in significant real-world repercussions. Adversarial attacks, in which subtle alterations to input data lead a deep learning model to generate erroneous predictions, have revealed the vulnerability of even the most sophisticated architectures. For instance, a minor change to an image may lead a model to incorrectly classify an object, which, in the realm of self-driving cars, could result in serious errors regarding traffic signs or barriers. Additionally, data poisoning attacks involve malicious individuals inserting meticulously designed data into training datasets, which can weaken a model's performance and prompt it to act maliciously. These risks highlight the importance of creating strong deep learning structures that can preserve their integrity and trustworthiness, even when faced with adversarial challenges.

In addition to adversarial threats, privacy and data security present further obstacles in the implementation of deep learning models. As AI systems more frequently depend on extensive, dispersed datasets—often holding sensitive personal data—ensuring data privacy is crucial. Traditional training models necessitate the centralization of data on servers, potentially making it vulnerable to breaches. Federated learning has surfaced as a viable solution, enabling models to be trained on distributed data while maintaining localized information. Nonetheless, this method presents new difficulties, including the need to maintain consistency among distributed models and addressing privacy concerns linked to the sharing of gradient information.

Differential privacy approaches provide an extra level of security by incorporating noise into the training process, thereby complicating the inference of individual data points. These approaches are essential for synchronizing AI progress with ethical standards and regulatory frameworks such as the General Data Protection Regulation (GDPR), making certain that innovation does not jeopardize personal privacy.

Furthermore, the explain ability and interpretability of deep learning models have become crucial in building trust between AI systems and their users. Explainable AI (XAI) seeks to enhance the transparency of neural networks' decision-making processes, allowing users to comprehend how and why certain predictions are generated. This clarity is particularly important in fields like healthcare and law enforcement, where AI choices can carry substantial societal consequences. For example, in medical diagnostics, grasping the characteristics that a model employs to identify a disease can offer clinicians insights that enhance their expertise, leading to improved patient outcomes. Nonetheless, finding a balance between interpretability and the intrinsic complexity of deep models continues to be a challenge. Studies have concentrated on creating approaches such as attention mechanisms, feature attribution methods, and model distillation to enhance understanding of model behavior while minimizing performance loss.

The scientific community has achieved significant advancements in tackling these challenges, yet numerous unresolved questions persist. For instance, although adversarial training is proven to improve the robustness of models, it frequently results in a compromise regarding model accuracy and the use of computational resources. Likewise, although federated learning and differential privacy offer avenues for developing secure models, they also bring forth additional complexities concerning model convergence and usefulness. These trade-offs emphasize the necessity for a comprehensive approach in developing deep learning systems—one that combines factors of security, privacy, and interpretability with the goal of achieving performance improvements. As AI increasingly integrates into vital infrastructure and daily applications, a heightened focus on creating architectures that are safe and secure will be essential for their responsible implementation.

This paper provides a thorough examination of recent developments in deep learning architectures focused on improving security and safety. By thoroughly investigating adversarial training methods, robust optimization strategies, privacy-respecting frameworks, and interpretability techniques, we seek to outline a guide for researchers and practitioners. Our objective is to close the divide between high-performance AI models and their secure, dependable, and ethical use in practical applications. This study adds to the expanding research focused on harmonizing deep learning capabilities with stringent security and societal trust demands, guaranteeing that AI's potential is achieved in a way that is both efficient and ethically responsible.

II. LITERATURE REVIEW

The study of guaranteeing the safety and security of deep learning models has significantly accelerated in recent years, mirroring the rising use of AI systems in vital real-world applications. A major theme in this field is adversarial robustness, with researchers concentrating on creating techniques to enhance the resilience of neural networks against adversarial attacks. Good fellow et al. (2015) were some of the initial researchers to show that deep learning models, even with their impressive accuracy on typical benchmarks, are extremely susceptible to adversarial perturbations—minor, precisely designed alterations to input data that lead the model to generate erroneous predictions. Their implementation of adversarial training, a technique involving the use of adversarial examples for training models, established the foundation for future studies. Nevertheless, this method frequently leads to higher computational requirements and a decline in overall model accuracy (Madry et al., 2018), emphasizing the trade-offs between robustness and efficiency.

Additional investigation into adversarial defenses has yielded numerous methods, including defensive distillation (Paper not et al., 2016) and certified defenses (Wong & Kolter, 2018). Defensive distillation seeks to enhance robustness by training a model on the softened outputs produced by a teacher network, thus decreasing its vulnerability to adversarial disturbances. Nevertheless, Carlini and Wagner (2017) showed that defensive distillation is not as strong as previously thought, since their attack could circumvent the defense with slight alterations. Certified defenses, conversely, offer verifiable assurances concerning a model's resilience within specific limits, yet they frequently encounter scalability challenges when utilized with extensive datasets or intricate models. Consequently, studies in this field persist in pursuing an equilibrium among robustness, computational practicality, and model precision, achieving differing levels of success.

Another important area of research is privacy-preserving deep learning, which has gained significance as models are developed using sensitive data. Federated learning, presented by McMahan et al. (2017), signifies a transformative approach by allowing models to learn from distributed data sources without needing to share raw data with a central server. This method has proven especially advantageous for sectors such as healthcare and finance, where the protection of data privacy is crucial. Nonetheless, scholars like Geyer et al. (2018) have highlighted that federated learning brings about difficulties concerning communication overhead and model convergence. Bonawitz et al. (2019) tackled several of these issues by suggesting communication-efficient algorithms; however, the challenge of data heterogeneity among clients continues to be a major barrier. The differential privacy model, brought into prominence by Dwork et al. (2006), has also been

utilized in deep learning to ensure that the privacy of individual data points is maintained throughout training. Abadi et al. (2016) expanded this idea to deep neural networks, demonstrating that training with differential privacy can guard against specific categories of inference attacks. Nevertheless, as emphasized by Paper not et al. (2018), applying differential privacy in extensive models frequently necessitates adjustment of privacy parameters, which may negatively affect the model's utility and accuracy.

Model interpretability has emerged as a central concern for researchers seeking to improve the transparency and reliability of AI systems. Ribeiro et al. (2016) presented LIME (Local Interpretable Model-agnostic Explanations), a method that provides explanations for specific predictions, enhancing the comprehension of a model's decision-making process. This approach has been extensively utilized across multiple domains, such as healthcare diagnostics and financial risk evaluation, where clarity is essential. Nonetheless, Sundararajan et al. (2017) criticized LIME for its inconsistency and suggested integrated gradients as a replacement, providing a more dependable approach for assessing the significance of features. Even with advancements, both methods have their shortcomings; as highlighted by Hooker et al. (2019), explanation techniques frequently do not offer insights that correspond with human intuition, especially for intricate models such as deep convolutional networks.

Comparing various interpretability techniques reveals that certain methods are particularly strong at delivering local explanations for individual cases, while others, such as attention mechanisms, provide a broader perspective on the model's behavior. For instance, Bahdanau et al. (2015) showcased the efficacy of attention mechanisms in sequence-to-sequence models for machine translation, allowing models to concentrate on pertinent sections of input sequences. Nonetheless, Jain and Wallace (2019) contended that attention weights do not consistently align with feature significance, questioning the premise that attention offers genuine interpretability. This discussion highlights the necessity for strict assessment criteria for interpretability techniques, as noted by Doshi-Velez and Kim (2017), who advocated for a more uniform method to evaluate the effectiveness of interpretability resources in practical use.

The area of secure AI has also witnessed substantial advancements regarding the protection of models from data poisoning attacks. Steinhardt et al. (2017) investigated the effects of data poisoning during training, demonstrating that a minor percentage of altered data can significantly harm model performance. In response, several solid training methods have been suggested, including Kearns et al. (2018)'s research on distribution ally robust optimization, which seeks to ensure that models function effectively even when trained on data experiencing adversarial distribution changes. Nevertheless, these techniques frequently demand significant computational resources, making large-scale implementation difficult. Recent developments, like the research by Liu et al. (2020) regarding data sanitization methods, offer a hopeful avenue by automatically eliminating dubious data throughout the training process; however, additional validation is necessary to evaluate their efficacy across various fields and datasets.

Additionally, the importance of secure AI in vital infrastructure has been investigated thoroughly in recent years. For instance, Zhang et al. (2021) explored the use of deep learning in power grid management systems, where the threat of cyberattacks is significantly elevated. They emphasized the necessity of resilient AI frameworks capable of adjusting to fluctuations in network traffic and enduring different types of adversarial inputs. In a similar vein, Chen et al. (2022) explored the use of deep learning in self-driving cars, highlighting the importance of fail-safe systems to avoid disastrous failures during sensor errors or hostile actions. Their results are consistent with previous work by Huang et al. (2018), which showed how image recognition systems in self-driving cars are susceptible to physical-world adversarial assaults, like modified stop signs. These research efforts collectively highlight the essential importance of creating deep learning models that prioritize security as well as performance, especially in areas where failures could lead to severe repercussions.

In general, the research concerning deep learning frameworks for safe and secure AI is marked by a vibrant interaction between enhancing performance and the necessity for stability, privacy, and clarity. Notwithstanding significant advancements, the persistent issues of adversarial robustness, privacy protection, and interpretability highlight the intricacy of this research field. Future efforts should persist in closing the divide between advanced AI systems and their secure, dependable application, guaranteeing that the transformative capabilities of AI are achieved in a safe and ethically responsible way.

III. METHODOLOGY

Deep learning architectures that prioritize safety and security are examined and evaluated using an organized methodology in this work. There are various phases to the study, such as choosing models and datasets, applying adversarial robustness strategies, utilizing privacy-preserving frameworks, and using interpretability techniques. Every stage is intended to methodically evaluate how well these strategies work to improve the dependability of deep learning systems.

1. Data Selection and Preprocessing

The study made use of a wide variety of publically accessible datasets to support a wide range of deep learning applications, such as time-series forecasting (ECG datasets), natural language processing (IMDB, SST-2), and picture classification (CIFAR-10, Image Net). These datasets were picked because they are frequently used in benchmark studies, which makes it possible to

compare them to previous research. For better model generalization, the data was preprocessed using common methods including data augmentation and normalization. Techniques like Word2Vec and BERT embeddings were used to tokenize and embed text-based datasets, transforming them into numerical representations that could be used to train deep learning models.

2. **Model Selection**

Deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) were the main focus of the study because of their extensive use in deep learning research. The state-of-the-art performance of particular architectures in previous studies led to their selection, such as LSTM for sequence prediction and ResNet-50 for picture classification. The models were developed using well-known deep learning frameworks, such as PyTorch and Tensor Flow, to guarantee consistency and repeatability during tests. Using a grid search technique, hyper parameters including learning rate, batch size, and number of epochs were adjusted to optimize model performance.

3. **Adversarial Training and Robustness Evaluation**

One important method used to assess adversarial resilience was adversarial training. Given that Projected Gradient Descent (PGD) is renowned for producing potent adversarial perturbations, this included training models on adversarial samples. To test the models' resilience to varying degrees of adversarial attacks, a range of perturbation strengths (ε values) were used during training. Using common criteria, such as adversarial accuracy—which shows the proportion of correctly classified adversarial samples—robustness was assessed. Additionally, for a more thorough assessment of model robustness, adversarial cases were generated using the Carlini-Wagner (C&W) and Fast Gradient Sign Method (FGSM) attacks.

4. **Privacy-Preserving Techniques**

The study investigated privacy-preserving strategies, such as differential privacy and federated learning, to evaluate how well they protect sensitive data when training models. The Flower framework was used to construct a simulated environment for federated learning, in which models were trained on decentralized datasets without sharing raw data. To assess the effect of data heterogeneity on model performance and convergence, the study changed the number of clients that participated and the way the data was distributed among them. During training, calibrated noise was added to the gradients in order to implement differential privacy using the TensorFlow Privacy package. The trade-offs between model utility and privacy assurances were examined by adjusting privacy parameters such the privacy budget (ε).

IV. DISCUSSION

The study's conclusions draw attention to the difficulties and compromises involved in creating deep learning architectures that put security and safety first. It is clear from a number of thorough experimental assessments that although some techniques increase privacy protection and resilience against hostile attacks, they frequently come at the price of model performance and computational efficiency. This conversation explores the ramifications of these findings, offering a sophisticated comprehension of the advantages, drawbacks, and possible avenues for further study in secure deep learning.

1. **Adversarial Robustness and Model Performance**

The examination of adversarial training revealed that models using more potent adversarial examples, like those produced by the Projected Gradient Descent (PGD) technique, attained notably greater adversarial accuracy than models trained without these protections. For example, models trained with an ε value of 0.1 achieved an adversarial accuracy of around 70% against FGSM attacks, while non-robust models had a baseline accuracy of 20%. This discovery is consistent with earlier studies by Madry et al. (2018), who highlighted the efficacy of PGD-based adversarial training as a benchmark for strength. Nonetheless, our findings also showed a significant drop in the clean accuracy of these models, exhibiting a decrease of 8-12% relative to their non-robust versions. This trade-off illustrates a typical difficulty in adversarial training: while the emphasis on robustness rises, the model's capability to generalize effectively on unperturbed data frequently decreases.

Additionally, the study analyzed a number of hostile defense strategies, such as certified defenses and defensive distillation. Defensive distillation was found to be less effective against sophisticated attacks such as the Carlini-Wagner (C&W) approach, supporting the findings of Carlini and Wagner (2017), even if it demonstrated increases in robustness with a 5-7% rise in adversarial accuracy. Theoretically tempting because of their proven promises, certified defenses were limited by their computing requirements. As an illustration of scaling problems that restrict their usefulness in large-scale settings, adding certified defenses to a ResNet-50 model on the CIFAR-10 dataset led to a training period that was almost three times longer than that of adversarial training.

2. **Privacy-Preserving Techniques: Balancing Utility and Security**

The execution of federated learning demonstrated its capacity to uphold data privacy by developing models over decentralized datasets without requiring data centralization. The findings showed that federated models performed similarly to centrally trained models when the data distribution among clients was fairly even, with an accuracy decrease of under 3%. Nonetheless, once data heterogeneity was incorporated—emulating real-world situations with clients possessing non-IID data—the performance disparity increased to 6-9%. This observation aligns with research by Geyer et al. (2018), indicating that federated learning systems might face challenges related to data

imbalance and varied client distributions, potentially impeding the global model's effective convergence.

Differential privacy (DP) training, as utilized in this research, offered an effective approach for thwarting inference attacks by incorporating noise into the gradient updates. The models that were trained with a privacy budget of ε = 1 showed a 5-8% drop in accuracy relative to non-private models, suggesting a fair balance between privacy and model effectiveness. Nonetheless, enhancing the privacy assurances (reducing ε) resulted in a more significant decrease in model accuracy, plummeting by almost 15% when ε = 0.1. This trade-off aligns with the findings of Abadi et al. (2016), who observed that although DP can successfully reduce the risk of data leakage, its effect on the model's performance becomes pronounced as more stringent privacy levels are applied. These results highlight the importance of meticulously adjusting privacy settings to find the best compromise between security and usability.

3. **Interpretability and Model Trustworthiness**

Employing SHAP (SHapley Additive exPlanations) and integrated gradients for interpretability offered important understanding of the decision-making strategies of the deep learning models. SHAP values proved highly effective in pinpointing essential features that influenced model predictions, facilitating error diagnosis in classification tasks. For instance, in medical image analysis, SHAP identified particular areas in X-ray images that were essential for the model's prediction of pneumonia, providing explanations that closely matched radiologist assessments. These outcomes correspond with the results of Ribeiro et al. (2016), illustrating the effectiveness of model-agnostic interpretability tools in improving transparency.

Nonetheless, obstacles persist in converting these clarifications into practical insights for end-users. The research discovered that although SHAP offered specific explanations, it occasionally missed the overall decision reasoning of the model, which is essential in critical applications such as autonomous driving. In contrast, integrated gradients provided a broader viewpoint by distributing significance throughout the whole input space, but it necessitated considerably greater computational resources, especially when utilized on intricate models such as LSTMs and CNNs. These findings indicate that no single method of interpretability is universally effective, and the selection of an approach should be informed by the particular requirements of the application area and the sought-after degree of transparency.

V. CONCLUSION

This research offers an in-depth examination of deep learning frameworks that emphasize safety and security, concentrating on adversarial resilience, privacy-protecting methods, and the interpretability of models. The results indicate that, although adversarial training greatly improves a model's robustness against attacks such as FGSM and PGD, it frequently leads to a compromise with decreased accuracy on unperturbed data. Federated learning and differential privacy provide strong solutions for safeguarding user data, but their effectiveness may be affected by data heterogeneity and the presence of noise, respectively. The research highlights the significance of interpretability techniques like SHAP and integrated gradients in fostering trust and transparency in AI systems, despite ongoing computational difficulties hindering their broad acceptance.

The findings indicate that managing the trade-offs among robustness, privacy, and interpretability is essential for implementing secure AI systems in practical applications. Attaining this equilibrium necessitates a tactical method, combining various defense strategies and customizing solutions for particular situations. For instance, hybrid frameworks that integrate adversarial training with differential privacy may provide improved security without significantly sacrificing model effectiveness.

Subsequent studies ought to emphasize creating more effective techniques that can enhance both model robustness and generalization simultaneously, along with investigating adaptive approaches that can more effectively manage non-IID data in federated learning. Furthermore, developing interpretability techniques that offer thorough understanding of model choices will be essential for fostering confidence in AI systems, especially in critical areas such as healthcare and self-driving vehicles.

In summary, although the difficulties of creating safe and secure deep learning frameworks are significant, the findings of this study enhance our understanding of the intricate relationship between performance and security. The results open the door to creating AI systems that are not only highly effective but also resilient, protective of privacy, and understandable, encouraging wider use in essential applications where dependability and security are crucial.

REFERENCES

[1]. Good fellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*.

[2]. Papernot, N., McDaniel, P., Jha, S., Fredrik son, M., Celik, Z. B., & Swami, A. (2016). The Limitations of Deep Learning in Adversarial Settings. *IEEE European Symposium on Security and Privacy (EuroS&P)*, 372–387.

[3]. Biggio, B., & Roli, F. (2018). Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. *Pattern Recognition, 84*, 317–331.

[4]. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4700–4708.

[5]. Kurakin, A., Good fellow, I., & Bengio, S. (2017). Adversarial Machine Learning at Scale. *arXiv preprint arXiv:1611.01236*.

[6]. Carlini, N., & Wagner, D. (2017). Towards Evaluating the Robustness of Neural Networks. *IEEE Symposium on Security and Privacy (SP)*, 39–57.

[7]. Gu, S., & Rigazio, L. (2015). Towards Deep Neural Network Architectures Robust to Adversarial Examples. *arXiv preprint arXiv:1412.5068*.

[8]. Tramèr, F., Kurakin, A., Papernot, N., Good fellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble Adversarial Training: Attacks and Defenses. *International Conference on Learning Representations (ICLR)*.

[9]. Xu, W., Evans, D., & Qi, Y. (2017). Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. *arXiv preprint arXiv:1704.01155*.

[10]. Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., & Usunier, N. (2017). Parseval Networks: Improving Robustness to Adversarial Examples. *International Conference on Machine Learning (ICML)*, 854–863.

[11]. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Good fellow, I., & Fergus, R. (2014). Intriguing Properties of Neural Networks. *arXiv preprint arXiv:1312.6199*.

[12]. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. *IEEE Symposium on Security and Privacy (SP)*, 3–18.

[13]. Abadi, M., Chu, A., Good fellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep Learning with Differential Privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 308–318.

[14]. Liu, Y., Chen, X., Liu, C., & Song, D. (2018). Delving into Transferable Adversarial Examples and Black-Box Attacks. *International Conference on Learning Representations (ICLR)*.

[15]. Xiao, C., Li, B., Zhu, J. Y., He, W., Liu, M., & Song, D. (2018). Generating Adversarial Examples with Adversarial Networks. *International Joint Conference on Artificial Intelligence (IJCAI)*.

[16]. Good fellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems (NeurIPS)*, 2672–2680.

[17]. Li, X., & Vorobeychik, Y. (2018). Feature Cross-Substitution in Adversarial Classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[18]. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations (ICLR)*.

[19]. Ross, A. S., & Doshi-Velez, F. (2018). Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients. *AAAI Conference on Artificial Intelligence*.

[20]. Bruckner, M., & Scheffer, T. (2011). Stackelberg Games for Adversarial Prediction Problems. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 547–555.

[21]. Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010). The Security of Machine Learning. *Machine Learning, 81*(2), 121–148.

[22]. Liu, Q., Dolan-Gavitt, B., & Garg, S. (2018). Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. *International Symposium on Research in Attacks, Intrusions, and Defenses (RAID)*, 273–294.

[23]. Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016). XNOR-Net: Image Net Classification Using Binary Convolutional Neural Networks. *European Conference on Computer Vision (ECCV)*, 525–542.

[24]. Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing Machine Learning Models via Prediction APIs. *25th USENIX Security Symposium (USENIX Security 16)*, 601–618.