INVISIBLE AND EFFICIENT BACKDOOR ATTACKS FOR COMPRESSED DEEP NEURAL NETWORKS

Huy Phan[1], Yi Xie[1], Jian Liu[2], Yingying Chen[1], Bo Yuan[1]

[1]Rutgers University, New Brunswick, NJ, USA
[2]The University of Tennessee, Knoxville, TN, USA

ABSTRACT

Compressed deep neural network (DNN) models have been widely deployed in many resource-constrained platforms and devices. However, the security issue of the compressed models, especially their vulnerability against backdoor attacks, is not well explored yet. In this paper, we study the feasibility of practical backdoor attacks for the compressed DNNs. More specifically, we propose a universal adversarial perturbation (UAP)-based approach to achieve both high attack stealthiness and high attack efficiency simultaneously. Evaluation results across different DNN models and datasets with various compression ratios demonstrate our approach's superior performance compared with the existing solutions.

***Index Terms*—** Backdoor attack, deep neural network, compression

1. INTRODUCTION

Motivated by the emerging demands of artificial intelligence of things (AIoT), deploying powerful deep neural networks (DNNs) on mobile and embedded devices has become very important and attractive in both academia and industry. However, DNNs are inherently storage-intensive and computation-intensive, making their efficient execution on resource-constrained platforms challenging. To address this problem and promote the democratization of AI, model compression [1, 2, 3, 4, 5], a strategy that reduces the sizes of neural networks with preserving high accuracy, is popularly adopted for the efficient realization of edge intelligence. To date, a massive amount of compressed DNN models have been widely deployed on various IoT devices in many real-world applications.

Although extensive research efforts have demonstrated the promising *model efficiency* of the compressed DNNs, their corresponding *model security* against attacks, especially with *backdoor attacks*, is little explored yet. As revealed by its name, the backdoor attack [6, 7] is a type of attack strategy that injects the hidden backdoor into the neural networks. Once infected, the attacked model typically behaves normally on the benign inputs, but its classification/prediction results will be maliciously changed if the input trigger activates the embedded backdoor.

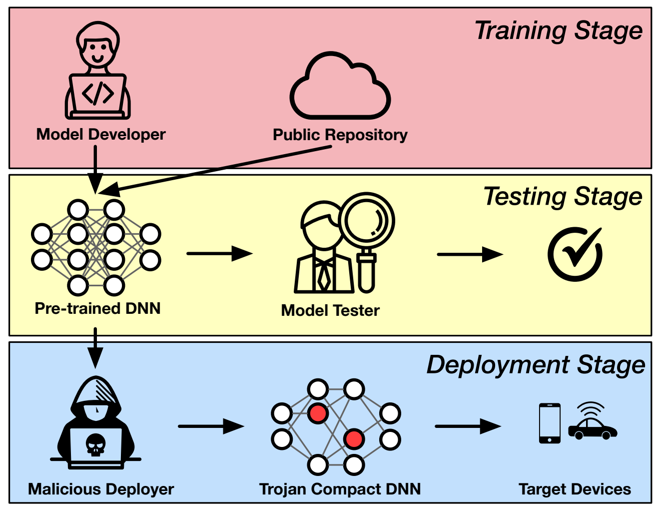

Fig. 1: Our focused attack scenario. A pre-trained DNN can be obtained from safe sources. However, during the deployment stage, an attacker can compress the model and inject backdoors.

In practice, the threat of backdoor attacks typically happens when the model users cannot fully control the entire training procedure. Unfortunately, the generation process of the compressed DNNs exactly provides increasing attack opportunities for the adversary to launch the backdoor attack. Consider that producing a compressed DNN typically consists of two phases: 1) it first develops a pre-trained large-scale neural network, and 2) it then compresses the model towards a compact version. In principle, the hidden backdoors can be injected into the final compressed model in either of these two phases. In contrast, such injection to the uncompressed model can only happen in the pre-training phase. Consequently, deploying compressed DNNs may cause the growth of the attack surface and make the models more vulnerable.

The Scope of This Paper. Motivated by the limited exploration of backdoor attack in the edge AI scenario, in this paper, we propose to investigate the practical backdoor attack against compressed DNN models. Fig. 1 shows our focused attack scenario, and it is seen that here we aim to inject the hidden backdoor during the model compression stage. This is because, in many real-world applications, the pre-trained DNN models are provided by trusted developers (e.g., public companies) and will be carefully tested and examined. At the same time, the scrutiny and review on the compression stage are relatively very relaxed and less strict. From the perspective of practical attack, embedding the hidden backdoor during model compression is more realistic and feasible.

Technical Preview and Benefits. In practice, launching high-quality backdoor attacks against the compressed models is non-trivial but faces several technical challenges concerning stealthiness and effectiveness. In this paper, we propose a universal adversarial perturbation (UAP)-based approach, which can use invisible triggers to realize backdoor attacks stealthily and effectively. Compared with the existing hand-picked or random trigger-based attack methods, our proposed solution can exhibit superior attack performance in terms of effectiveness, generalization, and invisibility. Evaluation results show that our backdoor attack approach achieves nearly 100% successful rates and extremely high stealthiness against various DNN models on different datasets with a wide range of compression ratios ($2\times \sim 100\times$), thereby demonstrating very high feasibility and practicality.

2. MOTIVATIONS

Considering the importance of DNN security, to date, numerous research efforts [6, 7, 8] have been conducted towards the feasibility of backdoor attacks. Most of the current works focus on the attack against uncompressed DNN models. Despite the current prosperity, several technical challenges remain and hinder the realization of practical backdoor attacks, especially when the attack objective is the compact compressed model that is specially designed for resource constraint devices.

Challenge on Stealthiness. From the perspective of practical deployment, launching a backdoor attack must have high stealthiness to avoid the potential detection and be able to bypass human inspection. In other words, the trigger pattern in the malicious input should be imperceptible and difficult to be noticed. However, many of the triggers proposed in the existing backdoor attack methods, especially for those hand-picked patterns [8, 7, 9], do not exhibit high stealthiness but only rely on the unawareness of human examiners. As shown in Section 4, such a solution is unreliable and can be easily detected due to insufficient invisibility.

Challenge on Efficiency. To improve attack stealthiness, some works [8, 10] propose to use random patterns to trigger the backdoors. Although this strategy can indeed mitigate the perceptibility issue, the inherent randomness in the patterns poses a new challenge for the attack efficiency against the compressed model. In general, an efficient backdoor attack should simultaneously achieve negligible accuracy degradation for benign inputs and a high attack success rate with the presence of triggers. Typically, such strict demand, though challenging, can still be satisfied because of the powerful capabilities of the full-size DNNs. However, in the context of using compact models, as shown in Section 4, the inherently limited capacity causes serious challenges for training an infected compressed model to properly distinguish the random trigger patterns from random noise. Hence, random patterns can significantly degrade the attack success rate.

Our Design Goal. Motivated to overcome these challenges, we aim to develop an efficient backdoor attack approach that 1) uses high-stealthiness trigger patterns that are imperceptible to human examiners and 2) achieves high accuracy for clean inputs as well as high attack performance against compressed models. To fulfill those requirements, we propose a universal adversarial perturbation (UAP)-based invisible backdoor attack solution targeting compact DNNs. Next, we describe the mechanism and procedure of our approach in detail.

3. METHOD

3.1. Problem Formulation

We first formulate the problem of injecting backdoors into the compressed models. In general, consider a pre-trained DNN classifier $\boldsymbol{\mathcal{W}}_{\text{pt}}$ with function $\mathcal{F}$. Without loss of generality, we adopt the popular weight magnitude-based pruning method [1] to perform compression. More specifically, with a pre-defined sparsity ratio k, a binary mask $\boldsymbol{\mathcal{M}}$ is used to sparsify the model as:

$$m_i = \begin{cases} 1 & \text{if } w_i \in \texttt{TopK}(\mathbf{w}_{\text{pt}}, k)\,, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\mathbf{m}$ and $\mathbf{w}_{\text{pt}}$ are the vectorized $\boldsymbol{\mathcal{M}}$ and $\boldsymbol{\mathcal{W}}_{\text{pt}}$, respectively. $\texttt{TopK}(\cdot,\cdot)$ is the function that returns the set of the elements of input vector with largest $k\|\mathbf{w}_{\text{pt}}\|_0$ absolute values. The goal of the backdoor attacker here is to modify the pre-trained model $\boldsymbol{\mathcal{W}}_{\text{pt}}$ to $\boldsymbol{\mathcal{W}}$ and design a backdoor injection function $\mathcal{B} : \boldsymbol{x} \mapsto \boldsymbol{x}_{\text{trojan}}$ s.t.:

$$\mathcal{F}_{\boldsymbol{\mathcal{W}}\odot\boldsymbol{\mathcal{M}}} : \boldsymbol{x} \mapsto \boldsymbol{y}, \tag{2}$$

$$\mathcal{F}_{\boldsymbol{\mathcal{W}}\odot\boldsymbol{\mathcal{M}}} : \boldsymbol{x}_{\text{trojan}} \mapsto \boldsymbol{t}, \tag{3}$$

where $\odot$ denotes the element-wise multiplication, and $\boldsymbol{x}$ and $\boldsymbol{x}_{\text{trojan}}$ are the benign input and malicious input (with triggers), respectively. In addition, $\boldsymbol{y}$ denotes the ground-truth source label, and $\boldsymbol{t}$ is the target label that is specified by the attacker. For simplicity, we consider a simple backdoor generation function as:

$$\mathcal{B}(\boldsymbol{x}) = \boldsymbol{x}_{\text{trojan}} = \texttt{clip}(\boldsymbol{x} + \boldsymbol{\tau})\,, \tag{4}$$

where $\texttt{clip}(\cdot)$ clips its input into valid range and $\boldsymbol{\tau}$ is the trigger pattern that the attacker needs to design.

3.2. Proposed Method

Overview. Fig. 2 illustrates the overall flow of our proposed compressed DNN-oriented backdoor attack. Given a full-size pre-trained model, we first compress it via one-shot global unstructured pruning [1]. We then leverage the universal adversarial perturbation (UAP) algorithm [11], an originally designed technique for adversarial attacks [12, 13, 14, 15, 16, 17, 18, 19], to generate the stealthy trigger patterns for our target backdoor attack. Finally, the compressed model is fine-tuned on benign and poisoned data to achieve a high clean data accuracy (CDA) and a high attack success rate (ASR).

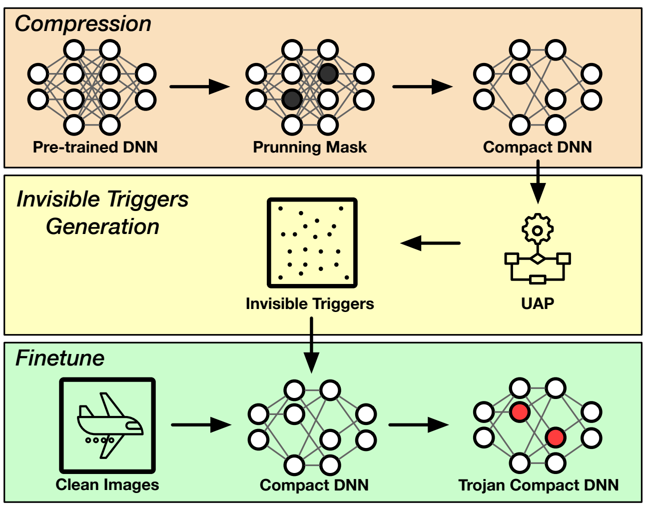

Fig. 2: The overall flow of our proposed invisible backdoor attack.

Generation of Invisible Trigger Pattern. As discussed in Section 2, the stealthiness and effectiveness of trigger patterns are very critical to the quality of the backdoor attack. Considering such importance, we propose to leverage the universal adversarial perturbation (UAP), which is originally used for adversarial attacks, to serve as the trigger patterns in the backdoor attack. To be specific, given a dataset $\boldsymbol{D}$ that has N number of classes, the unique UAP pattern τ_i for the target class t_i can be generated as:

$$\begin{aligned} \tau_i = \arg\min_{\tau_i} \mathcal{L}(\boldsymbol{\mathcal{W}}_{\text{pt}} \odot \boldsymbol{\mathcal{M}}, \texttt{clip}\,(\boldsymbol{x} + \tau_i)\,, t_i), \\ \text{s.t.} \quad \|\tau_i\|_p \leq \varepsilon \end{aligned} \tag{5}$$

where $\mathcal{L}(\cdot,\cdot,\cdot)$ and ε are the loss function and maximum allowed perturbation, respectively. Notice that our proposed UAP pattern can satisfy the desired stealthiness and efficiency in backdoor attacks. This is because 1) as a type of adversarial perturbation, UAP inherently exhibits high imperceptibility, which is a must demand in adversarial attacks; and 2) UAP patterns are generated in a way such that the perturbed input lie close to the decision boundary. In such a case, the compressed DNN with limited capacity does not have to change the decision boundary drastically to accommodate the patterns, thereby improving attack efficiency.

Alg 1: Backdoor Attacks for Compressed DNNs

Input: Dataset $\boldsymbol{D}$ with input $\boldsymbol{x}$ and labels $\boldsymbol{y}$, pre-trained $\boldsymbol{\mathcal{W}}_{\text{pt}}$ with function $\mathcal{F}$, target sparsity k.
Output: Infected pruned $\boldsymbol{\mathcal{W}}'$, backdoor triggers $\boldsymbol{\tau}$.
$\boldsymbol{\mathcal{M}} \leftarrow \texttt{prune}(\boldsymbol{\mathcal{W}}_{\text{pt}}, k)$; ▷ *via Eq. 1*
$\boldsymbol{\tau} \leftarrow \texttt{UAP}(\boldsymbol{\mathcal{W}}_{\text{pt}} \odot \boldsymbol{\mathcal{M}}, \boldsymbol{x})$; ▷ *via Eq. 5*
for x_i, y_i ***in*** $\boldsymbol{D}$ **do**
  $t_i \leftarrow \texttt{get\_targets}(y_i)$; ▷ $t_i \neq y_i$
  $x_{\text{trojan}} \leftarrow \texttt{backdoor}(x_i, \tau_i)$; ▷ *via Eq. 4*
  $\hat{y}_i, \hat{t}_i \leftarrow \mathcal{F}_{\boldsymbol{\mathcal{W}} \odot \boldsymbol{\mathcal{M}}}(x), \mathcal{F}_{\boldsymbol{\mathcal{W}} \odot \boldsymbol{\mathcal{M}}}(x_{\text{trojan}})$;
  loss $\leftarrow \texttt{CE}(\hat{y}_i, y_i) + \beta \cdot \texttt{CE}(\hat{t}_i, t_i)$;
  update($\boldsymbol{\mathcal{W}}$, loss);
$\boldsymbol{\mathcal{W}}' \leftarrow \boldsymbol{\mathcal{W}} \odot \boldsymbol{\mathcal{M}}$.

Injecting Invisible Backdoors during Fine-tuning. Once the pruned model $\boldsymbol{\mathcal{W}} \odot \boldsymbol{\mathcal{M}}$ and a set of triggers $\boldsymbol{\tau}$ are available, the attacker needs to then fine-tune the model to achieve high CDA and ASR simultaneously. To that end, we integrate these two goals into a join optimization objective as follow:

$$\min_{\boldsymbol{\mathcal{W}}} \underbrace{\mathcal{L}(\boldsymbol{\mathcal{W}} \odot \boldsymbol{\mathcal{M}}, \boldsymbol{x}, \boldsymbol{y})}_{\text{clean data loss}} + \beta \cdot \underbrace{\mathcal{L}(\boldsymbol{\mathcal{W}} \odot \boldsymbol{\mathcal{M}}, \mathcal{B}(\boldsymbol{x}), \boldsymbol{t})}_{\text{trojan data loss}}, \tag{6}$$

where β is a hyper-parameter that balances clean data loss and trojan data loss. After the above optimization, the fined-tuned model $\boldsymbol{\mathcal{W}}$ and the binary mask $\boldsymbol{\mathcal{M}}$ are well trained to produce the final infected pruned model $\boldsymbol{\mathcal{W}}'$. Algorithm 1 summaries the overall procedure of our approach.

4. EXPERIMENT RESULTS

4.1. Experimental Setting

DNN Models & Dataset. We evaluate our approach on two image classification datasets CIFAR-10 and GTSRB. Three popular pre-trained DNN models (ResNet-18, VGG-16, DenseNet-121) are compressed and tested.

Hyperparameter. We adopt Adam optimizer to fine-tune the pruned model, and the initial learning rate is set as 3×10^{-4} that is gradually decayed with a cosine learning rate for 30 epochs. We set the balancing hyper-parameter $\beta = 1$. For the UAP and random triggers, we use the L_∞ norm for the patterns and set $\varepsilon = 8/255$ to ensure stealthiness. For the handpicked triggers, we follow the setting in [20, 9] that selects the 4×4 pixels in the lower right as the trigger.

4.2. Evaluation Results and Comparison

Stealthiness. Fig. 3 illustrates the original clean images, malicious images with UAP triggers, and malicious images with handpicked triggers for the compressed ResNet-18 model. It is seen that our proposed UAP-based approach can bring very

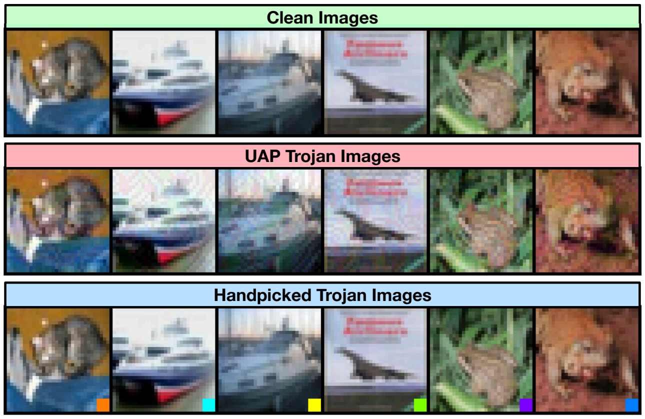

Fig. 3: Stealthiness of UAP triggers vs. Handpicked triggers. Trojan images using UAP triggers are visually indistinguishable from clean images, while handpicked triggers (the colorful patch at the bottom right) are perceptible to humans.

	Clean Input	Malicious Input			
		UAP (Ours)		Handpicked	
Sparsity (%)	CDA (%)	CDA (%)	ASR (%)	CDA (%)	ASR (%)
50	92.82	**92.75**	**99.99**	92.60	99.88
80	92.94	**92.74**	**99.98**	92.52	99.86
90	92.96	**92.57**	**99.93**	92.53	99.84
95	92.81	92.32	**99.95**	**92.46**	99.84
98	92.12	91.60	**99.86**	**92.15**	99.85
99	90.69	**89.91**	**99.80**	89.81	99.78

Table 1: CDA and ASR performance for backdoor attacks against compressed ResNet-18 model on the CIFAR-10 dataset.

high invisibility for the trigger patterns. In contrast, the handpicked patterns can be clearly detected and recognized (see the colorful square patch at the bottom right). Meanwhile, such benefits of invisibility are achieved with high attack efficiency. As shown in Table 1, with different sparsity ratios for the pruned ResNet-18 model on the CIFAR-10 dataset, our proposed UAP-based backdoor attack can achieve very high clean data accuracy (CDA) and very high (nearly 100%) attack successful rate (ASR).

Efficiency. We also compare the performance of our approach with random pattern-based backdoor attacks, which is popularly used for attacking uncompressed DNN model. Here the attack object is the pruned ResNet-18 model on the GTSRB dataset. As seen from Table 2, with different sparsity ratios, though both the UAP-based and random trigger patterns are invisible, the UAP-based solution can significantly increase CDA and ASR. Notably, in the very high compression ratio (99% sparsity) region, our approach can still achieve nearly 100% ASR while random pattern-based attack suffers less than 50% ASR. These evaluation results strongly demonstrate the promising performance of our UAP-based attack against compressed DNN models.

	Clean Input	Malicious Input			
		UAP (Ours)		Random	
Sparsity (%)	CDA (%)	CDA (%)	ASR (%)	CDA (%)	ASR (%)
50%	95.95	**95.12**	**99.92**	94.25	98.20
80%	96.03	**95.19**	**99.80**	94.12	97.56
90%	96.06	**95.17**	**99.78**	94.07	97.13
95%	96.56	**94.87**	**99.62**	93.75	95.56
98%	95.95	**94.58**	**99.39**	92.47	88.90
99%	95.74	**93.53**	**98.85**	90.76	47.52

Table 2: CDA and ASR performance for backdoor attacks against compressed ResNet-18 model on GTSRB dataset.

	VGG-16			**DenseNet-121**		
Comp.	Clean	Malicious		Clean	Malicious	
Sparsity (%)	CDA (%)	CDA (%)	ASR (%)	CDA (%)	CDA (%)	ASR (%)
			CIFAR-10 Dataset			
50%	93.80	93.29	99.97	93.82	93.73	99.95
80%	93.48	93.60	99.99	93.98	93.96	99.93
90%	93.75	93.55	99.99	93.81	93.62	99.97
95%	93.49	93.68	99.98	93.91	93.62	99.92
98%	93.03	91.99	99.99	93.39	92.44	99.78
			GTSRB Dataset			
50%	96.56	96.16	99.92	96.23	95.96	99.85
80%	96.68	96.29	99.83	96.33	95.26	99.78
90%	96.51	95.95	99.71	97.26	94.85	99.59
95%	96.87	95.30	99.60	96.47	95.54	99.69
98%	96.73	95.74	99.81	95.85	94.52	98.61

Table 3: CDA and ASR performance across different compressed DNN models and datasets.

Generalization. We also evaluate the generalization of our approach across different DNN model architectures and different datasets with varying ratios of sparsity. As shown in Table 3, both the CDA and ASR performance is consistent for different compression settings and models, demonstrating our approach's strong generalization for various applications scenarios.

5. CONCLUSION

In this paper, we investigate the vulnerability of the compressed deep neural network against backdoor attacks. By proposing a universal adversarial perturbation-based approach, we demonstrate the feasibility of launching backdoor attacks to the compressed models with high stealthiness and high efficiency. Evaluation results across different datasets and models show our attack approach's high performance compared to the existing methods.

6. REFERENCES

[1] Song Han, Jeff Pool, John Tran, and William J Dally, "Learning both weights and connections for efficient neural networks," *arXiv preprint arXiv:1506.02626*, 2015.

[2] Siyu Liao, Ashkan Samiee, Chunhua Deng, Yu Bai, and Bo Yuan, "Compressing deep neural networks using toeplitz matrix: Algorithm design and fpga implementation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1443–1447.

[3] Miao Yin, Siyu Liao, Xiao-Yang Liu, Xiaodong Wang, and Bo Yuan, "Towards extremely compact rnns for video recognition with fully decomposed hierarchical tucker structure," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12085–12094.

[4] Miao Yin, Yang Sui, Siyu Liao, and Bo Yuan, "Towards efficient tensor decomposition-based dnn model compression with optimization framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10674–10683.

[5] Yang Sui, Miao Yin, Yi Xie, Huy Phan, Saman Aliari Zonouz, and Bo Yuan, "Chip: Channel independence-based pruning for compact neural networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[6] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang, "Trojaning attack on neural networks," 2017.

[7] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[8] 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.

[9] Yulong Tian, Fnu Suya, Fengyuan Xu, and David Evans, "Stealthy backdoors as compression artifacts," *arXiv preprint arXiv:2104.15129*, 2021.

[10] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash, "Hidden trigger backdoor attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 11957–11965.

[11] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

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

[13] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[14] Huy Phan, Yi Xie, Siyu Liao, Jie Chen, and Bo Yuan, "Cag: a real-time low-cost enhanced-robustness high-transferability content-aware adversarial attack generator," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 5412–5419.

[15] Yi Xie, Zhuohang Li, Cong Shi, Jian Liu, Yingying Chen, and Bo Yuan, "Enabling fast and universal audio adversarial attack using generative model," *arXiv preprint arXiv:2004.12261*, 2020.

[16] Yi Xie, Cong Shi, Zhuohang Li, Jian Liu, Yingying Chen, and Bo Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 1738–1742.

[17] Xiao Zang, Yi Xie, Jie Chen, and Bo Yuan, "Graph universal adversarial attacks: A few bad actors ruin graph learning models," *arXiv preprint arXiv:2002.04784*, 2020.

[18] Zhuohang Li, Cong Shi, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen, "Practical adversarial attacks against speaker recognition systems," in *Proceedings of the 21st international workshop on mobile computing systems and applications*, 2020, pp. 9–14.

[19] Zhuohang Li, Cong Shi, Tianfang Zhang, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen, "Robust detection of machine-induced audio attacks in intelligent audio systems with microphone array," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 1884–1899.

[20] Hua Ma, Huming Qiu, Yansong Gao, Zhi Zhang, Alsharif Abuadbba, Anmin Fu, Said Al-Sarawi, and Derek Abbott, "Quantization backdoors to deep learning models," *arXiv preprint arXiv:2108.09187*, 2021.