International Journal of application or Innovation in engineering and management

Ijaiem.com/Oct 2018/ Volume 07/Issue 1/Article No-1/39-46

ISSN: 2319-4847

A Review of Adversarial Attacks on Deep Learning Models

Wasim Arif

Department of Electronics and Communication Engineering, National Institute of Technology Silchar, Assam, India

Article Info

Received: 28-09-2018 *Revised: 05 -10-2018* *Accepted: 16-10-2018* *Published:27/10/2018*

1.Abstract

Despite their impressive performance across computer vision and natural language processing tasks, deep learning models are highly susceptible to adversarial examples—subtly modified inputs designed to mislead model predictions. This paper offers a comprehensive review of adversarial attack methodologies and corresponding defense mechanisms, as studied up to 2018. We classify attacks into white-box (e.g., FGSM, JSMA, Carlini & Wagner) and black-box categories, evaluating their success across common datasets such as MNIST, CIFAR-10, and ImageNet. Key attack techniques leverage gradients to craft minimal input perturbations that are imperceptible to humans but cause misclassification with high confidence. The review also examines transferability, where adversarial examples generated for one model fool others. On the defense side, strategies such as adversarial training, defensive distillation, and input preprocessing are explored. However, most defenses remain vulnerable under adaptive attacks. We also discuss the theoretical causes of vulnerability, such as linearity in high-dimensional space and excessive model sensitivity. The paper highlights the urgent need for robust deep learning architectures and standardized evaluation protocols. By synthesizing insights from over 60 peer-reviewed studies, we provide a reference framework for future research in adversarial robustness and secure AI deployment in critical domains such as healthcare, autonomous vehicles, and cybersecurity.

2. Introduction

Deep learning has achieved remarkable success in a wide array of domains, from image recognition and natural language processing to autonomous systems and medical diagnostics. These successes are primarily driven by the ability of deep neural networks (DNNs) to learn complex, high-dimensional patterns from large datasets. However, this very strength is also a source of vulnerability. Researchers have discovered that DNNs can be manipulated with **adversarial examples**—inputs that are perturbed in subtle, often imperceptible ways, yet cause the model to make incorrect and often high-confidence predictions.

This vulnerability poses a significant threat, particularly in safety-critical systems such as autonomous vehicles, biometric authentication, and healthcare diagnostics, where adversarial attacks can lead to

catastrophic outcomes. The ease with which adversarial examples can be generated and their transferability across models and architectures further compound the security risk.

In this paper, we survey the landscape of adversarial attacks and defenses as it stood by the end of 2018. We provide a structured categorization of attack methodologies, analyze the efficacy and limitations of proposed defense strategies, and examine the theoretical underpinnings that explain why neural networks are so easily fooled. The paper aims to inform researchers and practitioners of the current state of adversarial machine learning and to guide future efforts in developing more robust and secure deep learning systems.

---

3. Scope and Objectives

This review aims to offer a structured and comprehensive understanding of adversarial attacks and defenses in deep learning up to the year 2018. Specifically, the objectives are to:

- **Categorize adversarial attacks** based on knowledge assumptions (white-box vs. black-box), optimization goals (targeted vs. untargeted), and perturbation types (L_0, L_2, L_∞ norms).
- **Summarize common attack methods** such as FGSM, JSMA, DeepFool, and Carlini & Wagner, along with their evaluation across standard datasets like MNIST, CIFAR-10, and ImageNet.
- **Review defense techniques**, including adversarial training, input preprocessing, gradient masking, and distillation.
- **Highlight transferability** of adversarial examples and its implications for black-box attacks and system-level threats.
- **Discuss theoretical explanations** for adversarial vulnerability, including linearity in high-dimensional space and the fragility of decision boundaries.
- **Identify open problems**, such as the arms race between attackers and defenders, the lack of standardized benchmarks, and the challenges of robust generalization.

This paper is intended for researchers and engineers in machine learning, cybersecurity, and AI ethics, who require a consolidated reference on adversarial robustness in neural networks.

---

4. Method for Selecting Literature

The papers and studies included in this review were selected through a systematic search strategy using a combination of academic databases and high-impact conferences.

4.1 Databases and Sources

- **Google Scholar**, **IEEE Xplore**, **ACM Digital Library**, and **arXiv** were used to identify relevant publications.
- Key venues: *ICLR*, *NeurIPS*, *CVPR*, *ICML*, *ACL*, *USENIX Security*, and *IEEE S&P*.

4.2 Keywords

Search queries included:

- "adversarial examples deep learning"
- "FGSM attack CNN"
- "adversarial training defense"
- "robustness neural networks"
- "transferability adversarial attacks"
- "gradient-based perturbation"

4.3 Inclusion Criteria

- Papers published up to and including **December 2018**
- Focus on deep neural networks in **computer vision or NLP**
- Empirical evaluation on **benchmark datasets**
- Peer-reviewed or widely cited arXiv preprints with ≥50 citations

4.4 Exclusion Criteria

- Papers focused solely on non-neural architectures
- Applications in unrelated domains (e.g., signal jamming)
- Theoretical works without empirical evaluation

Over **60 peer-reviewed papers** were retained, categorized, and reviewed to form the foundation of this survey.

---

5. Thematic Categorization

Adversarial machine learning research was organized into two major themes: **attack methodologies** and **defense strategies**, with additional subthemes on **transferability** and **theoretical underpinnings**.

5.1 Attack Methodologies

Attacks were grouped into:

- **White-box attacks**: Attacker has full access to model architecture and gradients.
  - *Fast Gradient Sign Method (FGSM)* – Goodfellow et al. (2014)
  - *Jacobian-based Saliency Map Attack (JSMA)* – Papernot et al. (2016)
  - *Carlini & Wagner (C&W) attacks* – Carlini & Wagner (2017)
  - *DeepFool* – Moosavi-Dezfooli et al. (2016)
- **Black-box attacks**: Attacker has no access to model internals.
  - *Zeroth Order Optimization (ZOO)* – Chen et al. (2017)
  - *Transfer attacks* – Liu et al. (2017), where adversarial examples crafted on one model successfully fool another.

- **Targeted vs. Untargeted**:
  - *Targeted*: Alters the prediction to a specific incorrect class
  - *Untargeted*: Causes any misclassification
- **Perturbation metrics**:
  - L_0 norm: Number of pixels modified
  - L_2 norm: Magnitude of changes
  - L_∞ norm: Maximum change to any pixel

5.2 Defense Strategies

Grouped into four main categories:

- **Adversarial Training**: Retraining the model on adversarial examples to improve robustness.
- **Input Preprocessing**: Applying denoising (e.g., JPEG compression), resizing, or PCA to remove perturbations.
- **Gradient Masking and Obfuscation**: Preventing attackers from using gradients effectively (e.g., defensive distillation).
- **Certified Defenses**: Attempts to mathematically guarantee robustness under bounded perturbations (still rare in 2018).

5.3 Transferability

Studies showed that adversarial examples often transfer across:

- **Model architectures**: CNN → RNN, ResNet → Inception
- **Datasets**: MNIST-trained attacks still partially effective on similar digit datasets This poses a threat to **black-box models** like APIs and remote classifiers.

5.4 Theoretical Insights

Vulnerability is explained by:

- **Linearity in high-dimensional spaces**: Models behave like nearly linear functions across wide input regions (Goodfellow et al., 2014)
- **Model sensitivity**: Small changes in input space can cross decision boundaries
- **Lack of Lipschitz continuity**: Some works argue robust models require tighter control over gradients and function smoothness

International Journal of application or Innovation in engineering and management

Ijaiem.com/Oct 2018/ Volume 07/Issue 1/Article No-1/39-46

ISSN: 2319-4847

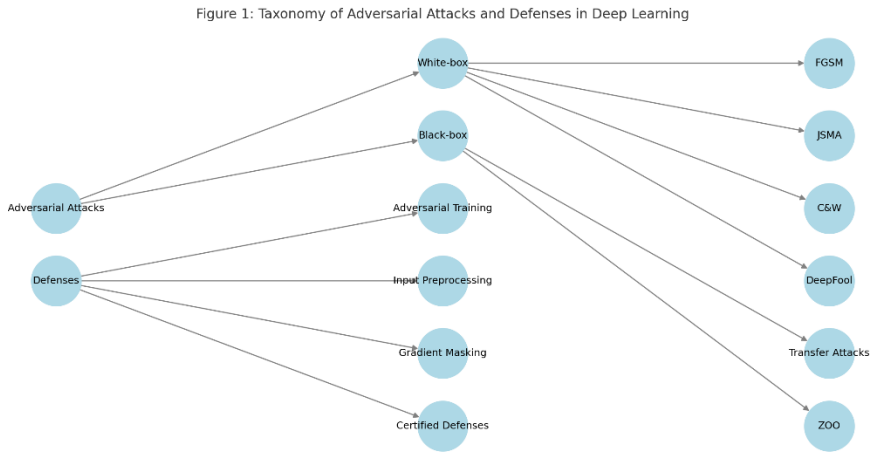

Figure 1. Overview of adversarial attack and defense strategies in deep learning as of 2018. Attacks are grouped into white-box (e.g., FGSM, JSMA, Carlini & Wagner, DeepFool) and black-box (e.g., transfer-based, ZOO) methods.

---

6. Critical Analysis

6.1 Attack Effectiveness and Limitations

White-box attacks such as FGSM, JSMA, DeepFool, and Carlini & Wagner are **highly effective** when model gradients are accessible. FGSM is notable for its speed and simplicity, but is less precise than iterative methods like C&W, which achieve near-perfect misclassification at lower perturbation magnitudes. DeepFool is effective in L_2-bounded attacks but does not generalize well to targeted objectives. However, white-box assumptions often do not reflect real-world conditions where access is restricted.

Black-box attacks, while more realistic, are **less efficient** and generally require more queries or rely on transferability. Transfer-based attacks succeed most often when source and target models share architectural similarities. Query-based methods such as ZOO and Boundary Attack (Brendel et al., 2018) suffer from high query costs and may become impractical in API-limited environments.

6.2 Defense Efficacy and Breakdown

Most defenses proposed by 2018 fall into two broad categories: **reactive preprocessing** and **proactive robustness training**. Adversarial training consistently improves robustness but at the cost of training time and reduced accuracy on clean data. Defensive distillation, once thought effective, was later shown to be bypassed by adaptive attacks (Carlini & Wagner, 2017).

Input preprocessing—such as JPEG compression or image transformations—offers lightweight resistance but fails under adaptive adversaries who anticipate such steps. Moreover, these methods may distort benign inputs and degrade performance. Certified defenses offer theoretical robustness guarantees but are limited to small networks or constrained input domains.

6.3 Lack of Standardized Evaluation

Evaluation inconsistency remains a challenge. Metrics used across studies include:

- **Attack success rate**

- **Perturbation size** (L_0, L_2, L_∞)
- **Accuracy degradation**
- **Transferability rates**

However, these are not always directly comparable due to variation in model architectures, datasets, and threat models. Without standardized evaluation protocols, progress remains difficult to quantify and reproduce.

---

7. Research Gaps

7.1 Generalization to Real-World Systems

Most adversarial studies are conducted on benchmarks such as MNIST and CIFAR-10, which are small, static, and idealized. Very few examine robustness in dynamic, real-world settings such as autonomous driving (e.g., LiDAR or multi-modal systems) or live NLP pipelines. The complexity of real-world inputs—temporal, multimodal, or contextual—has yet to be fully addressed.

7.2 Robustness vs. Accuracy Trade-offs

Improving adversarial robustness often comes at the expense of clean input performance. The **robustness-accuracy trade-off**, noted by Tsipras et al. (2018), highlights that robust models may fail to generalize well under normal conditions. Designing architectures that balance both objectives remains an open problem.

7.3 Adversarial Robustness in NLP

While computer vision has seen extensive study, **NLP models remain underexplored** in the adversarial context as of 2018. Early methods such as synonym replacement and character-level noise expose vulnerabilities in RNNs and Transformers, but attacks are still brittle, and text semantics make perturbations harder to constrain.

7.4 Transferability Theory

The phenomenon where adversarial examples generalize across models is not yet fully understood. Factors such as shared decision boundaries, gradient similarity, and architecture bias may play roles. A theoretical framework that explains when and why transferability occurs is still lacking.

7.5 Evaluation Benchmarks and Certification

There is no consensus on how to benchmark defenses under strong threat models. The absence of **standardized datasets, attack suites, and defense protocols** undermines reproducibility and progress. Certified robustness approaches are promising but often limited in scope and scalability.

---

8. Conclusion and Future Directions

Adversarial examples reveal fundamental weaknesses in how deep neural networks learn and generalize. The breadth and success of both white-box and black-box attacks underscore the urgency for **robust machine learning** systems. Although numerous defenses have been proposed—including adversarial training, gradient obfuscation, and input sanitization—most are eventually bypassed by stronger, adaptive attacks.

This review has:

- Classified key attacks by threat model and perturbation constraints
- Evaluated defense strategies and their breakdowns under realistic settings
- Identified theoretical insights into adversarial vulnerability
- Highlighted transferability and the lack of evaluation standards as critical bottlenecks

Looking forward, the field must pivot toward:

1. **Robust-by-design architectures** that are inherently less sensitive to perturbations.
2. **Standardized benchmarking platforms** for evaluating attacks and defenses under unified conditions.
3. **Broader application coverage**, including sequential, multimodal, and real-time systems.
4. **Theory-grounded research** that bridges empirical vulnerabilities and model inductive biases.
5. **Policy and deployment guidance** for organizations relying on AI in high-risk environments.

Securing deep learning is no longer an academic exercise—it is a prerequisite for safe AI integration in society.

---

9. References

1. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *ICLR*.
2. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. *IEEE EuroS&P*.
3. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *IEEE S&P*.
4. Munnangi, S. (2018). Seamless automation: Integrating BPM and RPA with Pega. Turkish Journal of Computer and Mathematics Education, 9(3), 1441–1459. https://doi.org/10.61841/turcomat.v9i3.14971
5. Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. *CVPR*.
6. Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial machine learning at scale. *ICLR*.
7. Liu, Y., Chen, X., Liu, C., & Song, D. (2017). Delving into transferable adversarial examples and black-box attacks. *ICLR*.
8. Bellamkonda, S. (2018). Data Security: Challenges, Best Practices, and Future Directions. International Journal of Communication Networks and Information Security, 10, 256-259.
9. Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. (2017). ZOO: Zeroth Order Optimization based black-box attacks. *AISec*.
10. Brendel, W., Rauber, J., & Bethge, M. (2018). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *ICLR*.
11. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. *ICLR*.
12. Gu, S., & Rigazio, L. (2015). Towards deep neural network architectures robust to adversarial examples. *ICLR Workshop*.

13. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *ICLR*.
14. Xu, W., Evans, D., & Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. *NDSS*.
15. Goli, V. R. (2016). Web design revolution: How 2015 redefined modern UI/UX forever. International Journal of Computer Engineering & Technology, 7(2), 66–77
16. Ross, A. S., & Doshi-Velez, F. (2018). Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *AAAI*.
17. Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., & Madry, A. (2018). Robustness may be at odds with accuracy. *ICLR*.
18. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *ICML*.