Baltic Journal of MultiDisciplinary Research - BJMR

Pages: 57-63

Volume-II, Issue-III (2025)

Self-Supervised Robustness Enhancement for Multimodal Neural Networks Under Cross-Domain Adversarial Perturbations

Author: [1]Rohan Sharma, [2]Aarav Sharma

Corresponding Author: rohan126578@gmail.com

Abstract

Multimodal neural networks have become foundational in artificial intelligence, enabling systems to learn and reason from diverse modalities such as vision, language, and audio. However, these models remain highly vulnerable to cross-domain adversarial perturbations, where subtle but carefully crafted manipulations across one or more modalities lead to significant performance degradation. Traditional supervised defense mechanisms struggle to address these threats, primarily due to the lack of labeled adversarial data and the complexity of multimodal interactions. This paper proposes self-supervised robustness enhancement as a promising defense paradigm. By leveraging self-supervised pretext tasks and representation learning, multimodal models can learn modality-invariant, semantically consistent features that improve resilience against adversarial inputs. We explore contrastive learning, masked prediction, and redundancy-driven objectives as self-supervised strategies to fortify robustness. The analysis highlights how these methods mitigate cross-domain perturbations without explicit adversarial labels, while also improving generalization and interpretability. The discussion further outlines challenges in scalability, stability, and real-world deployment, positioning self-supervised learning as a crucial pathway toward robust and trustworthy multimodal AI.

Keywords: Self-supervised learning, multimodal neural networks, adversarial robustness, cross-domain perturbations, contrastive learning, representation learning, defense mechanisms

[1]Indian Institute of Technology (IIT) Bombay, Mumbai, India

[2]International Institute of Information Technology (IIIT), Hyderabad, India

I. Introduction

Multimodal neural networks represent a significant leap in the development of artificial intelligence, enabling machines to integrate and process heterogeneous data sources such as text, images, audio, and video. By leveraging complementary information across modalities, these systems achieve superior performance in tasks like visual question answering, speech-vision recognition, medical diagnosis, and autonomous navigation. However, this capability also exposes them to unique adversarial vulnerabilities. Unlike unimodal models, multimodal networks are susceptible to cross-domain adversarial perturbations, where attackers manipulate multiple modalities simultaneously or exploit correlations between them. These perturbations are particularly dangerous because they exploit the very cross-modal synergies that multimodal models rely on, often leading to cascading misclassifications or erroneous predictions.

Conventional defenses against adversarial attacks typically rely on supervised approaches, such as adversarial training or defensive distillation. These methods require access to labeled adversarial examples, which are expensive to generate, limited in scope, and ineffective against unseen perturbations. In multimodal contexts, this reliance on labeled data is even more problematic, as the attack space expands exponentially with the number of modalities. An attacker could manipulate visual features, textual embeddings, or their interactions, making it impractical to anticipate and label every possible adversarial variation. As a result, supervised defenses often fail to generalize across domains and modalities[1].

Self-supervised learning has recently emerged as a transformative approach in representation learning, offering a way to harness large-scale unlabeled data to learn rich, generalizable features. Through pretext tasks—such as predicting missing tokens, reconstructing occluded image patches, or contrasting positive and negative pairs—models can learn modality-invariant representations that capture semantic consistency across domains. When applied to multimodal networks, self-supervised learning provides an avenue to enhance robustness without requiring explicit adversarial supervision. Instead of defending against specific perturbations, models learn to emphasize stable, semantically aligned features that are less sensitive to cross-domain manipulations.

For example, contrastive self-supervised methods encourage representations of paired modalities (e.g., an image and its corresponding caption) to remain close in embedding space, while unrelated pairs are pushed apart. This alignment discourages adversaries from easily disrupting cross-modal coherence with small perturbations. Similarly, masked prediction tasks force models to infer missing information within or across modalities, reinforcing redundancy and error-correction capabilities that are critical under attack. By combining such self-supervised objectives with multimodal architectures, it becomes possible to construct models that are inherently more resilient to adversarial perturbations, even in the absence of labeled adversarial data[2].

The potential of self-supervised robustness enhancement lies not only in security but also in improved generalization, interpretability, and efficiency. However, challenges remain in scaling these methods to high-dimensional multimodal data, ensuring stability during adversarial encounters, and validating their effectiveness in real-world applications where attack strategies evolve rapidly. This paper explores the application of self-supervised techniques to enhance the robustness of multimodal neural networks against cross-domain adversarial perturbations, examining current strategies, their limitations, and promising directions for future research.

II. Cross-Domain Adversarial Perturbations in Multimodal Neural Networks: Challenges and Limitations

The vulnerability of multimodal neural networks to cross-domain adversarial perturbations arises from their dependence on joint feature spaces that fuse heterogeneous modalities. Attackers exploit this fusion by introducing perturbations that appear benign within individual modalities but, when combined, distort the shared representation space. For instance, a slightly altered sentence paired with a minimally perturbed image can mislead a vision-language model into producing incorrect answers. Such perturbations are effective because they target the alignment mechanisms that bind modalities together, creating inconsistencies that propagate throughout the model's decision pipeline. Addressing this challenge requires defenses that are modality-aware and capable of reinforcing stable, semantically grounded representations[3].

Self-supervised learning provides several mechanisms to achieve this resilience. Contrastive learning, one of the most widely adopted self-supervised paradigms, is particularly effective in multimodal contexts. By aligning positive pairs across modalities and separating negative ones, contrastive learning encourages the model to focus on features that are robust and semantically consistent. For example, in a multimodal image-text dataset, contrastive objectives ensure that perturbations that fail to preserve semantic coherence are less likely to influence the representation space. This discourages attackers from introducing subtle yet harmful mismatches, as the model becomes trained to disregard irrelevant variations[4].

Masked prediction is another self-supervised strategy that enhances robustness. By randomly masking portions of the input—such as image patches, audio segments, or textual tokens—and requiring the model to reconstruct them, multimodal networks develop redundancy-aware features. This redundancy acts as a natural defense against adversarial attacks, as the model learns to infer missing or corrupted information using context from other modalities. For example, if adversarial perturbations compromise a section of the image, the model can rely on accompanying textual cues to reconstruct the intended meaning, thereby reducing the attack's effectiveness[5].

Moreover, self-supervised pretext tasks can explicitly encourage cross-modal consistency. For instance, cycle-prediction tasks—where the model predicts modality A from B and vice versa—reinforce bidirectional dependencies that make it harder for adversaries to disrupt both directions simultaneously. Such tasks strengthen the semantic binding between modalities, ensuring that perturbations in one domain are detected through inconsistencies in the other. This redundancy-driven resilience mirrors principles in error-correcting codes, where robustness is achieved through distributed representation and redundancy[6].

Beyond these individual strategies, combining multiple self-supervised tasks can further enhance defense. A hybrid approach that integrates contrastive learning, masked prediction, and cycle-consistency enables multimodal systems to learn robust representations from multiple angles. These representations are not tailored to a specific adversarial pattern but instead capture fundamental semantic structures that are difficult to manipulate without introducing detectable inconsistencies. By reducing sensitivity to perturbations and reinforcing cross-modal coherence,

self-supervised learning offers a defense mechanism that is inherently adaptable and scalable to diverse multimodal settings[7].

III. Self-Supervised Learning Paradigms for Robustness Enhancement and Adaptive Defense

While self-supervised robustness enhancement offers promising directions, several challenges and open research questions remain in applying these techniques to multimodal adversarial defense. One of the primary issues is scalability. Multimodal data is often high-dimensional and heterogeneous, requiring complex architectures to capture joint representations. Training self-supervised models with multiple pretext tasks on large-scale datasets demands significant computational resources, which may limit their applicability in real-world systems. Efficient approximations and lightweight architectures will be necessary to bring self-supervised robustness techniques into practical deployment[8].

Another challenge lies in stability. Adversarial perturbations are dynamic and often adapt to the defense strategies employed. While self-supervised objectives improve resilience against generic perturbations, attackers may design adversarial examples specifically to exploit weaknesses in these objectives. For example, adversaries could craft perturbations that preserve contrastive alignment while still causing misclassification, undermining the defense. This necessitates continuous refinement of self-supervised tasks and the development of adaptive strategies that evolve alongside adversarial tactics[9].

Interpretability is also a critical concern. Multimodal systems already face scrutiny for their black-box nature, and introducing self-supervised objectives adds another layer of complexity. To build trust in these defenses, it is important to develop interpretable frameworks that explain how self-supervised tasks contribute to robustness. Causal modeling and attention visualization are promising directions for shedding light on how self-supervised learning redistributes focus across modalities during adversarial encounters. By making the defense mechanisms transparent, researchers can better evaluate their reliability and identify potential blind spots[10].

Despite these challenges, the integration of self-supervised learning with adversarial robustness opens exciting opportunities for real-world applications. In healthcare, multimodal diagnostic systems that combine imaging, clinical notes, and genomic data can benefit from self-supervised tasks that enforce cross-modal consistency, reducing the risk of adversarial manipulation in sensitive domains. In autonomous driving, where sensor data, images, and textual instructions converge, self-supervised robustness can ensure reliable performance under adversarial conditions such as sensor spoofing or manipulated signage. Similarly, in social media platforms, where multimodal content is subject to misinformation campaigns, self-supervised defenses can detect inconsistencies between modalities, helping to curb adversarial exploitation[11].

Future research must also explore hybrid defenses that combine self-supervised learning with traditional adversarial training. While self-supervised approaches reduce reliance on labeled adversarial data, integrating them with supervised methods can yield complementary benefits. For instance, adversarial training can provide robustness against known attacks, while self-supervised tasks enhance generalization to unseen perturbations. This synergy can create a layered defense framework that is more resilient to evolving adversarial strategies[12].

Ultimately, self-supervised robustness enhancement for multimodal neural networks represents a paradigm shift from reactive defenses to proactive resilience. By leveraging unlabeled data and reinforcing semantic coherence, these methods address the complexity of cross-domain adversarial perturbations without requiring exhaustive supervision. Although scalability, stability, and interpretability remain ongoing challenges, the potential benefits of this approach position self-supervised learning as a cornerstone of robust and trustworthy multimodal AI[13].

IV. Conclusion

Self-supervised learning presents a powerful approach to enhancing the robustness of multimodal neural networks under cross-domain adversarial perturbations. By focusing on semantic consistency, redundancy, and cross-modal coherence, self-supervised tasks such as contrastive learning, masked prediction, and cycle-consistency provide a defense mechanism that does not depend on labeled adversarial data. While significant challenges remain in scalability, adaptability, and transparency, the integration of self-supervised robustness strategies offers a

promising pathway toward securing multimodal AI systems. As adversarial threats grow more sophisticated, self-supervised learning stands as a critical foundation for building resilient, generalizable, and trustworthy multimodal intelligence.

References:

[1] 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.

[2] S. K. Patel, "Attack detection and mitigation scheme through novel authentication model enabled optimized neural network in smart healthcare," *Computer Methods in Biomechanics and Biomedical Engineering,* vol. 26, no. 1, pp. 38-64, 2023.

[3] E. Song, K. Byun, and H.-G. Kang, "ExcitNet vocoder: A neural excitation model for parametric speech synthesis systems," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019: IEEE, pp. 1-5.

[4] 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.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156-3164.

[6] 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.

[7] I. Salehin *et al.*, "AutoML: A systematic review on automated machine learning with neural architecture search," *Journal of Information and Intelligence,* vol. 2, no. 1, pp. 52-81, 2024.

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

[9] H. Azmat and Z. Huma, "Designing Security-Enhanced Architectures for Analog Neural Networks," *Pioneer Research Journal of Computing Science,* vol. 1, no. 2, pp. 1-6, 2024.

[10] 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.

[11] G. Alhussein, M. Alkhodari, A. Khandoker, and L. Hadjileontiadis, "Novel speech-based emotion climate recognition in peers' conversations incorporating affect dynamics and temporal convolutional neural networks," *Available at SSRN 4846084*.

[12] 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.

[13] Z. Huma and A. Mustafa, "Hardware Security in Energy-Constrained Neural Processing Units," *Journal of Data and Digital Innovation,* vol. 1, no. 1, pp. 1-7, 2025.