Adversarial Attacks and Defenses in Capsule Networks: A Critical Review of Robustness Challenges and Mitigation Strategies

Milind Shah[1(✉)], Kinjal Gandhi[2], Seema Joshi[3], Mudita Dave Nagar[4], Ved Patel[4], and Yash Patel[5]

[1] Department of Computer Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Changa 388421, Gujarat, India
milindshahcomputer@gmail.com

[2] Department of Computer Science and Engineering, School of Computer Science Engineering and Technology (SoCSET), ITM (SLS) Baroda University, Vadodara, Gujarat, India
kinjal445@gmail.com

[3] Graduate School of Engineering and Technology (GSET), Gujarat Technology University (GTU), Ahmedabad, Gujarat, India
ap_seema@gtu.edu.in

[4] Department of Computer Science & Engineering, School of Computer Science Engineering & Technology (SoCSET), ITM (SLS) Baroda University, Vadodara, Gujarat, India
mudita.nagar@gmail.com, ved17patel@gmail.com

[5] Department of Computer Science & Engineering with Cyber Security & Networking, School of Computer Science Engineering & Technology (SoCSET), ITM (SLS) Baroda University, Vadodara, Gujarat, India
yashp1734@gmail.com

Abstract. Capsule Networks (CapsNets) have gained significant attention in recent years due to their potential for improved representation learning and robustness. However, their vulnerability to adversarial attacks poses challenges for their deployment in safety-critical applications. This paper provides a critical review of the robustness challenges faced by CapsNets and explores various mitigation strategies proposed in the literature. The review includes an analysis of the adversarial attacks targeting CapsNets, such as manipulating primary capsule votes and direct targeting of CapsNets' votes. The computational cost of applying existing attack methods designed for Convolutional Neural Networks (CNNs) to CapsNets is also examined. To enhance the robustness of CapsNets, the incorporation of detection-aware attacks and innovative defense mechanisms is discussed. The effectiveness and efficiency of these defense strategies are evaluated through extensive experiments. The findings reveal the superiority of certain defense mechanisms in mitigating adversarial attacks on CapsNets. However, it is acknowledged that further research is needed to explore more robust attacks and approvals and to compare the robustness of CapsNets with CNNs. This critical review aims to provide insights into the current state of adversarial attacks and defenses in Capsule Networks, facilitating future research and development in this field.

Keywords: Neural Network · Capsule Networks · Cybersecurity · Adversarial Attacks · Defenses · Deep Learning · Evasion Attack

R. C. Bansal et al. (Eds.): ACTET 2023, CCIS 2000, pp. 11–33, 2024.
https://doi.org/10.1007/978-3-031-54162-9_2

1 Introduction

In recent times, advancements in machine learning and capsule neural networks have paved the way for tackling various practical challenges. These include but are not limited to tasks such as image classification, video analysis, text processing, and more.

Nevertheless, the susceptibility of the majority of contemporary machine learning classifiers to adversarial examples remains a critical concern. Adversarial examples refer to input data instances that have been intentionally modified to deceive a machine learning classifier. Often, these alterations can be imperceptible to human observers, yet the classifier still produces erroneous results.

Adversarial examples pose a security risk as they can be exploited to launch attacks on machine learning algorithms, even in cases where the adversary does not have direct access to the underlying model.

Furthermore, it has been observed that adversarial attacks are viable even when targeting machine learning algorithms that operate in real-world scenarios and rely on imperfect sensor inputs rather than precise digital data. It is important to note that the power and efficacy of machine learning and AI algorithms are expected to continue advancing in the future.

Exploiting vulnerabilities in machine learning security, similar to adversarial instances, can potentially lead to compromising and gaining control over highly powerful AIs. Therefore, ensuring robustness against adversarial instances is a crucial aspect of addressing the AI safety problem.

The field of adversarial attack and defense research presents several challenges. One of these challenges lies in evaluating potential attacks or defenses. Traditional machine learning approaches rely on training and test sets, where the performance is assessed based on the loss on the test set. However, in adversarial machine learning, defenders face the difficulty of dealing with inputs from an unknown distribution sent by attackers. Evaluating a defense against a single attack or a predetermined set of attacks is insufficient because a new attack can still bypass the defense. The complex nature of machine learning and capsule neural networks makes it challenging to conduct conceptual analysis, emphasizing the need for empirical proof of a defense's effectiveness. To address these challenges, competitions are often organized, pitting defenses against attacks developed by different teams. This competition-based evaluation, while not as conclusive as theoretical proof, simulates real-life security scenarios more effectively than a subjective review by the defense proposer [1].

Most of this research has focused on creating more robust models to defend against adversarial attacks if the input image is accurately categorized as the original class rather than the attacker's target class. Better defenses have led to stronger attack algorithms to break them. After multiple defensive creations and breaking iterations, some research concentrated on adversarial attack detection. Instead of classifying adversarial attacks as actual data, detection methods detect them. However, a defense-aware attack destroyed several state-of-the-art adversarial attack detection systems shortly after publication [2].

It is possible to predict adversarial samples markedly differently than clean samples, but the predictions are typically incomprehensible to humans. In various applications or under varying constraints, the model's susceptibility to adversarial attacks was discovered. It is possible to initiate adversarial attacks under various constraints, such as

assuming attackers have limited knowledge of target models, assuming a higher level of generalization for the attack, and imposing various real-world constraints on the attack. Given these developments, several concerns could be addressed. First, are these developments relatively independent of one another, or is there another perspective from which we can see their commonalities? Second, should adversarial samples be viewed as careless edge cases that can be resolved by applying patches to models, or are they embedded in the internal working mechanism of models that they are difficult to eliminate? [3].

To evaluate the efficiency of Adversarial Vector Loss (AVL), a series of black box attacks were conducted to analyze the resilience of both the standard Capsule network and AdvCapsNet. These networks were compared with commonly used vanilla neural networks on the CIFAR10 and Imagenette datasets. For thispurpose, we selected AlexNet, VGG, ResNet101, and DenseNet121. The AdvCapsNet model was trained exclusively using paired adversarial examples generated through the ResNet50-based FGSM attack. The attack had a magnitude of 0.3 and Lav set to 0.1, Conwayconsistent with the settings used in all of our experiments. Specifically, we perform two separate comparisons to analyze the performance of models under varying levels of attack intensity as well as the performance of different attacks with equal levels of intensity. Comparisons are performed on both datasets. In the first evaluation, we analyze the resilience of these models by utilizing FGSM, MI-FGSM, and PGD strategies, wherein the magnitude of noise varies from 0 to 0.5. In contrast, the findings indicate that our proposed model exhibits superior resilience against perturbations of greater magnitude when compared to both vanilla CNNs and Capsule networks. The main idea behind our implementation of adversarial regularization lies in its ability to promote the acquisition of an integrated representation by levying regularization on the optimization of model parameters. In the subsequent experiment, we analyze the resilience of various models in the presence of distinct attack models. The adversarial examples are generated by utilizing a consistent magnitude of $\varepsilon = 0.12$ for the FGSM, MI-FGSM, and PGD techniques, which are based on ResNet50, ResNet101, DenseNet121, VGG, and AlexNet models. The findings indicate that the success rate of attacks on AdvCapsNet is significantly lower compared to the vanilla models. This suggests that the parameters trained with AVL possess the ability to effectively defend against adversarial attacks from unfamiliar models. In conclusion, our experimental findings indicate that the AdvCapsNet we have proposed demonstrates greater resilience against adversarial attacks when compared to vanilla models. This is likely because our model encourages the learning of unchanged features in input images, thereby eliminating the impact of adversarial attacks [20].

2 Adversarial Attacks on Capsule Networks

Adversarial attacks are a type of attack used to convince machine learning models, such as capsule networks. Adversarial attacks operate by introducing small, imperceptible modifications to an initial image, which may result in misclassification by the model.Capsule networks are a form of neural network designed to acquire hierarchical object representations. It has been demonstrated that they are more resistant to adversarial attacks than ordinary neural networks, such as convolutional neural networks (CNN). Recent research has shown, however, that capsule networks keep more vulnerable to adversarial attacks.

Researchers have proposed a new technique for generating adversarial attacks that are designed to fool capsule networks. The researchers showed that their method effectively fooled capsule networks in a range of image classification tasks.

1) **Fast Gradient Sign Method (FGSM)** - Popular and straightforward adversarial attack technique FGSM computes the gradients of the loss function concerning the input and then disrupts the input in the direction of the sign of the gradients. This attack technique is also applicable to capsule networks.
2) **Basic Iterative Method (BIM) -** The BIM algorithm is superior to the FGSM algorithm. The BIM algorithm operates by adding little perturbations iteratively to the input image until the model is fooled.
3) **Projected Gradient Descent (PGD)** - The PGD algorithm is more effective than FGSM and BIM. The PGD algorithm operates by adding perturbations iteratively to the input image while simultaneously projecting the image back into the feasible area.
4) **One-Pixel Attack:** This attack focuses on modifying just a few pixels in the input image to cause misclassification. It searches for the most influential pixels and modifies their color values to deceive the model.
5) **Universal Adversarial Perturbations:** In this attack, a single perturbation is crafted to be applied to multiple input images, causing them to be misclassified. The perturbation is carefully calculated to be imperceptible to human observers but effective at fooling the model [25, 26].

The security of machine learning algorithms is severely compromised by adversarial attacks. Capsule networks are vulnerable to adversarial attacks, but are stronger than ordinary neural networks. However, additional research is required to develop more effective methods for protecting capsule networks from adversarial attacks.

The following areas of research are being analyzed for protecting capsule networks against adversarial attacks:

- ***Data Augmentation*** - Data augmentation is a method for making machine learning models less vulnerable to adversarial attacks. Data augmentation involves producing new data points that are comparable to the existing data points to augment the size of the training dataset.
- ***Robust Optimization*** - Robust optimization is a method for training machine learning models that are more resistant to adversarial attacks. Robust optimization algorithms are intended to find solutions that are insensitive to minor changes in the input data.
- ***Adversarial Training -*** Adversarial training is a technique that uses adversarial instances to train a machine learning model. Training against adversarial data can make machine learning models more resistant to adversarial attacks.
- ***Capsule Routing Security:*** Capsule networks rely on dynamic routing algorithms to determine the instantiation parameters of capsules. By introducing additional security measures into the routing process, such as limiting the number of routing iterations or applying noise, the capsule network can become more resilient against adversarial attacks [25, 26].

The analysis of adversarial attacks is still in its earliest stages. However, prior research indicates that adversarial attacks represent a significant risk to the security of

machine learning models. New methods for protecting machine learning models against adversarial attacks require additional research [6].

3 What are the Potential Consequences of Adversarial Attacks on Machine Learning Algorithms?

This section will cover the potential consequences of adversarial attacks on machine learning algorithms. Specific attack strategies will be utilized based on various application scenarios, conditions, and the capabilities of adversaries.

3.1 Untargeted vs Targeted Attack

The classification of threat models can be identified into two categories: targeted and untargeted, based on the objectives implemented by attackers. In the context of targeted attacks, the objective is to intentionally manipulate a model's prediction to direct it towards a predetermined class, concerning a given instance. The objective of an untargeted attack is to inhibit a model's ability to assign a particular label to a given instance. In certain situations, the two previous types of attack are alternatively referred to as the false positive attack and the false negative attack. The primary objective of the first approach is to encourage models to incorrectly classify negative instances as positive, whereas the latter aims to mislead models into classifying positive instances as negative. The terms "false positive attack" and "false negative attack" are occasionally referred to as Type-I attack and Type-II attack, respectively [3].

3.2 One Shot vs Iterative Attack

Based on practical limitations, adversaries can launch either one-shot or iterative attacks to target models. The one-shot attack method allows for the generation of adversarial samples in a single attempt, providing a single chance to achieve the desired outcome. On the other hand, the iterative attack approach allows for multiple steps to be taken to explore and identify a more optimal direction for generating adversarial samples. The utilization of an iterative attack has been found to generate adversarial samples that are more effective in comparison to a one-shot attack. Nevertheless, this approach requires a greater number of queries to the target model and involves additional computational resources to initiate each attack. Consequently, its practicality may be constrained in computational-intensive tasks [3].

3.3 White Box and Black Box Attack

In the context of white-box attacks, it is believed that attackers demonstrate comprehensive knowledge regarding the target model. This knowledge encompasses various aspects such as the model's architecture, weights, hyper-parameters, and potentially even the training data. The utilization of white-box attacks facilitates the detection of related vulnerabilities within the target model. In ideal circumstances, this scenario represents the most challenging situation that defenses may encounter. The black-box attack

methodology operates under the assumption that attackers have the same level of access to the model's output as regular end users. This assumption holds greater practicality in real-world scenarios. Despite the lack of comprehensive information regarding models, the black-box attack remains a significant concern for machine learning systems. This is primarily due to the transferability property demonstrated by adversarial samples [3].

4 Research Questions

This review paper discusses the following research questions.

Q1: What are the different types of adversarial attacks that can be used against capsule networks?
Q2:What are the existing research limitations in adversarial attacks and defenses?
Q3: What are the open challenges and future directions in adversarial attacks and defenses?
Q4: What are the future research directions for improving the robustness of capsule networks to adversarial attacks?
Q5: Can the robustness of capsule networks against adversarial attacks be improved by combining multiple defense mechanisms, such as adversarial training, input transformation, and ensemble methods?

5 Review of Literature

Till now, a lot of research has been done to solve the challenges of adversarial attacks using capsule neural networks.

In [1] Alexey Kurakin et al., Google Brain organized a competition at NIPS 2017 to encourage the development of innovative strategies to create and defend against adversarial examples. The primary objective of the competition was to expedite research on adversarial instances and enhance the robustness of machine learning classifiers. This chapter provides an overview of the competition's format, organization, and solutions devised by top-ranking teams. The competition sought to raise awareness of the issue and inspire scholars to devise original approaches. Participants were challenged to explore novel techniques and enhance existing solutions through competitive engagement. The competition results showcased significant progress made by all three tracks compared to the baselines established. Notably, the winning entry in the defense tracking competition achieved an impressive 95% accuracy in classifying all threatening images generated by different attacks. These findings suggest that practical applications can attain a satisfactory level of resilience against adversarial cases, even though the worst-case accuracy did not match the exceptional average accuracy achieved.

In [2] Yao Qin et al., Present a novel method for breaking out of this loop, one in which adversarial attacks are "deflected" by forcing the attacker to provide input that semantically resembles the class that is the focus of the attack. This would put an end to the cycle. We propose a more robust defense based on Capsule Networks that integrates three detection algorithms to provide state-of-the-art detection performance against both conventional and defense-aware attacks. This can be accomplished by achieving state-of-the-art detection against both types of attacks. After that, we show that undetected

attacks against our defense frequently appear perceptually the same as the opposed target class by having human participants label images that were created by the attack. The term "adversarial" can no longer be used to describe these attack pictures since our network classifies them in a manner that is comparable to how humans do. As a first step toward putting an end to the conflict between defenses and attacks, you should implement a novel method that can redirect impacts from your adversaries. We offer an innovative cycle consistency loss to drive the winning capsule reconstruction of the CapsNet to closely resemble the class-conditional distribution. This was done to improve accuracy. We can identify common adversarial attacks on SVHN and CIFAR-10 with a low False Positive Rate since we have three detection algorithms and three independent distance measurements at our disposal. We present a defense-aware attack as a means of explicitly attacking our detection measures, and we discover that our model achieves considerably lower undetected attack rates than the most cutting-edge approaches currently available for defense-aware attacks. In addition, a significant percentage of attacks that go undetected are redirected by our model in such a way that they take on the characteristics of the adversarial target class but do not succeed in becoming malicious. An analysis conducted by humans reveals that 70% of undiscovered black-box adversarial attacks are uniformly identified as the target class on SVHN. This was discovered as a result of the inquiry.

In [3] Ninghao Liu et al., This paper aims to analyze current research related to adversarial attacks and defenses, with a particular focus on the interpretation of machine learning. The process of interpretation can be categorized into two distinct types: interpretation at the feature level and interpretation at the model level. In the context of adversarial attacks and defenses, we provide an analysis of the potential applications of each interpretation method. Next, we will briefly elucidate additional connections between interpretation and adversaries. In conclusion, we will now analyze the challenges and possible methods related to the resolution of adversary concerns via the process of interpretation. In the analysis, we analyzed the potential applications of the interpretation within each category, specifically focusing on its utility in initiating adversarial attacks or formulating defensive strategies. Subsequently, we will look into further clarification of the interrelationships between interpretation and adversarial samples or robustness. In conclusion, the present discussion is related to the current challenges encountered in the process of constructing transparent and resilient models, alongside potential avenues for leveraging adversarial samples in forthcoming activities. Future research directions include the development of models with enhanced explainability, the exploration of adversarial attacks in real-world scenarios, and the enhancement of models through the utilization of adversarial samples.

In [4] Alberto Marchisio et al., Perform research to establish the level to which CapsNets is vulnerable to attacks from adversaries. These alterations, which are introduced as test inputs, are so small that human beings are unable to recognize them; however, they are capable of fooling the network into generating inaccurate predictions. We present a greedy technique as a means of automatically producing adversarial samples that cannot be detected in the context of a black-box attack. We show that such attacks, when applied to the German Traffic Sign Recognition Benchmark and CIFAR10 datasets, can

lead CapsNets into producing wrong classifications. This can be problematic for intelligent CPS, such as autonomous vehicles, which need accurate classifications to function well. In addition, we apply the identical adversarial attacks to a 5-layer CNN (LeNet), a 9-layer CNN (VGGNet), and a 20-layer CNN (ResNet), and then compare the findings to those of the CapsNetsto analyze the different ways in which the CapsNets react to the same adversarial attacks. In conclusion, the findings of this research show that the resilience of the CapsNet is equivalent to that of a CNN that is significantly deeper, such as the VGGNet. On the other hand, the LeNet is noticeably more vulnerable to linear transformations and adversarial attacks, and the robustness of the DeepCaps is greater than that of the ResNet. Therefore, we can make substantial progress in the protection of safety-critical applications by leveraging deep and complex networks, such as DeepCaps. To increase its robustness, it would be advantageous to make further improvements to the CapsNet algorithm to boost prediction accuracy. In this regard, the DeepCaps architecture appears to be more secure than the ResNet under comparable attack circumstances.

In [5] Richard Osuala et al., Highlight several unexplored solutions for analysis. A meta-analytic methodology called SynTRUST evaluates medical image synthesis study validation accuracy. 26 concrete completeness, reproducibility, usefulness, scalability, and durability metrics support SynTRUST. SynTRUST validates sixteen of the most promising cancer imaging challenge solutions and finds many enhancements. This effort aims to connect the clinical cancer imaging group's demands to the artificial intelligence group's data synthesis and adversarial network research. Finally, GANs' adversarial learning is flexible and modality-independent. This survey lists numerous cancer imaging difficulties that adversarial networks can handle. Unsupervised domain adaptation, patient privacy-preserving distributed data synthesis, adversarial segmentation mask discrimination, and multi-modal radiation dosage estimation are GAN/adversarial training solutions. Before considering GAN and adversarial training, we analyzed research on cancer imaging challenges in radiology and non-radiology techniques. After screening and analysis of cancer imaging issues, we categorized them into Data Scarcity and Usability, Data Access and Privacy, Data Annotation and Segmentation, Detection and Diagnosis, and Treatment and Monitoring. We found 164 relevant publications on adversarial networks in cancer imaging and categorized them by cancer imaging challenge. Finally, we analyze each challenge and GAN-related papers to analyze if GANs and adversarial training can solve it. Improving SynTRUST for medical image synthesis research dependability. SynTRUST evaluates 16 well-chosen cancer imaging challenge solutions. Despite these findings' rigor and validity, we may recommend trustworthy improvements for future research. We also recommend data synthesis and adversarial training techniques for challenges that the literature has not addressed.

In [6] Alberto Marchisio et al., Analyze Capsule Networks' vulnerability to dangerous attacks. These test input issues are invisible to humans but can fool the network into producing inaccurate predictions. A greedy algorithm generates targeted, undetected adversarial instances automatically in a black-box attack scenario. When launched against the German Traffic Sign Recognition Benchmark (GTSRB), similar attacks might deceive Capsule Networks. We also apply adversarial attacks on 5-layer

and 9-layer CNNs and compare their behavior to Capsule Networks. This research develops a unique method to autonomously produce focused, undetectable, and robust threatening cases and compares CapsuleNet, a 5-layer CNN, and a 9-layer CNN under these adversarial instances. Finally, they developed a black box adversarial attack technique. Using the GTSRB dataset, we tested our approach against CapsuleNets, 5-layer LeNets, and 9-layer VGGNets. Our findings show that the CapsuleNet resists attack better than the LeNet but less than the VGGNet. Our approach makes traffic signal pixel alterations less obvious in the CapsuleNet than in the VGGNet. CapsuleNet output probabilities are less than VGGNet predictions. CapsuleNet output probabilities fluctuate less than VGGNet output probabilities. Adding prediction confidence to the CapsuleNet technique might improve its resilience.

In [7] Muhammad Shafique et al., In both the cloud environment during the ML training phase and at the peripheral during the ML inference phase, this study presents viable defenses and strategies to overcome these vulnerabilities. This chapter examines the effects of a resource-constrained design on system reliability and security. It defines verification methods to ensure accurate system behavior and outlines unresolved research issues in building secure and dependable Machine Learning (ML) algorithms for both edge computing and cloud platforms. This review covers three main aspects: 1) the significant security and reliability issues faced by machine learning algorithms, 2) the measures taken to safeguard these systems, and 3) the formal technique employed to validate specific neural networks (NNs). The research also includes a summary of the major challenges that currently hinder the development of effective machine-learning algorithms.

In [8] Jindong Gu et al., Analyze the reliability of CapsNets under adversarial conditions, specifically focusing on how the internal processes of CapsNets are affected when the output containers are targeted. Initially, adversarial instances manipulate the primary capsule votes to deceive CapsNets. However, due to the computationally intensive routing mechanism, applying multi-step attack methods developed for CNNs to target CapsNets results in a high computational cost. Motivated by these observations, we propose an innovative vote attack that directly aims at the votes of CapsNets. By bypassing the routing procedure, our vote attack is both effective and efficient. Furthermore, we integrate our vote attack into the detection-aware attack paradigm, which effectively evades the class-conditional reconstruction-based detection method. Extensive experiments confirm that our vote attack on CapsNets outperforms other attack methods. Although CapsNets exhibit higher resistance to our stronger Vote-Attack compared to CNNs, it is premature to conclude that CapsNets are less vulnerable. We assume that the robust accuracy of CapsNets can still be further reduced. Future research will explore more robust attacks and validations to compare the resilience of CNNs and CapsNets.

In [9] Boxi Wu et al., This research demonstrates that adversarial attacks can be disrupted by small disturbances. Even a slight random noise added to adversarial instances can render their incorrect predictions invalid, even on models that have been trained to defend against adversarial attacks. This vulnerability was found in all state-of-the-art attack methods. Building upon this observation, we propose more effective defensive disturbances to counteract attackers. Our defensive disturbances employ adversarial training to decrease the local Lipschitzness in the ground-truth class. By targeting all

classes simultaneously, we can rectify incorrect predictions that have higher Lipschitzness. Empirical and theoretical evaluations of linear models validate the effectiveness of our defensive perturbation. CIFAR10 enhances the performance of the state-of-the-art model from 66.16% to 72.66% against four AutoAttack methods, including a boost from 71.76% to 83.30% against the Square attack. Additionally, employing a 100-step PGD approach improves FastAT's top-1 robust accuracy on ImageNet from 33.18% to 38.54%. This work makes two contributions: 1) It reveals that adversarial attacks can be disrupted, and 2) inspired by this finding, we introduce Hedge Defense as a more effective means to counter attacks and enhance adversarial-trained models. Both empirical and theoretical findings provide evidence for the efficacy of our technique. Our work not only attracts attacks using the same technique but also sheds light on new defense strategies. Further research could explore alternative criteria for selecting specific predictions rather than targeting all classes. With Hedge Defense, defenders may not need to ensure that the model can correctly classify all local cases; instead, they can focus on meeting specific requirements, such as reducing the local Lipschitzness in the ground-truth class, to identify better scenarios.

In [10] Abhijith Sharma et al., In this survey, we present a comprehensive overview of existing techniques employed in adversarial patch attacks. We aim to enable researchers interested in this field to quickly familiarize themselves with the latest advancements. Additionally, we discuss the current methods used for detecting and defending against adversarial patches. This serves to enhance the community's understanding of this discipline and its practical applications. In conclusion, we offer a clear and in-depth analysis of adversarial patch attacks and defenses in the context of vision-based tasks, providing readers with insights into their strengths and limitations. While challenges such as scalability and real-time capabilities persist, it is noteworthy that most research in adversarial patch attacks focuses on classification and object detection. Exploring the application of adversarial patch attacks in language or translation models could be intriguing. Considering the lack of explainability in DNN-based black box models, could adversarial patch attacks offer a new perspective on model predictions? If so, could they contribute to the development of more robust real-world models? We are enthusiastic about investigating these issues and supporting future solutions to advance this field and benefit society.

In [11] Jindong Gu et al., This paper introducesSegPGD, an impactful attack technique specifically designed for segmentation models. Through convergence analysis, we demonstrate that SegPGD generates more potent adversarial instances compared to PGD, even when both methods employ the same number of attack iterations. We recommend incorporating SegPGD into segmentation adversarial training as it produces more effective adversarial examples, ultimately enhancing the resilience of segmentation models. Our proposals are validated through experiments conducted on widely used segmentation model architectures and standard datasets. However, it is worth noting that further exploration of segmentation adversarial training methods may lead to even more effective and efficient approaches. This research serves as a foundation for future endeavors aimed at improving the robustness of segmentation models.

In [12] Alberto Marchisio et al., By conducting a systematic analysis and evaluation, we compare CapsNets to traditional Convolutional Neural Networks (CNNs) and investigate the factors influencing the robustness of CapsNets. In this comprehensive

comparison, we examine two CapsNet models and two CNN models across various datasets, including MNIST, GTSRB, CIFAR10, and their affine-transformed counterparts. Through this extensive analysis, we identify the key properties that contribute to the enhancement of robustness in these architectures, as well as their limitations. Generally, CapsNets exhibit greater resistance to both adversarial examples and affine transformations compared to CNNs with an equal number of parameters. Similar conclusions hold when comparing CapsNets and CNNs with increased depth. Surprisingly, our findings indicate that dynamic routing, a distinguishing feature of CapsNets, does not significantly improve their robustness. Instead, the capsule-based hierarchical feature learning within CapsNets plays a primary role in generalization. In summary, this paper introduces a method for analyzing the resilience of CapsNets against affine transformations and adversarial attacks. We examine the differences between CapsNets and CNNs in terms of improving robustness. ShallowCaps, despite requiring a significant number of parameters, exhibit superior resistance to adversarial attacks but struggle to generalize well on complex datasets. They also demonstrate better resistance to adversarial attacks compared to affine transformations. However, the DeepCaps model, despite having fewer parameters, mitigates the disparity between transformed and untransformed datasets. In MNIST classification, DeepCaps shows lower resilience to adversarial attacks compared to ShallowCaps. On the CIFAR10 dataset, they outperform a CNN with a similar architecture and the ResNet20 model. The resilience of DeepCaps is further enhanced through adversarial training. When considering the affCIFAR dataset, DeepCaps outperforms ResNet20 in terms of handling affine modifications. Our results indicate that dynamic routing does not significantly enhance the robustness of CapsNets. This comprehensive study provides valuable insights for future CapsNet designs in addressing safety-critical applications by considering potential attackers, as well as opening up avenues for exploring new adversarial attacks.

In [13] Bader Rasheed et al., This paper presents a novel approach called multiple adversarial domain adaptation (MADA) that tackles the problem of adversarial attacks by treating it as a domain adaptation task to identify resilient features. Our method utilizes adversarial learning to discover domain-invariant features across multiple adversarial domains and a clean domain. To evaluate the effectiveness of MADA, we conducted experiments on the MNIST and CIFAR-10 datasets using various adversarial attacks during both the training and testing phases. The results demonstrate that MADA outperforms adversarial training (AT) by an average of 4% on adversarial samples and 1% on clean samples. The objective of this paper is to enhance the generalization of adversarial training on both adversarial and clean samples by formulating the problem as a multiple-domain adaptation task, with adversarial domains representing the target domains. Our work introduces a domain adaptation-based strategy to enhance adversarial training specifically for adversarial data. By aligning the distributions of adversarial domains with the clean distribution in the feature embedding space, we effectively reduce the impact of adversarial attacks. This approach not only improves the interpretability of features in the embedding space but also enhances model generalization in adversarial environments. Furthermore, instead of relying solely on the Wasserstein distance, alternative methods for aligning distributions could be explored in future research.

In [14] Junjie Mao et al., This paper aims to evaluate the security and robustness of existing face antispoofing models, particularly multimodality models, against various types of attacks. The study focuses on assessing the resilience of multimodality models to both white-box and black-box attacks, specifically targeting adversarial examples. To enhance the security of these models, a novel approach is proposed, which combines mixed adversarial training with differentiable high-frequency suppression modules. Experimental results reveal that when exposed to adversarial examples, the accuracy of a multimodality face antispoofing model decreases significantly from over 90% to approximately 10%. However, the suggested defense method successfully maintains an accuracy of over 80% on attack examples and over 90% on original examples. The research includes an analysis of advanced single-modality and multimodality face antispoofing models, evaluating their susceptibility to white-box and black-box attacks using RGB, Depth, and IR images. The evaluation encompasses attacks on a multimodality model with a single-input stream, and the results demonstrate the model's resilience against attacks focused on a single modality in experimental scenarios. Additionally, the security of single-modality and multimodality models against various patch attacks is examined. By incorporating hybrid adversarial training and diffusible high-frequency suppression modules, the security of both single-modality and multimodality models is enhanced. Experimental outcomes highlight that multimodality models offer superior security compared to single-modality models. Furthermore, this paper presents the first proposal for adversarial attack research on multimodality models.

In [15] Lin Hiu et al., In this study, we presented an initial attack model called the AMR technique, which achieves high recognition accuracy. Moreover, we proposed a transferable attack technique that utilizes feature gradients to increase signal disruption in the feature space. Additionally, we introduced a novel attack strategy that employs two original signal samples and one adversarial target signal sample as inputs for the triplet loss, aiming to achieve stronger attack effectiveness and greater transferability. To evaluate the efficacy of our proposed attack technique, we introduced signal-characteristic indicators. Our feature gradient-based adversarial attack technique surpasses existing gradient attack methods in terms of attack effectiveness and transferability. The main contribution of this research lies in the introduction of a transferable attentive technique that focuses on informative and discriminative feature regions, introducing disruption at the feature level to mimic more realistic adversarial scenarios. We conducted comprehensive experiments using a new indicator system that aligns better with signal characteristics, and most of the indicators outperformed the label gradient approach. We propose two novel approaches, AL-BIM and AL-MIM, which optimize the triplet loss for performing regional attacks on stable features extracted from AMR signals. Our methods surpass label-based adversarial attack techniques in terms of effectiveness. Experimental results on public datasets demonstrate that our feature gradient-based attack method outperforms label gradient-based methods in both black-box and white-box attack scenarios, achieving higher attack success rates and improved transferability. Furthermore, the disruptions caused by our feature gradient-based attacks are smoother and less noticeable. To quantify signal distortion and migration rate, we introduced four signal character

indicators (ACR, APD, PSR, TR), which outperform previous attack techniques. Additionally, we explored techniques to minimize attack disruption and restrict the impact of the attack.

6 Methodological Comparison

See Table 1.

Table 1. Comparative Analysis

Author Name	Publication with Year	Techniques Used	Dataset Used	Accuracy	Technologies Used	Findings
Xu Han et al. [16]	Wiley Hindawi, 2022	Natural Language Processing (NLP), Deep Neural Network (DNN)	–		Neural network	Text attacks. Adversarial scenarios can inform backdoor attacks, robustness testing, and defense. Readability depends on the objective. Attacks require sophistication. DNN applications will increase the robustness of research
Xiaopeng Fu et al. [17]	Wiley Hindawi, 2021	Visual Similar Word Replacement Algorithm (VSWA)	Yelp Review Dataset and Amazon Review Dataset	Bi-LSTM has 95.64% accuracy and LSTM 95.69 for Yelp Review. For Amazon Review Dataset, LSTM has 88.48% accuracy and BiLSTM 88.55%	Python, LSTM & Bi-LSTM	Utilize the VSWR methodology to generate adversarial instances on datasets utilized for sentiment analysis, thereby launching attacks on pre-trained deep learning classification models
Heng Yin et al. [18]	Wiley Hindawi, 2021	Adam Iterative Fast Gradient	NIPS 2017 Adverarial Competition	95%	Python	In black-box circumstances, including adversarial trained networks, the gradient-based method is superior to gradient-based alternatives We also targeted an ensemble of networks with novel adversarial example transferability strategies

(*continued*)

Table 1. (*continued*)

Author Name	Publication with Year	Techniques Used	Dataset Used	Accuracy	Technologies Used	Findings
Murali Krishna Puttagunta et al. [19]	Springer, 2023	Deep Learning Models	MNIST and CIFAR-10		Python	To propose strong medical deep learning implementation decisions. Finally, this paper lists some unsolved research issues that need more research
Yueqiao Li et al. [20]	Elsevier, 2021	AdvCapsNet	CIFAR10	64.14%	Python	To analyze Capsule networks and other basic CNNs against more complicated transfer attacks on two interesting datasets. Offer an AdvCapsNet with AVL for adversarial attack threats. The Capsule network's unified efficiency framework might incorporate the extra loss with regularization losses
Taeyoung Hahn et al. [21]	NeurIPS Proceedings, 2019	Self Routing	CIFAR10, SVHN &SmallNORB	–	Python	Systematic evaluations of our self-routing. Our technique is outstanding at adversarial defense and perspective generalization, CapsNets' strengths. Our technology works better with more capsules per layer than older, inaccurate techniques. CapsNet may not need routing by agreement. Finding a mechanism to add residual connections to our models is interesting because residual networks operate as ensembles of networks with various depths. Our capsules are synergetic

(*continued*)

Table 1. (*continued*)

Author Name	Publication with Year	Techniques Used	Dataset Used	Accuracy	Technologies Used	Findings
Alberto Marchisio et al. [22]	IEEE Access, 2022	Neural Architecture Search algorithm	CIFAR10	86.07%	ROHNAS Framework	Analytical models of DNN and CapsNet layers, activities, visualization, and execution on specialized processors allow architectural modeling and quick hardware estimation. We analyze and select adversarial perturbations to speed up NAS (Neural Architecture Search) robustness evaluation with DNNs. These perturbations highlight DNN discrepancies under adversarial scenarios. We use the Non-dominated Sorting Genetic Algorithm II (NSGA-II) to create an evolutionary algorithm. This technique optimizes DNN adversarial resistance, energy efficiency, memory consumption, and latency via multi-objective Pareto-frontier selection

7 How Do Researchers Evaluate Potential Attacks or Defenses for Adversarial Machine Learning?

This section briefly introduces the basic idea of different defense strategies against adversaries.

7.1 Input Denoising

Adversarial perturbation refers to the introduction of imperceptible noise into data. To prevent this issue, a potential solution is to utilize filtering techniques or incorporate random transformations to counteract the effects of adversarial noise. It is noteworthy that the inclusion of f_x can occur either before the model input layer or as an internal component within the target model.

In the context of the former scenario, where an instance z* is potentially subject to adversarial influence, our objective is to develop a mapping f_x that satisfies the condition $f(f_x(z^*)) = f(z_0)$. In the latter scenario, the concept remains comparable, with the exception that the function f is substituted by the output h of a specific intermediate layer [3].

7.2 Model Robustification

Another commonly utilized strategy is to improve the model's preparation against potential threats from adversaries. There are two potential approaches to improving the model's refinement: altering the training objective or adjusting the model structure. Examples of previous approaches include using adversarial training and substituting real-world training loss with robust training loss. The underlying rationale is to proactively address the potential impact of adversarial samples during a model's training, thereby improving the model's resilience. Instances of model modification encompass various techniques such as model distillation, the implementation of layer discretization, and the regulation of neuron activations. In a formal context, let f^l represent the robust model. The objective is to ensure that $f^l(z^*) = f^l(z_0) = y$ [3].

7.3 Adversarial Detection

In contrast to the previous two approaches that attempt to determine the accurate label of a given instance, adversarial detection focuses on determining whether the given instance has been infected by adversarial perturbation. The primary objective is to construct an additional predictor, denoted as *fd*, which assigns a value of *1 to x* if it has been infected and a value of *0* otherwise. The process of establishing *fd* can be conducted using the conventional approach of constructing a binary classifier. Input denoising and model robustification methods are utilized to prevent the effects of external influences on the accuracy of correction predictions. The adversarial attack involves manipulating the input data and model architectures to achieve the desired outcome. Adversarial detection methods utilize a reactive approach to determine whether the model should proceed with making predictions. To avoid being manipulated, one should be suspicious of the information provided. The implementation of proactive strategies typically presents greater challenges compared to reactive strategies [3].

8 Existing Defense Mechanisms in Adversarial Attack

Existing defense mechanisms have been designed to mitigate the effects of adversarial attacks. These techniques want to improve the robustness of deep learning models, such as capsule networks, against adversarial instances. Here are some common defensive techniques:

1) **Adversarial Training** - Adversarial training is a defensive approach that involves incorporating adversarial instances into the training data. During training, by introducing the model to adversarial disturbances, the model becomes more robust and resistant to such attacks. Adversarial training can enhance the model's accuracy in adversarial instances, but its performance on pure examples may suffer as a result.

2) **Defensive Distillation** - Training a model on reduced logs rather than precise class designations constitutes defensive distillation. Initially, the logs are altered using a temperature parameter, which eliminates the decision boundaries and reduces the model's sensitivity to minor disturbances. This technique has been demonstrated to offer some protection against adversary attacks.
3) **Gradient Masking**- Gradient masking involves obfuscating or concealing the gradients of the model to make it more difficult for adversaries to generate effective adversarial examples. This can be achieved by introducing noise or disturbances into the gradients during backpropagation. Recent research has shown, however, that gradient masking only is not an effective defense.
4) **Ensemble Defense** - Ensemble methods integrate the predictions of multiple models to make more robust decisions. By training multiple models with distinct architectures or random initialization, the ensemble can capture diverse perspectives and mitigate the effects of adversarial attacks. It is less likely that adversarial examples that fool one model will fool the entire ensemble.
5) **Certified Defenses -** Certified defenses provide formal assurances regarding the model's resistance to adversarial attacks. These techniques utilize mathematical checks or bounds to ensure that the model's predictions are robust over a certain range of disturbances. Certified defenses offer more robust guarantees, but they frequently involve additional computational costs.
6) **Input Preprocessing** - Applying input preprocessing techniques, such as input normalization or denoising, can help make the model more resilient to adversarial perturbations. These techniques can reduce the effectiveness of small changes introduced by attackers, making it more difficult to deceive the model [25, 26].
7) **Adversarial Detection and Filtering** - Implementing mechanisms to detect and filter adversarial inputs can help identify potential attacks and prevent them from influencing the model's decisions. This can involve monitoring input data for characteristics indicative of adversarial examples and rejecting or flagging suspicious samples [25, 26].

While these defense mechanisms can provide some protection against adversarial attacks, they may not be universally efficient or applicable in all circumstances. The evolution of adversarial attacks and defense strategies is an ongoing research topic, as is the development of more robust and reliable defense mechanisms against adversarial instances.

9 Existing Research Limitations

Existing research has mainly focused on a limited number of capsule networks, which is one of its primary limitations. Therefore, it is unknown how well the results of this research apply to other capsule networks. Moreover, the majority of research in this field has been conducted with relatively smaller datasets. This makes it challenging to evaluate the resilience of capsule networks against adversarial attacks on large datasets.

Existing research has also been limited by its focus on a relatively small amount of adversarial attacks. Therefore, it is unknown how well the findings of this research apply to other adversarial attacks. In addition, the majority of research in this field has focused

on relatively straightforward adversarial attacks. This makes it difficult to evaluate the resilience of capsule networks against more sophisticated adversary attacks.

The majority of research in this field has been focused on developing defenses against adversarial attacks. However, research into the development of methods for identifying adversarial attacks is missing. This is a crucial area of research, as it is possible to create defenses that are efficient against some adversarial attacks but vulnerable to others.

10 How Effective are the Current Defenses Against Adversarial Attacks on Machine Learning Algorithms?

The concerns arising from adversarial attacks are related to the reduction of confidence in the true output class and the possibility of misclassification. The strategies utilized to counter adversarial attacks typically aim to achieve one of two objectives: 1) enhance the ability to be detected the attack, ensuring that clean and malicious inputs can be visually differentiated, or 2) improve the resilience of the deep neural network (DNN) against the attack, thereby minimizing its impact. One potential defense strategy against evasion-based adversarial attacks that are developed using input gradients is to utilize a technique known as gradient masking, which involves minimizing these gradients. The utilization of this technique results in a decrease in the reliability of output classification through the process of retraining the deep neural network (DNN) using the output probability vector. Adversarial training is a frequently utilized defense mechanism in which a trained deep neural network (DNN) undergoes training using adversarial inputs alongside their corresponding correct output labels. This improvement enhances the precision of the system when dealing with a recognized attack.An additional method utilized in the majority of practical machine learning (ML) systems involves the implementation of input pre-processing. The defense mechanism utilized in this scenario involves the process of smoothing, transforming, and reducing the noise before its input into the deep neural network (DNN). This defensive measure reduces adversarial noise, thereby decreasing the likelihood of a successful attack [7].

Adversarial training, a highly effective defense strategy, was proposed as a means for reducing the vulnerability of classification models. This strategy involves the creation and injection of adversarial examples into the training data during the training process. An effective strategy to improve the resilience of segmentation models is the implementation of adversarial training techniques. However, the process of generating efficient segmentation adversarial examples during the training phase can be a time-intensive effort. In this research, we provide proof that shows our SegPGD method is both effective and efficient in addressing this particular challenge. The utilization of SegPGD as the underlying attack method in adversarial training has been found to significantly improve the resilience of segmentation models by generating significant adversarial examples. It is noteworthy to mention that multiple strategies utilizing single-step attacks have been proposed in the context of adversarial training, aiming to address the efficiency of adversarial training in the field of classification. However, single-step attacks do not effectively mislead segmentation models as the adversarial examples they generate are not sufficiently significant [11].

At present, several defense strategies that have been proven effective in countering black-box and gray-box attacks are vulnerable to adaptive white-box attacks. In the 2018 International Conference on Learning Representations (ICLR2018), it was observed that a majority of the heuristic defenses, specifically seven out of nine, were found to be compromised by the adaptive white-box attacks. The application of adversarial attack algorithms, such as Projected Gradient Descent (PGD) and Carlini and Wagner (C&W), to the physical world presents two significant challenges that must be addressed, despite the proven efficiency of these algorithms in the digital domain. One primary challenge is the potential disruption caused by environmental noise and natural transformations, which can compromise the integrity of adversarial perturbations computed in the digital world. The second challenge is related specifically to machine learning tasks that involve images and videos. In these tasks, only the pixels that correspond to specific objects can be altered in the physical world [23].

11 How Do Capsule Neural Networks Differ from Other Types of Neural Networks in Their Susceptibility to Adversarial Attacks? Are There any Current Solutions or Defenses Against Adversarial Attacks on Machine Learning Algorithms?

Capsule Networks can maintain hierarchical spatial relationships among objects, which suggests the possibility of outperforming traditional Convolutional Neural Networks (CNNs) in tasks such as image classification [6].

Convolutional Neural Networks (CNNs) commonly indicate vulnerability to small quasi-imperceptible artificial perturbations, resulting in their vulnerability to being deceived. The vulnerability of convolutional neural networks (CNNs) can present possible risks to applications that prioritize security, such as face verification and autonomous driving. Moreover, the presence of adversarial images serves as evidence that the object recognition mechanism utilized by Convolutional Neural Networks (CNNs) differs significantly from that observed in the human brain. Therefore, there has been a growing interest in adversarial examples since their release [8].

Convolutional neural networks (CNNs) have demonstrated remarkable performance in various domains, emerging as the leading approach. However, recent research revealed a vulnerability in these models, revealing their vulnerability to adversarial perturbations. The presence of gradient calculation instability can contribute to the enhancement of this phenomenon across multiple layers within the network. Nevertheless, it is widely acknowledged that deep neural networks are vulnerable to adversarial inputs, which appear as minimal perturbations introduced to images that are unnoticeable by human observers. Adversarial noise has the potential to deceive convolutional neural networks (CNNs) and other types of neural network architectures, resulting in these models producing inaccurate predictions with a significant level of certainty. The presence of adversarial attacks implements constraints on the utilization of neural networks in tasks that are crucial for security. One possible reason for the efficiency of adversarial samples is that Convolutional Neural Networks (CNNs) show a high degree of linearity within feature spaces of significant dimensionality. While Convolutional Neural Networks (CNNs) can transform feature vectors using non-linear functions, it has been observed that basic

activation functions like softmax lack the necessary level of non-linearity to effectively counter adversarial attacks. In contrast, it is worth noting that a Capsule network can generate significantly more complex non-linearities, thereby reducing the vulnerability to adversarial attacks. To address this issue, we present a novel AdvCapsNet model based on Capsule and incorporating a considerably more complicated non-linear function. This model aims to provide robust protection against adversarial attacks [20].

The Capsule network utilizes a dynamic routing mechanism to acquire knowledge about the constituent elements that constitute a particular entity in its entirety. In contrast to deep neural networks, which are limited to modeling local feature knowledge, Capsule networks show the ability to not only model knowledge about local features but also simulate their relationships. Hence, it can be shown that Capsule networks are more effectively designed for image processing, thereby showing superior performance in tasks such as image classification and other related activities. When analyzing the robustness of the Capsule network, it has been observed that it shows greater resilience compared to other frequently utilized neural networks when subjected to certain fundamental white-box adversarial attacks like FGSM and BIM. The Capsule network demonstrates superior classification accuracy compared to general Convolutional Neural Networks (CNNs) in both untargeted and targeted white-box attacks. This indicates that the Capsule network's architecture shows greater effectiveness in terms of adversarial robustness compared to conventional CNN networks [20].

12 Open Challenges and Future Directions

For adversarial attacks and defenses, there are several unresolved issues and potential directions. Among the most significant challenges are:

- ***Developing more robust machine learning models*** - It is becoming more and more challenging to develop machine learning models that are robust against adversarial attacks as adversarial attacks become more sophisticated.
- ***Designing more effective defense mechanisms*** - Existing Defense mechanisms are frequently inefficient against evolving and novel adversary attacks. It is crucial to design Defense mechanisms that protect machine learning models from a broad range of adversarial attacks.
- ***Understanding the underlying causes of adversarial vulnerability***- The reason machine learning models are vulnerable to adversarial attacks is not yet fully understood. A more in-depth understanding of the fundamental causes of adversarial vulnerability could result in the development of more effective defense mechanisms.

Future research directions in adversarial attacks and defenses include the following:

- ***Developing adversarial attack and defense techniques for new machine learning applications***- In the context of image classification, adversarial attacks, and defenses have been extensively investigated. However, it is essential to develop adversarial attack and defense techniques for other applications of machine learning, such as natural language processing and speech recognition.
- ***Developing adversarial attack and defense techniques that are robust to real-world conditions*** **-** Typically, laboratory-developed adversarial attacks and Defenses are

not robust under real-world conditions. It is crucial to develop adversarial attack and defense techniques that are resilient to a wide range of real-world scenarios, such as noise, lighting variations, and broad devices.

- ***Developing adversarial attack and defense techniques that are efficient and scalable***- Attack and defense techniques that are adversarial can be computationally intensive. It is essential to develop efficient and scalable adversarial attack and defense techniques so that they can be implemented in real-world applications.

The research and analysis of adversarial attacks and defenses is a discipline that is undergoing rapid development. There are many open challenges and potential directions, but there is also a great deal of opportunity for advancement. We can expect the growth of more robust machine learning models and more effective Defense mechanisms as research in this area continues.

13 Conclusion and Future Work

The present state of research on adversarial attacks and defenses in capsule networks is analyzed in this paper. This paper has also discussed the various forms of adversarial attacks that have been proposed, as well as the various defense mechanisms that have been developed to counteract them. And challenges and limitations of existing research, as well as potential directions for future research in this field.

This paper concluded that capsule networks are more robust to adversarial attacks than ordinary neural networks. However, they remain vulnerable to certain forms of attack. There is a need for additional research to develop more efficient defense mechanisms for capsule networks.

This paper also concludes that there is no single defense mechanism that is effective against every type of adversarial attack. It is essential to utilize a combination of defense mechanisms to provide the maximum amount of protection possible against adversarial attacks.

We expect that this review will assist researchers in understanding the challenges and limitations of the existing research on adversarial attacks and defenses in capsule networks. We also expect that this review will assist in the development of more efficient defense mechanisms for capsule networks.

References

1. Kurakin, A., et al.: Adversarial attacks and defenses competition, pp. 195–231 (2018). https://doi.org/10.1007/978-3-319-94042-7_11
2. Qin, Y., Frosst, N., Raffel, C., Cottrell, G., Hinton, G.: Deflecting Adversarial Attacks, no. ICML (2020). http://arxiv.org/abs/2002.07405
3. Liu, N., Du, M., Guo, R., Liu, H., Hu, X.: Adversarial attacks and defenses: an interpretation perspective (2020). http://arxiv.org/abs/2004.11488
4. Marchisio, A., Nanfa, G., Khalid, F., Hanif, M.A., Martina, M., Shafique, M.: SeVuc: a study on the security vulnerabilities of capsule networks against adversarial attacks. Microprocess. Microsyst. **96**, 104738 (2023). https://doi.org/10.1016/j.micpro.2022.104738

5. Osuala, R., et al.: Data synthesis and adversarial networks: a review and meta-analysis in cancer imaging. Med. Image Anal. **84**, 102704 (2023). https://doi.org/10.1016/j.media.2022.102704
6. Marchisio, A., Nanfa, G., Khalid, F., Hanif, M.A., Martina, M., Shafique, M.: CapsAttacks: robust and imperceptible adversarial attacks on capsule networks, pp. 1–10 (2019). http://arxiv.org/abs/1901.09878
7. Shafique, M., et al.: Robust machine learning systems: challenges, current trends, perspectives, and the road ahead. IEEE Des. Test **37**(2), 30–57 (2020). https://doi.org/10.1109/MDAT.2020.2971217
8. Gu, J., Wu, B., Tresp, V.: Effective and efficient vote attack on capsule networks, pp. 1–16 (2021). http://arxiv.org/abs/2102.10055
9. Wu, B., et al.: Attacking adversarial attacks as a defense (2021). http://arxiv.org/abs/2106.04938
10. Sharma, A., Bian, Y., Munz, P., Narayan, A.: Adversarial patch attacks and defenses in vision-based tasks: a survey, pp. 1–15 (2022). http://arxiv.org/abs/2206.08304
11. Jindong, G., Zhao, H., Tresp, V., Torr, P.H.S.: SegPGD: an effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, Part XXIX, pp. 308–325. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19818-2_18
12. Marchisio, A., De Marco, A., Colucci, A., Martina, M., Shafique, M.: RobCaps: evaluating the robustness of capsule networks against affine transformations and adversarial attacks, pp. 1–9 (2023). http://arxiv.org/abs/2304.03973
13. Rasheed, B., Khan, A., Ahmad, M., Mazzara, M., Kazmi, S.M.A.: Multiple adversarial domains adaptation approach for mitigating adversarial attacks effects. Int. Trans. Electr. Energy Syst. **2022** (2022). https://doi.org/10.1155/2022/2890761
14. Mao, J., Weng, B., Huang, T., Ye, F., Huang, L.: Research on multimodality face antispoofing model based on adversarial attacks. Secur. Commun. Netw. **2021** (2021). https://doi.org/10.1155/2021/3670339
15. Hu, L., et al.: Transferable adversarial attacks against automatic modulation classifier in wireless communications. Wirel. Commun. Mob. Comput. **2022** (2022). https://doi.org/10.1155/2022/5472324
16. Han, X., Zhang, Y., Wang, W., Wang, B.: Text adversarial attacks and defenses: issues, taxonomy, and perspectives. Secur. Commun. Netw. **2022** (2022). https://doi.org/10.1155/2022/6458488
17. Fu, X., Gu, Z., Han, W., Qian, Y., Wang, B.: Exploring security vulnerabilities of deep learning models by adversarial attacks. Wirel. Commun. Mob. Comput. **2021** (2021). https://doi.org/10.1155/2021/9969867
18. Yin, H., Zhang, H., Wang, J., Dou, R.: Boosting adversarial attacks on neural networks with better optimizer. Secur. Commun. Networks **1**, 2021 (2021). https://doi.org/10.1155/2021/9983309
19. Puttagunta, M.K., Ravi, S., Nelson Kennedy Babu, C.: Adversarial examples: attacks and defenses on medical deep learning systems. Multimed. Tools Appl. (2023). https://doi.org/10.1007/s11042-023-14702-9
20. Li, Y., Su, H., Zhu, J.: AdvCapsNet: to defense adversarial attacks based on Capsule networks. J. Vis. Commun. Image Represent. **75**, 103037 (2021). https://doi.org/10.1016/j.jvcir.2021.103037
21. Hahn, T., Pyeon, M., Kim, G.: Self-routing capsule networks. In: Advances in Neural Information Processing Systems, vol. 32, no. NeurIPS (2019)
22. Marchisio, A., Mrazek, V., Massa, A., Bussolino, B., Martina, M., Shafique, M.: RoHNAS: a neural architecture search framework with conjoint optimization for adversarial robustness

and hardware efficiency of convolutional and capsule networks. IEEE Access **10**, 109043–109055 (2022). https://doi.org/10.1109/ACCESS.2022.3214312

23. Lau, C.P., Liu, J., Lin, W.A., Souri, H., Khorramshahi, P., Chellappa, R.: Adversarial attacks and robust defenses in deep learning. Handb. Stat. **48**, 29–58 (2023). https://doi.org/10.1016/bs.host.2023.01.001
24. Austin Short, A.G., Pay, T.L.: Adversarial examples, DLSS, vol. SAND2019-1, pp. 1–6 (2019). https://www.osti.gov/servlets/purl/1569514
25. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 86–94 (2017). https://doi.org/10.1109/CVPR.2017.17
26. Arvidsson, V., Al-Mashahedi, A., Boldt, M.: Evaluation of defense methods against the one-pixel attack on deep neural networks. In: 35th Annual Workshop Swedish Artificial Intelligence Society, SAIS 2023, vol. 199, pp. 49–57 (2023). https://doi.org/10.3384/ecp199005