A study on cyber attack targeted for semantic communication systems

Mehwish Ali Naqvi, Insoo Sohn*

Division of Electronics & Electrical Engineering, Dongguk University, Seoul, Republic of Korea

Abstract

Semantic communication (SeCom) systems aim to transmit task-relevant meaning rather than raw data, offering enhanced efficiency for next-generation networks. However, their reliance on deep learning and generative models introduces new and largely unexplored security vulnerabilities. This paper examines four recent and representative cyber attacks that compromise the semantic layer of communication systems. These include embedding-level adversarial perturbations, GAN-based semantic jamming, latent backdoor injection, and covert prompt interception attacks. Together, they expose novel vulnerabilities inherent to generative AI-driven semantic communication. By analyzing the attack goals, mechanisms, and impact across different modalities, this study highlights the urgent need to reassess security paradigms in AI-native communication environments.

Keywords: Semantic Communication, Cybersecurity, Generative AI

1. Introduction

Semantic communication (SeCom) represents a fundamental shift in wireless communication, aiming to transmit meaning rather than raw symbols. By leveraging deep learning models and shared knowledge bases, SeCom improves efficiency in bandwidth-constrained and task-driven environments such as autonomous driving, edge computing, and the Metaverse [2, 1]. The integration of generative artificial intelligence (GenAI) models—including LLMs, diffusion networks, and VAEs—further enhances semantic fidelity, enabling multi-modal compression and prompt-based content reconstruction [3]. However, this reliance on AI-driven pipelines introduces new attack surfaces at the semantic layer, which remain insufficiently protected by traditional bit-level security models [6, 4, 8]. Similar concerns arise in other AI-enabled domains, such as intelligent transportation systems, where our group demonstrated a Time Tampering Black-Box Genetic Algorithm (TTB-GA) attack on digital twin models [7], and in energy harvesting (EH) networks, where AI-based security frameworks address threats like eavesdropping, data manipulation, and denial of service [5]. In this study, we focus exclusively on cyber attacks targeting the semantic layer, reviewing four recent and representative works on adversarial perturbations, GAN-based jamming, semantic backdoors, and covert prompt detection to highlight the evolving threat landscape of SeCom systems.

2. Attack Models in Semantic Communication

A white-box attack approach known as BERT manipulation of the embedding space is proposed by Hoang et al. [9]. Gradient-based perturbations are applied to the input text to generate adversarial examples that distort semantic interpretation for the receiver while preserving syntactic structure. These assaults specifically aim at communal knowledge-based semantic encoders, leading to considerable deterioration of semantic integrity. Experiments utilising the SNLI dataset demonstrated a decline in BLEU score from 0.72 to 0.46 under significant perturbation, but adversarial training only partially restored accuracy. The study highlights the vulnerability of semantic representation spaces and the simplicity with which they can be undermined.

Tang et al. [10] introduce a generative adversarial network (GAN)-based semantic jamming framework, where the attacker plays the role of a generator that learns to craft meaningful perturbations, while the

* Corresponding author:
Email addresses: MANaqvi9@dgu.ac.kr (Mehwish Ali Naqvi), isohn@dongguk.edu (Insoo Sohn)

SeCom receiver acts as a discriminator. This setup forms an adversarial game where the attacker's goal is to reduce semantic similarity without affecting signal detectability. The proposed attack strategy effectively reduces BLEU scores by up to 40% under varying interference levels and assumes access to model gradients. It demonstrates that intelligent jammers trained via adversarial learning can significantly impact semantic integrity in AI-native networks.

A hidden semantic backdoor attack (CSBA) developed for semantic communication in ICVs is introduced by Xu et al. [11]. In contrast to conventional input-level poisoning, CSBA embeds triggers into the latent semantic space, deliberately eliminating semantic attributes like road signs or people while maintaining visual fidelity. The assault employs Patchwise Latent Masking (PLaMa) and adversarial fine-tuning to guarantee that semantic tampering is undetectable. Assessment of the Cityscapes dataset indicates an attack success rate above 80% with no PSNR degradation, underscoring the stealth and accuracy of backdoors in the semantic feature space.

Du et al. [12] explore covert communication attacks in generative AI-aided semantic communication using multi-modal prompts. The attacker, modeled as a passive "warden," attempts to detect hidden semantic transmissions by observing structured prompt delivery. The system employs a benign jammer to mask transmission and uses a Generative Diffusion Model (GDM) to jointly optimize transmission power, jamming power, and diffusion steps for semantic reconstruction. The attack exploits binary hypothesis testing for detection, and experimental results show that without careful optimization, the warden can successfully identify prompt transmissions. The study highlights a new attack vector specific to prompt-based generative SeCom.

3. Conclusion

This research studied four recent cyberattacks that focus on the semantic layer of AI-native communication systems. In contrast to conventional threats that compromise signal quality or bit-level precision, these assaults directly alter meaning by embedding perturbations, generating adversarial noise, utilising latent backdoors, and employing covert prompt detection. Collectively, they unveil a varied and dynamic danger landscape in semantic communication, wherein adversaries might attain significant success rates without activating traditional alerts. With the growing implementation of SeCom systems in safety-critical and multi-modal applications, there is an imperative necessity to establish security frameworks that are intrinsically cognisant of semantic vulnerabilities. This work seeks to enhance the foundation by elucidating the assault surface and encouraging subsequent defence research.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00252328).

References

[1] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2678, 2021.

[2] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210–217, 2022.

[3] L. Xia, Y. Zhang, Y. Du, Y. Shi, Y.-C. Liang, and Z. Ding, "Generative AI for semantic communication: Architecture, challenges, and outlook," *IEEE Wireless Communications*, vol. 32, no. 1, pp. 60–67, 2025.

[4] Y. E. Sagduyu, T. Erpek, S. Ulukus, and A. Yener, "Is semantic communication secure? A tale of multi-domain adversarial attacks," *IEEE Communications Magazine*, vol. 61, no. 11, pp. 50–55, 2023.

[5] M. Mohammadi and I. Sohn, "Security in energy harvesting systems using artificial intelligence: A survey," *ICT Express*, vol. 9, no. 4, pp. 527–540, 2023.

[6] Z. Yang, D. Niyato, X. Xie, Y. Lyu, X. Cao, and L. Liang, "Secure semantic communications: Fundamentals and challenges," *IEEE Network*, vol. 38, no. 6, pp. 104–110, 2024.

[7] M. Pooyandeh, H. Liu, and I. Sohn, "Cybersecurity in digital twins of electric vehicle's LIBs: Unveiling a robust TTB-GA attack," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 4, pp. 5360–5381, 2025.

[8] Q. T. Do, D. Won, T. S. Do, T. P. Truong, and S. Cho, "Security and privacy challenges in semantic communication networks," in *Proc. Int. Conf. on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 1–6, 2025.

[9] V.-T. Hoang, V.-L. Nguyen, R.-G. Chang, P.-C. Lin, R.-H. Hwang, T. Q. Duong, Adversarial Attacks Against Shared Knowledge Interpretation in Semantic Communications, IEEE Transactions on Cognitive Communications and Networking, vol. 11, no. 2, pp. 1024–1040 (2025).

[10] R. Tang, D. Gao, M. Yang, T. Guo, H. Wu, G. Shi, GAN-inspired Intelligent Jamming and Anti-jamming Strategy for Semantic Communication Systems, 2023 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1623–1628 (2023).

[11] X. Xu, Y. Chen, B. Wang, Z. Bian, S. Han, C. Dong, C. Sun, W. Zhang, L. Xu, P. Zhang, CSBA: Covert Semantic Backdoor Attack Against Intelligent Connected Vehicles, IEEE Transactions on Vehicular Technology, vol. 73, no. 11, pp. 17923–17928 (2024).

[12] H. Du, G. Liu, D. Niyato, J. Zhang, J. Kang, Z. Xiong, B. Ai, D. I. Kim, Generative AI-aided Joint Training-free Secure Semantic Communications via Multi-modal Prompts, ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 12896–12900 (2024).