Baltic Journal of MultiDisciplinary Research - BJMR

Pages: 43-49

Volume-II, Issue-III (2025)

Cross-Domain Adversarial Attack Taxonomy for Multimodal Neural Networks: Threat Landscape and Open Challenges

Author: [1]Hadia Azmat, [2]Ifrah Ikram

Corresponding Author: hadiaazmat728@gmail.com

Abstract

Multimodal neural networks have emerged as powerful frameworks for integrating diverse data modalities such as text, images, audio, and video, enabling advancements in fields ranging from autonomous systems to healthcare and security. However, their increasing adoption has exposed them to adversarial attacks that exploit vulnerabilities across multiple domains simultaneously. Unlike unimodal systems, multimodal networks present expanded attack surfaces due to cross-modal dependencies, complex fusion strategies, and heterogeneous data representations. This paper introduces a taxonomy for cross-domain adversarial attacks targeting multimodal neural networks, categorizing them by modality manipulation, attack vectors, and fusion-stage vulnerabilities. We highlight how these attacks undermine robustness, reliability, and interpretability, while also complicating detection and defense mechanisms. Through a detailed exploration of the threat landscape, we emphasize challenges such as cross-domain transferability of adversarial perturbations, coordinated attacks exploiting multimodal fusion, and real-time adversarial adaptation. Finally, we identify open challenges and research directions for developing resilient, interpretable, and trustworthy multimodal systems capable of withstanding evolving adversarial threats.

[1]University of Lahore, Pakistan

[2]COMSATS University Islamabad, Pakistan

Keywords: Adversarial attacks, multimodal neural networks, cross-domain vulnerabilities, threat taxonomy, robustness, transferability, deep learning security, interpretability, data fusion, trustworthy AI

Introduction

The integration of multiple data modalities into deep learning architectures has transformed the capabilities of artificial intelligence systems. Multimodal neural networks, which jointly process modalities such as vision, language, and audio, have enabled breakthroughs in domains including autonomous driving, medical diagnosis, human–computer interaction, and security monitoring. By fusing complementary sources of information, these systems achieve superior performance compared to unimodal models, offering richer representations and improved generalization. However, the complexity inherent in multimodal architectures has also introduced new vulnerabilities, particularly in the context of adversarial machine learning[1].

Adversarial attacks—small, carefully crafted perturbations designed to mislead neural networks—have been extensively studied in unimodal domains, most notably in computer vision. In these cases, imperceptible changes to an input image can cause a model to misclassify it with high confidence. Extending such attacks to multimodal settings, however, poses novel challenges and opportunities for adversaries. Unlike unimodal models, multimodal neural networks rely on cross-modal fusion, where information from different modalities is combined. This creates new points of attack, including modality-specific manipulations, fusion-stage vulnerabilities, and coordinated cross-domain perturbations[2].

For example, in a video–text retrieval system, an adversary may manipulate the visual modality by introducing subtle noise to frames while simultaneously injecting misleading textual cues. Similarly, in autonomous driving systems, both sensor data such as LiDAR and cameras and linguistic instructions such as navigation commands may be manipulated, causing catastrophic failures. Such cross-domain adversarial strategies significantly complicate the task of defense, as robustifying one modality may not protect the system as a whole[3].

To address these challenges, a systematic understanding of the threat landscape is required. This paper introduces a taxonomy of cross-domain adversarial attacks for multimodal neural networks, categorizing them by attack surface, modality, and stage of execution. The taxonomy captures not only attacks that exploit single modalities but also those that leverage coordinated strategies across multiple inputs, thereby targeting the system's fusion and reasoning processes. By establishing this classification, we aim to provide researchers and practitioners with a framework for identifying, analyzing, and mitigating vulnerabilities in multimodal AI systems[4].

The need for such a taxonomy is amplified by the growing deployment of multimodal AI in safety-critical environments. Adversarial attacks on healthcare diagnostic tools can jeopardize patient outcomes, while attacks on multimodal surveillance or autonomous systems can threaten public safety. Furthermore, adversarial robustness is closely tied to issues of fairness, interpretability, and trust. Defending against attacks requires more than brute-force robustness; it demands a principled understanding of how multimodal models reason across domains and how adversaries exploit their dependencies[5].

This paper proceeds by first presenting a taxonomy of cross-domain adversarial attacks, discussing categories, mechanisms, and representative examples. It then analyzes the broader threat landscape and highlights open challenges, focusing on transferability, interpretability, and real-time adversarial dynamics. Finally, it concludes by outlining directions for developing resilient and trustworthy multimodal neural networks capable of defending against evolving adversarial threats[6].

Cross-Domain Adversarial Attack Taxonomy

Cross-domain adversarial attacks in multimodal neural networks can be systematically classified according to modality manipulation, target stage, and attack coordination. The first class involves modality-specific attacks, where adversaries manipulate only one input domain while leaving others unchanged. For example, an image-based perturbation in a vision–language model may cause incorrect captioning despite the text input being intact. Similarly, adversarial text crafted for a question-answering system can exploit linguistic vulnerabilities while visual

features remain unaffected. These attacks are deceptively simple but highly effective, as downstream fusion often amplifies even subtle perturbations[7].

The second category comprises coordinated attacks that manipulate multiple modalities simultaneously. Here, adversaries design perturbations that reinforce one another, achieving higher effectiveness than single-modality attacks. An attacker may combine adversarial noise in visual data with misleading textual prompts, leading to synergistic misclassifications that bypass unimodal defenses. Such attacks are particularly dangerous in domains like autonomous navigation, where both sensor inputs and command signals influence system behavior[8].

A third dimension of attacks targets the fusion process itself. Multimodal systems typically rely on early, intermediate, or late fusion strategies, and each stage presents unique vulnerabilities. Early fusion can be disrupted by subtle manipulations in raw feature concatenation, while intermediate fusion that uses attention mechanisms may be tricked into aligning misleading cross-modal signals. Late fusion attacks manipulate decision-level integration, skewing outcomes by overpowering contributions from one modality. Attacking the fusion stage is especially insidious because it undermines the very mechanism designed to improve system robustness[9].

Another significant category involves transferability-based attacks. Adversarial perturbations have long been known to transfer across models, but in multimodal systems, this property extends across modalities as well. Perturbations crafted for one domain may influence another through shared embedding spaces. For example, adversarial noise created for images may propagate into retrieval errors in text–image search systems. This cross-domain transferability complicates defense strategies, as protecting one modality in isolation may leave the system exposed at a different integration point[10].

Finally, real-time adaptive attacks exploit the dynamic nature of multimodal systems. In contexts such as video–audio analysis or real-time navigation, adversaries can continuously adjust perturbations by monitoring system responses. This iterative feedback loop allows them to maintain deception over time, rendering static defenses ineffective. Adaptive attacks highlight

the need for continuous, evolving defense mechanisms capable of responding to adversarial strategies as they unfold[11].

Representative case studies illustrate the severity of these threats. Visual question answering systems may be misled by perturbations in either the visual or textual modality, but coordinated manipulations across both inputs dramatically increase attack success rates. Similarly, multimodal biometric systems that combine facial recognition with voice authentication can be bypassed when adversaries exploit vulnerabilities across both modalities. These examples underscore the layered and complex nature of adversarial threats in multimodal neural networks, demonstrating the necessity of a structured taxonomy for understanding and addressing them[12].

Threat Landscape and Open Challenges

The adversarial threat landscape for multimodal neural networks is rapidly expanding, driven by the increasing adoption of such systems in high-stakes applications. One of the most significant features of this landscape is the expanded attack surface. Each modality brings unique vulnerabilities, and when combined, they create additional pathways for exploitation. An adversary who fails to compromise one modality can redirect efforts toward another, ensuring persistent threats to system integrity. This combinatorial attack surface challenges defenders to design comprehensive strategies that protect all domains simultaneously.

Another defining feature of the threat landscape is the transferability of adversarial examples across domains. Perturbations crafted for one modality can have cascading effects in others, undermining the fusion process. This cross-domain transferability complicates defense because it blurs the distinction between unimodal robustness and multimodal vulnerability. Addressing this issue requires a deeper understanding of how shared representations propagate perturbations across modalities[13].

The tension between robustness and interpretability adds further complexity. Defenses often rely on adversarial training, regularization, or complex fusion mechanisms, which can obscure interpretability. Yet interpretability is crucial for trust in systems deployed in critical environments. At the same time, efforts to improve interpretability, such as attention

visualization, may reveal pathways that attackers exploit. Balancing robustness with transparency thus remains an unresolved challenge.

Real-time adaptive threats represent another frontier of the adversarial landscape. Multimodal systems used in dynamic environments, such as autonomous driving or live surveillance, must contend with adversaries capable of iteratively adapting their strategies. These threats render static defenses inadequate, highlighting the need for adaptive, online learning approaches that evolve alongside attacks[14].

The absence of standardized benchmarks for evaluating multimodal adversarial robustness further exacerbates the situation. Unlike unimodal adversarial research, which benefits from widely adopted datasets and evaluation protocols, multimodal research remains fragmented. Without common baselines, comparing defenses or assessing vulnerabilities is difficult, slowing progress in the field. Establishing multimodal adversarial benchmarks is therefore a critical research priority.

Another open challenge lies in the computational overhead associated with defending multimodal systems. Adversarial training across multiple modalities demands significant resources, often making it impractical for large-scale deployment. Organizations with limited computational budgets may be unable to adopt state-of-the-art defenses, leaving their systems exposed. Lightweight, scalable defense mechanisms are urgently needed to bridge this gap[15].

Finally, the societal and ethical implications of adversarial threats in multimodal systems cannot be overlooked. Attacks on healthcare systems can lead to misdiagnoses, while those on surveillance or autonomous platforms can jeopardize safety and public trust. The propagation of adversarially manipulated multimedia also fuels misinformation campaigns, undermining societal confidence in AI technologies. Addressing these issues requires technical innovation alongside ethical frameworks and regulatory oversight.

Together, these factors define a complex and evolving threat landscape. They demonstrate that adversarial challenges in multimodal neural networks extend beyond technical robustness to encompass interpretability, scalability, benchmarking, and ethics. Overcoming these challenges

will require a holistic approach that integrates advances in machine learning, human–AI interaction, and governance[16].

Conclusion

Multimodal neural networks, while transformative in their ability to fuse diverse sources of information, face profound adversarial risks due to their expanded attack surfaces and complex fusion strategies. The taxonomy presented in this paper highlights the breadth of adversarial methods, ranging from modality-specific perturbations to coordinated, fusion-stage, transferability-based, and real-time adaptive attacks. The broader threat landscape underscores pressing challenges, including cross-domain transferability, the robustness–interpretability trade-off, and the lack of benchmarking standards. Addressing these challenges demands not only technical defenses but also ethical, regulatory, and societal considerations. As multimodal AI continues to permeate safety-critical domains, ensuring resilience against adversarial attacks is vital for building secure, interpretable, and trustworthy systems.

References:

[1] J. Barach, "Federated Learning for Privacy-Preserving Employee Performance Analytics," *IEEE Access,* 2025.

[2] J. P. Wahle, N. Ashok, T. Ruas, N. Meuschke, T. Ghosal, and B. Gipp, "Testing the generalization of neural language models for COVID-19 misinformation detection," in *International Conference on Information*, 2022: Springer, pp. 381-392.

[3] J. Barach, "Cross-Domain Adversarial Attacks and Robust Defense Mechanisms for Multimodal Neural Networks," in *International Conference on Advanced Network Technologies and Intelligent Computing*, 2024: Springer, pp. 345-362.

[4] J. P. Wahle, T. Ruas, N. Meuschke, and B. Gipp, "Incorporating Word Sense Disambiguation in Neural Language Models," *arXiv preprint arXiv:2106.07967,* 2021.

[5] J. Barach, "AI-Driven Causal Inference for Cross-Cloud Threat Detection Using Anonymized CloudTrail Logs," in *2025 Conference on Artificial Intelligence x Multimedia (AIxMM)*, 2025: IEEE, pp. 45-50.

[6] G. Singh, A. Mallik, Z. Iqbal, H. Revalla, S. Chao, and V. Nagasamy, "Systems and methods for detecting deep neural network inference quality using image/data manipulation without ground truth information," ed: Google Patents, 2023.

[7] J. Barach, "Enhancing intrusion detection with CNN attention using NSL-KDD dataset. In 2024 Artificial Intelligence for Business (AIxB)(pp. 15-20)," ed: IEEE, 2024.

[8] S. K. Patel, "Improving intrusion detection in cloud-based healthcare using neural network," *Biomedical Signal Processing and Control,* vol. 83, p. 104680, 2023.

[9] J. Barach, "Integrating AI and HR Strategies in IT Engineering Projects: A Blueprint for Agile Success," *Emerging Engineering and Mathematics,* pp. 1-13, 2025.

[10] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational AI," in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018, pp. 1371-1374.

[11] J. Barach, "Towards Zero Trust Security in SDN: A Multi-Layered Defense Strategy," in *Proceedings of the 26th International Conference on Distributed Computing and Networking*, 2025, pp. 331-339.

[12] S. Diao, C. Wei, J. Wang, and Y. Li, "Ventilator pressure prediction using recurrent neural network," *arXiv preprint arXiv:2410.06552,* 2024.

[13] B. Dai and D. Lin, "Contrastive learning for image captioning," *Advances in Neural Information Processing Systems,* vol. 30, 2017.

[14] C. Chun, K. M. Jeon, C. Leem, B. Lee, and W. Choi, "Comparison of CNN-based Speech Dereverberation using Neural Vocoder," in *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2021: IEEE, pp. 251-254.

[15] I. R. a. Kelley, "Data management in dynamic distributed computing environments," Thesis (Ph.D.), Cardiff University, 2012. [Online]. Available: http://orca.cf.ac.uk/44477/

[16] N. Khan, "Semi-Supervised Generative Adversarial Network for Stress Detection Using Partially Labeled Physiological Data," *arXiv preprint arXiv:2206.14976,* 2022.