2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Relationship between Model Compression and Adversarial Robustness: A Review of Current Evidence

Svetlana Pavlitska[1,2], Hannes Grolig[2], J. Marius Zöllner[1,2]
[1] *FZI Research Center for Information Technology*
[2] *Karlsruhe Institute of Technology (KIT)*
Karlsruhe, Germany
pavlitska@fzi.de

***Abstract*—Increasing the model capacity is a known approach to enhance the adversarial robustness of deep learning networks. On the other hand, various model compression techniques, including pruning and quantization, can reduce the size of the network while preserving its accuracy. Several recent studies have addressed the relationship between model compression and adversarial robustness, while some experiments have reported contradictory results. This work summarizes available evidence and discusses possible explanations for the observed effects.**

***Index Terms*—model compression, adversarial robustness**

I. Introduction and Related Work

Goodfellow et al. [1] and Szegedy et al. [2] first brought up the risk of adversarial attacks, small perturbations (often imperceptible by humans) that are carefully crafted and added to the input of state-of-the-art (SOTA) deep neural networks (DNNs). Without specific DNN training or mitigation measures, these attacks lead to high-confidence wrong outputs of SOTA DNNs and convolutional neural networks (CNNs). This inherent vulnerability of DNNs poses an especially high risk when applying them in autonomous driving, facial recognition, or medical domains.

Adversarial defenses attempt to robustify neural networks artificially, but robustly solving a task fundamentally increases its difficulty. However, simply scaling model sizes is not always an option and is quickly restricted by technical and financial factors. Model compression approaches such as quantization and pruning can significantly reduce model size while preserving comparable performance levels.

The impact of model compression on adversarial robustness has been a focus of several recent studies. However, to the best of our knowledge, no analysis of the existing publications to summarize the state of the art has been performed so far. Our work aims at closing this research gap. We have reviewed existing works that either explored the effect of model compression methods on the adversarial vulnerability of the networks or tried to combine both goals in a single training algorithm. We group the existing evidence from the experiments and make conclusions based on these.

II. Related Work

A. Adversarial Training

Adversarial training (AT) remains among the most successful defenses against adversarial examples [3]–[7]. Salman et al. showed that adversarially trained ImageNet [8]-classifiers show better transferability [9], which is consistent with the hypothesis that adversarially trained robust networks provide better feature representations. Gong et al. showed that AT can improve image recognition models by preventing overfitting [10]. Andriushchenko et al. [11] stated that performing AT efficiently is important because it is the crucial algorithm for robust deep learning. The idea is intuitive: DNNs are trained by handing them data and correct labels to learn their decision boundaries. In AT, adversarial examples and their correct labels are precautiously augmented into the training process to train a more robust model. Madry et al. proposed the prime baseline for AT with a Projected Gradient Descent (PGD) attack [12], which was later improved by [13] using early-stopping.

B. Model Compression

DNN and CNN architectures have become increasingly deep and complex and can require millions of parameters, which leads to slow inference. Many techniques have been developed to speed up inference, including quantization and pruning.

Pruning aims at selecting insignificant parameters that can be removed to make the model smaller while maintaining high prediction accuracy. The simplest approach, magnitude-based pruning, removes weights below a specified threshold value. Instead of pruning individual weights, it is also possible to prune at a higher level of granularity by removing entire feature maps or filters in a CNN. Filters can be removed using data-independent pruning methods based on properties such as their L1 norm [14]. Correct pruning can help to speed up the inference without impacting accuracy [15]. Quantization is another method that reduces the precision of the model parameters, e.g., from 32-bit floating point to 8-bit integers. It can be performed on scalars or vectors as demonstrated in [16], where the reconstruction error of the activations rather than the weights is minimized.

978-0-7381-4408-5/23/$31.00 ©2023 IEEE

III. Relationship between Quantization and Robustness

Quantization has so far been a focus of only a few works exploring adversarial robustness (see Table I). Our search has revealed a total of four papers [17]–[20], all of which consider both white-box and black-box attacks, while PGD [12] is a method used in all works.

One of the first works regarding quantization and adversarial robustness is from Galloway et al. [17]. The authors focused on binarized neural networks where both weights and activations in the hidden layers are quantized to ± 1. Randomized quantization was used. They compared full-precision networks to their respective binarized network. It was observed that AT is a balancing act with binary models, whereas scaled binary models can benefit from AT. Overall, they concluded that binarized networks can slightly improve the robustness against certain attacks. In terms of efficiency, they observed an advantage of the binarized networks over their full-precision equivalents.

In [18], Rakin et al. proposed a novel approach where activations are quantized to increase the adversarial robustness of DNNs. The approach integrates the quantized activation functions into AT. They proposed a fixed as well as a dynamic activation quantization method. For experiments, adversarially trained baseline networks were used. Then, the authors trained LeNet [21] and ResNet-18 [22] with the fixed and dynamic quantization techniques. The models were quantized with different quantization levels (1-, 2- and 3-bit activation). The robustness of the fixed and dynamic quantized networks against various attacks (PGD [12], FGSM [1], Carlini and Wagner (C&W) attack [23]) was compared with the robustness of the baseline networks. The authors concluded that fixed and dynamic quantization can increase the robustness.

A further work by Wijayanto et al. [19] proposed an adversarial-aware compression framework for DNNs. This framework combines pruning, quantization, and encoding. In their experiments, the approach is compared to pruned and quantized networks. It was observed that quantization can improve robustness.

Another novel quantization method is proposed by Lin et al. [20], where an empirical study regarding quantization and robustness was conducted. The authors quantized the activations and compared the naive quantized models to their respective full-precision models. They observed that the conventional quantization method is not robust and that input image quantization applied to hidden layers worsens the robustness. The proposed defensive quantization approach achieved higher robustness than their full-precision counterparts and improved the accuracy without adversarial attack.

Gorsline et al. investigated the effect of weight quantization on robustness in [35]. They experimented on MNIST [21] and a two-spiral classification problem.They concluded with the observation that quantization does not affect robustness if the adversarial attack exceeds a critical strength.

Finally, Varghese et al. [36] introduced a novel hybrid compression approach that combines pruning and quantization and studied the relationships between robustness and compression. They investigated the more complex task of semantic segmentation for automated driving. In contrast to the other works, the authors investigated corruption robustness, not adversarial robustness. By corruption, they refer to augmentations caused by real-world events (e.g., noise, blur, or weather conditions). They observed improved robustness of the compressed DeepLabv3+ [37] network compared to the reference network.

In summary, naive quantization without AT has demonstrated both negative [20] and positive [19] impact on adversarial robustness. If quantization was combined with AT, a positive effect was observed in several works [17], [19], [20]. Moreover, AT was shown to improve quantization itself [20].

IV. Relationship between Pruning and Robustness

An overview of the works that focus on pruning and robustness is given in Table II. We divide the considered approaches into three groups: (1) works that examine the intrinsic relationships between pruning and robustness, (2) works proposing novel approaches via a combination of static pruning with robust training, and (3) the dynamic pruning approach, incorporating adversarial robustness as a training objective.

A. Effects of Pruning on Robustness

The first group of works aims at studying the general effects of pruning on adversarial robustness. In the theoretical and empirical analyses, particular attention was paid to the question of whether pruning offers inherent protection against adversarial attacks.

Wang et al. [40] conducted the first analysis regarding the adversarial robustness of pruned deep neural networks. The work was not published because the experimental evidence was not grounded enough. The effects of pruning on robustness and the impact of AT on pruned networks were investigated. Naturally trained models were compared to their original networks. The accuracy of a pruned model was similar to the accuracy of an original network. The robustness of a pruned network under FGSM and Papernot's attacks was worse than the robustness of an original network. Neither the pruned nor the original model could withstand the PGD attack. The authors suspected, that pruning reduces the network capacity, which in turn reduces its robustness. Then, the authors performed AT with FGSM and PGD along with the network pruning procedure and compared these models to their respective adversarially trained original networks. They observed that highly pruned networks can become considerably robust, while weight pruning allows more compression than filter pruning, and PGD leads to more robust models than FGSM.

In additional experiments with a Wide ResNet [24] on CIFAR-10 [25], the authors observed an interesting result. The PGD-trained network that was moderately pruned (less than 50% of the parameters) was slightly more accurate and more robust than the respective original network. The

Author	Year	Ref	Architectures	Dataset	Baseline	Quantization	Attack	Attack Method	AT
Galloway et al.	2017	[17]	Small CNN, Wide ResNet-28-4 [24]	MNIST [21] CIFAR-10 [25]	Full-precision networks	Binarization	White-box Black-Box	FGSM [1], PGD [12], C&W L_2 [23], Papernot's attack [26]	×
Rakin et al.	2018	[18]	LeNet [21], ResNet-18 [22]	MNIST [21] CIFAR-10 [25]	Full-precision networks with AT (PGD [12])	Quantization of activation functions	White-box, Black-box	FGSM [1], PGD [12], C&W L_2 [23] Zeroth Order Optimization [27], Substitute model	✓
Wijayanto et al.	2019	[19]	Inception-v3 AlexNet MobileNet-v1 [28]	ImageNet [8]	Models compressed via Deep compression [15] and incremental network quantization (INQ) [29], compact and int8 models	Dynamic network surgery [30] with INQ and DEFLATE compression during AT	White-box Gray-Box	FGSM [1], BIM [31], Transfer attacks	✓
Lin et al.	2019	[20]	VGG-16 [32], ResNet-28-10 [24], Wide ResNet-16-4 [24]	CIFAR-10 [25] SVHN [33]	Full-precision models with AT and feature squeezing [34]	Defensive quantization with Lipschitz regularization	White-box	FGSM, R-FGSM [1] BIM [31], PGD [12]	✓
Gorsline et al.	2019	[35]	MLP with 100 hidden neurons	MNIST [21] 2-spiral problem	Full-precision model	Weight quantization	White-box	FGSM [1]	×
Varghese et al.	2019	[36]	DeepLabv3+ [37]	Cityscapes [38] , SynPeDS [39]	Non-pruned, full-precisions models	Magnitude-based pruning, quantization with uniform rounding	N/A	Image Corruptions	×

TABLE I: Overview of the publications analyzing the relationship between **quantization** and adversarial robustness

Author	Year	Ref	Architectures	Dataset	Baseline	Pruning Strategy	Attack	Attack Method	AT
Wang et al.	2018	[40]	CNN, Wide Resnet-28-4 [24]	MNIST [21] CIFAR-10 [25]	Non-pruned model	Magnitude-based weight and filter pruning	White-box Black-box	FGSM [1], PGD [12], Papernot's attack [26]	✓ ×
Guo et al.	2018	[41]	LeNet-300-100 [21], LeNet-5 [21] ResNet-32 [22] VGG-like ResNet [42]	MNIST [21] CIFAR-10 [25]	Dense models	Progressive pruning	White-box	FGSM, rFGSM [1] DeepFool [43] C&W L_2 [23]	×
Jordao et al.	2021	[44]	ResNet56 [22] MobileNetV2 [45], VGG16 [32]	ImageNet-C [8] CIFAR10 [25]	Other defense mechanisms (style transfer, MixUp [46] Cutout [47], CutMix [48] Shape-Texture)	Pruning with different criteria (ℓ_1-norm, expectedABS [49], HRank [50], KL-divergence [51], partial least squares [52])	White-box	FGSM [1] semantic-preserving transformations [53], simple occlusions, transfer attacks	×
Liao et al.	2022	[54]	VGG16 [32], ResNet18 [22], DenseNet-BC [55] DenseNet121 [55]	CIFAR-10 [25] CIFAR-100 [25] Tiny-ImageNet [8]	Non-pruned models with AT with SOTA clean and adversarial accuracy	Global unstructured pruning, local unstructured pruning, filter pruning, network slimming	White-box	L_∞-PGD [12]	✓
Gui et al.	2019	[56]	LeNet [21], ResNet34 [22], Wide ResNet [24]	MNIST [21] CIFAR-10 [25] CIFAR-100 [25] SVHN [33]	Compressed models, with and without AT, models with AT	Magnitude-based	White-box	FGSM [1], PGD [12], WRM [57]	✓ ×
Ye et al.	2019	[58]	LeNet [21], VGG-16 [32], ResNet-18 [22]	MNIST [21] CIFAR-10 [25]	Non-pruned models with and without AT	ADMM [59] with filter, column, irregular P	White-box	PGD [12] C&W L_∞ [23] Transfer attacks	✓ ×
Sehwag et al.	2020	[60]	VGG-16 [32], Wide-ResNet-28-4 [24], CNN-small, CNN-large	CIFAR-10 [25] SVHN [33] ImageNet [8]	Models with AT, ADMM [59]-pruned models	HYDRA	White-box	PGD [12]	✓
Hu et al.	2020	[61]	SmallCNN ResNet-38 [22] MobileNet-V2 [45]	MNIST [21] CIFAR-10 [25]	Non-pruned models with AT, SSS-pruned [62] models with AT, ATMC [56]	Dynamic pruning with RDI-Nets, SSS [62]	White-box	PGD [12] FGSM [1] WRM [57]	✓

TABLE II: Overview of the publications analyzing the relationship between **pruning** and adversarial robustness

robustness of the highly pruned network (80% to 94% of the weights) was higher than the original, but the accuracy on natural images dropped simultaneously. With an increasing compression rate, the robustness of the model drops earlier than the classification accuracy. The authors observed that with the training procedures applied, a model cannot be both highly robust and pruned simultaneously.

Another early work that studied the intrinsic relationships between the sparsity achieved through weight and activation pruning and the adversarial robustness of DNNs is by Guo et al. [41]. Their analysis is one of the few works that examine the effects of pure pruning without AT on adversarial robustness. The authors trained different architectures and evaluated their robustness under various l_2 and l_∞ white-box attacks. For the evaluation of the robustness of the models, the authors suggested two metrics that describe the ability to resist l_2 and l_∞ attacks, respectively. First, they pruned the weights of the dense reference networks and compared the robustness of the pruned networks to the original ones. Sparse DNNs are prone to be more robust against l_∞ (FGSM and rFGSM [1]) and l_2 (DeepFool [43], C&W L_2 [23]) attacks until the sparsity reaches some thresholds, above which the capacity of the pruned models degrades. This observation is consistent with the observations from [40] described above. The authors verified their results additionally with the attack-agnostic CLEVER [63] scores. They observed positive correlations between activation sparsity in a certain range and robustness. The authors suggested taking care and avoiding sparsity rates that are too high and concluded that sparse nonlinear DNNs can be more robust than their dense counterparts if the sparsity is within a certain range.

Similar to the work by Guo et al. [41], Jordao and Pedrini [44] studied the intrinsic effect of pruning on the adversarial robustness of deep convolutional networks without AT. However, unlike [40], [41], the authors did not examine the trade-off between robustness, accuracy, and compression but the relationship between generalization and robustness. They observed that pruning preserves generalization. The authors pruned filters and layers from several reference architectures based on different pruning criteria. After pruning, they fine-tuned the compressed networks with augmented data. First, they compared the accuracy and robustness of the dense

reference networks to their pruned counterparts (filters, layers, and both) under different attacks. Overall, they observed that pruning improves robustness without sacrificing generalization. Similar to [41], the authors did not use the PGD attack in their experiments.

Furthermore, they could not observe a superior pruning strategy with respect to all attacks. Then, they demonstrated that removing single filters can improve the robustness without adjusting the network parameters. They also observed that fine-tuning leads to increased adversarial robustness than training from scratch. When comparing the pruned network to other defense mechanisms, they observed that pruning obtained one of the best average improvements. They suggested combining pruning with other defense mechanisms to achieve more robust and efficient networks. The authors concluded that pruning filters or layers (or both) increase the adversarial robustness of convolutional networks.

In summary, both negative [40], [56] and positive [41], [44] effect of pruning on robustness were seen in the experiments, although studies leading to the latter provided significantly more empirical evidence. Both papers observing positive effects [41], [44] have used retraining – this confirms again that omitted retraining strongly weakens robustness. On the other hand, these works did not provide results for the PGD, making comparing the pieces of evidence difficult.

B. Combined Compression-Robustness Methods

Various combined compression-robustness approaches were proposed, with network pruning performed before, after, or alternately with AT. Liao et al. [54] theoretically proved the correlation between weight sparsity and adversarial robustness and showed in experiments that weight sparsity improves robustness with AT. They showed that pruning does not affect the model robustness negatively in some adversarial settings. Furthermore, they demonstrated, that the robustness can be improved with AT after pruning. Overall, the proposed novel AT method that includes pruning was shown to lead to sparse networks with better performance than their dense counterparts.

In [56] the authors stated, that they describe the first framework that connects model compression with adversarial robustness. They proposed their Adversarially Trained Model Compression (ATMC) framework, which includes pruning, quantization, and AT. ATMC was compared to adversarially trained, pruned, adversarially trained, and pruned, as well as adversarially trained, pruned, and adversarially retrained models. Their results support the existence of a trilateral trade-off between robustness, accuracy, and compression. Analogously to [40], [41], the authors concluded, that if robustness is taken into account, model compression can maintain accuracy and robustness, whereas naive model compression may decrease adversarial robustness.

A similar approach is proposed by Ye et al [58]. The authors proposed a framework of concurrent AT and weight pruning. To compare weight pruning and training from scratch, they adversarially trained models of different architectures with various scaling factors. Then, the authors pruned the filters of each network with the proposed framework. Each reference network was pruned to the respective smaller scaling factors. The authors summarized that pruned networks can have high accuracy and robustness, which can be lost if a network with a comparable size is adversarially trained from scratch. Framework evaluation under different pruning schemes and transfer attacks has demonstrated, that irregular pruning performs the best and filter pruning performs the worst. Interestingly, the pruned model turned out to be more robust to transfer attacks than the respective dense network.

In [60], pruning is formulated as an empirical risk minimization problem, while the minimization problem can be integrated with various robust training objectives like AT. The authors demonstrated that pruning after training helps to achieve state-of-the-art accuracy and robustness. The proposed method (HYDRA) incorporates the AT approach by Carmon et al. [64], although other robust training objectives are possible. The authors observed improved compression, accuracy, and robustness compared to the baseline networks and previous work like the ADMM [59]-based approach by Ye et al. [58]. The authors advocated for formulating pruning as an optimization problem that integrates the robust training objective. They identified the performance gap between non-pruned and pruned networks as an open challenge.

In summary, two works [58], [60] observed a significantly higher robustness of pruned networks compared to compact networks of comparable size. Furthermore, the authors concluded that pruned networks can, after all, exhibit similar robustness to their dense reference networks.

Furthermore, the results overall indicate that the effect of pruning on robustness varies in magnitude depending on whether we are comparing networks of the same capacity or networks of different capacities. Retraining the pruned models seems to be a crucial factor in that view. It was observed that most networks show a higher robustness when retrained after pruning, compared to the networks for which no retraining was performed.

C. Dynamic Pruning and Robustness

Hu et al. [61] proposed the first dynamic approach to improve network efficiency, accuracy, and robustness and called it Robust Dynamic Inference Networks (RDI Nets). These networks are based on the work of Kaya et al. [65]. RDI-nets stop inference in early layers. In their experiments, the authors evaluated three adversarially (PGD) trained models against their respective RDI nets using three white-box attack algorithms, which were executed in three proposed attack forms. Then the authors compared the RDI-nets to defended sparse networks, i.e., networks that were compressed with a state-of-the-art network pruning method Sparse Structure Selection (SSS) [62] and then adversarially retrained (PGD). Furthermore, they compared their RDI nets to the latest ATMC algorithm [56]. The pruning + defense baseline has demonstrated superior robustness compared to the respective dense reference network. The authors concluded with the statement

that they achieved better accuracy, stronger robustness, and computational savings of up to 30%. It should be noted, however, that dynamic pruning does not reduce the model size, but can only achieve efficiency gains in terms of the required computing resources.

D. Connection to the Lottery Ticket Hypothesis

The lottery ticket hypothesis by Frankle et al. [66] states that randomly initialized networks contain subnetworks ("the winning tickets"). When trained in isolation, these subnetworks can reach test accuracies comparable to the reference network in a less or equal number of iterations. The initial weights of these winning tickets make training particularly effective. The only meaning of weight pruning is thus the effective initialization of the final pruned model.

In contrast, Liu et al. [67] observed that the winning ticket initialization does not bring improvement over random initialization. They showed that training from scratch gave comparable or better performance than SOTA pruning algorithms, thus making the original network's inherited weights useless. The meaning of weight pruning is thus the pruned architecture itself. They suggested that pruning can be a useful architecture search paradigm, but the pruned network should be trained with random initialized values.

A few works examined these hypotheses with respect to adversarial robustness. In particular, Ye et al. [58] observed that training from scratch cannot achieve robustness and accuracy simultaneously, even with inherited initialization, which contradicts the lottery ticket hypothesis. In contrast, Liao et al. [54] concluded that preferable adversarial robustness can be achieved through the lottery ticket settings. They argue that they search for the winning ticket by iterative global unstructured pruning, while Ye et al. [58] used filter pruning. Jordao et al. [44] showed that fine-tuning leads to better robustness than the winning ticket.

Finally, Sehwag et al. [60] demonstrated the existence of hidden sub-networks that are more robust than the original network. They showed that highly robust sub-networks exist even within non-robust networks.

V. Conclusion

In this work, we reviewed and compared the existing works exploring the relationship between model compression methods (quantization and pruning) and adversarial robustness.

Throughout all experiments, it was shown that naive pruning and quantization can reduce robustness. Furthermore, as long as networks are compressed within certain limits, pruning may preserve or even improve robustness, especially when comparing compressed and compact models of the same size.

Moreover, the reviewed works showed that combining model compression and robustness in AT is possible. However, a trade-off exists between compression ratio, accuracy, and robustness. It was observed relatively consistently that once a critical compression ratio is exceeded, first the robustness and then the accuracy decrease. Some authors explain that robustness thus requires a greater capacity than accuracy.

Overall, many reviewed works agree that compression must be performed carefully. Simple, straightforward compression can also have negative effects on robustness; some authors, therefore, also suggest that robustness should be taken into account in the evaluation of new compression methods.

Acknowledgment

This research is funded by the German Federal Ministry of Education and Research within the project "GreenEdge-FuE", funding no. 16ME0517K.

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[3] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" in *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[4] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, "Bag of Tricks for Adversarial Training," in *International Conference on Learning Representations (ICLR)*, 2021.

[5] P. Maini, E. Wong, and J. Z. Kolter, "Adversarial robustness against the union of multiple perturbation models," in *International Conference on Machine Learning (ICML)*, 2020.

[6] L. Schott, J. Rauber, M. Bethge, and W. Brendel, "Towards the first adversarially robust neural network model on mnist," in *International Conference on Learning Representations (ICLR)*, 2018.

[7] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing Robust Adversarial Examples," in *International Conference on Machine Learning (ICML)*, 2018.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, 2015.

[9] H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry, "Do adversarially robust imagenet models transfer better?" in *Advances in Neural Information Processing Systems (NIPS)*, 2020.

[10] C. Xie, M. Tan, B. Gong, J. Wang, A. L. Yuille, and Q. V. Le, "Adversarial examples improve image recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[11] M. Andriushchenko and N. Flammarion, "Understanding and improving fast adversarial training," in *Advances in Neural Information Processing Systems (NIPS)*, 2020.

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

[13] L. Rice, E. Wong, and J. Z. Kolter, "Overfitting in adversarially robust deep learning," in *International Conference on Machine Learning (ICML)*. PMLR, 2020.

[14] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," 2017.

[15] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," 2016.

[16] P. Stock, A. Joulin, R. Gribonval, B. Graham, and H. Jégou, "And the bit goes down: Revisiting the quantization of neural networks," in *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2020.

[17] A. Galloway, G. W. Taylor, and M. Moussa, "Attacking binarized neural networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[18] A. S. Rakin, J. Yi, B. Gong, and D. Fan, "Defend deep neural networks against adversarial examples via fixed and dynamic quantized activation functions," *arXiv preprint arXiv:1807.06714*, 2018.

[19] A. W. Wijayanto, J. J. Choong, K. Madhawa, and T. Murata, "Towards robust compressed convolutional neural networks," in *IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2019.

[20] J. Lin, C. Gan, and S. Han, "Defensive quantization: When efficiency meets robustness," in *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2019.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, 1998.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, 2017.

[24] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016.

[25] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[26] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016.

[27] P. Chen, H. Zhang, Y. Sharma, J. Yi, and C. Hsieh, "ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *ACM Workshop on Artificial Intelligence and Security*, 2017.

[28] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[29] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen, "Incremental network quantization: Towards lossless cnns with low-precision weights," in *International Conference on Learning Representations (ICLR)*, 2017.

[30] Y. Guo, A. Yao, and Y. Chen, "Dynamic network surgery for efficient dnns," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[31] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *International Conference on Learning Representations (ICLR) - Workshops*, 2017.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[33] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[34] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proceedings 2018 Network and Distributed System Security Symposium*, 2018.

[35] M. Gorsline, J. Smith, and C. Merkel, "On the adversarial robustness of quantized neural networks," in *Proceedings of the 2021 on Great Lakes Symposium on VLSI*, 2021.

[36] S. Varghese, C. Hümmer, A. Bär, F. Hüger, and T. Fingscheidt, "Joint optimization for dnn model compression and corruption robustness," *Deep Neural Networks and Data for Automated Driving*, 2022.

[37] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision (ECCV)*. Springer, 2018.

[38] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] T. Stauner, F. Blank, M. Fürst, J. Günther, K. Hagn, P. Heidenreich, M. Huber, B. Knerr, T. Schulik, and K. Leiß, "Synpeds: A synthetic dataset for pedestrian detection in urban traffic scenes," in *Computer Science in Cars Symposium, CSCS*. ACM, 2022.

[40] L. Wang, G. W. Ding, R. Huang, Y. Cao, and Y. C. Lui, "Adversarial robustness of pruned neural networks," *Preprint*, 2018.

[41] Y. Guo, C. Zhang, C. Zhang, and Y. Chen, "Sparse dnns with improved adversarial robustness," *Advances in Neural Information Processing Systems (NIPS)*, vol. 31, 2018.

[42] K. Neklyudov, D. Molchanov, A. Ashukha, and D. P. Vetrov, "Structured bayesian pruning via log-normal multiplicative noise," *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017.

[43] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[44] A. Jordao and H. Pedrini, "On the effect of pruning on adversarial robustness," in *International Conference on Computer Vision (ICCV)*, 2021.

[45] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[46] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[47] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[48] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[49] J. T. C. Min and M. Motani, "Dropnet: reducing neural network complexity via iterative pruning," in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 9356–9366.

[50] M. Lin, R. Ji, Y. Wang, Y. Zhang, B. Zhang, Y. Tian, and L. Shao, "Hrank: Filter pruning using high-rank feature map," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[51] J.-H. Luo and J. Wu, "Neural network pruning with residual-connections and limited-data," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 1455–1464.

[52] A. Jordao, F. Yamada, and W. R. Schwartz, "Deep network compression based on partial least squares," *Neurocomputing*, 2020.

[53] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations (ICLR)*, 2019.

[54] N. Liao, S. Wang, L. Xiang, N. Ye, S. Shao, and P. Chu, "Achieving adversarial robustness via sparsity," *Machine Learning*, 2022.

[55] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[56] S. Gui, H. Wang, H. Yang, C. Yu, Z. Wang, and J. Liu, "Model compression with adversarial robustness: A unified optimization framework," *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[57] A. Sinha, H. Namkoong, and J. C. Duchi, "Certifying some distributional robustness with principled adversarial training," in *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018.

[58] S. Ye, K. Xu, S. Liu, H. Cheng, J.-H. Lambrechts, H. Zhang, A. Zhou, K. Ma, Y. Wang, and X. Lin, "Adversarial robustness vs. model compression, or both?" in *International Conference on Computer Vision (ICCV)*, 2019.

[59] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in Neural Information Processing Systems (NIPS)*, vol. 28, 2015.

[60] V. Sehwag, S. Wang, P. Mittal, and S. Jana, "Hydra: Pruning adversarially robust neural networks," *Advances in Neural Information Processing Systems (NIPS)*, 2020.

[61] T. Hu, T. Chen, H. Wang, and Z. Wang, "Triple wins: Boosting accuracy, robustness and efficiency together by enabling input-adaptive inference," in *International Conference on Learning Representations (ICLR)*, 2020.

[62] Z. Huang and N. Wang, "Data-driven sparse structure selection for deep neural networks," in *European Conference on Computer Vision (ECCV)*, 2018.

[63] T. Weng, H. Zhang, P. Chen, J. Yi, D. Su, Y. Gao, C. Hsieh, and L. Daniel, "Evaluating the robustness of neural networks: An extreme value theory approach," in *International Conference on Learning Representations (ICLR)*, 2018.

[64] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang, "Unlabeled data improves adversarial robustness," *Advances in Neural Information Processing Systems (NIPS)*, vol. 32, 2019.

[65] Y. Kaya, S. Hong, and T. Dumitras, "Shallow-deep networks: Understanding and mitigating network overthinking," in *International Conference on Machine Learning (ICML)*. PMLR, 2019.

[66] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *International Conference on Learning Representations (ICLR)*, 2019.

[67] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the value of network pruning," in *International Conference on Learning Representations (ICLR)*, 2019.