The Model Inversion Eavesdropping Attack in Semantic Communication Systems

Yuhao Chen, Qianqian Yang[†], Zhiguo Shi, Jiming Chen
College of Control Science and Engineering, Zhejiang University, Hangzhou 310007, China
College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310007, China
The State Key Laboratory of Industrial Control Technology, Hangzhou 310007, China
{csechenyh, qianqianyang20[†], shizg, cjm}@zju.edu.cn

Abstract—In recent years, semantic communication has been a popular research topic for its superiority in communication efficiency. As semantic communication relies on deep learning to extract meaning from raw messages, it is vulnerable to attacks targeting deep learning models. In this paper, we introduce the model inversion eavesdropping attack (MIEA) to reveal the risk of privacy leaks in the semantic communication system. In MIEA, the attacker first eavesdrops the signal being transmitted by the semantic communication system and then performs model inversion attack to reconstruct the raw message, where both the white-box and black-box settings are considered. Evaluation results show that MIEA can successfully reconstruct the raw message with good quality under different channel conditions. We then propose a defense method based on random permutation and substitution to defend against MIEA in order to achieve secure semantic communication. Our experimental results demonstrate the effectiveness of the proposed defense method in preventing MIEA.

I. Introduction

Recently, semantic communication has been widely believed to be one of the core technologies for the sixth generation (6G) of wireless networks because of its high communication efficiency [?]. Compared with the current research on communication which focuses on transmitting mapped bit sequences of the raw message [?], [?], [?], semantic communication systems transmit compacted semantic features. Existing literature in semantic communication mainly exploits the deep learning (DL) techniques to extract the semantic features from the raw message. For instance, Han et al. [?] proposed to extract the text-related features from the speech signal as the semantic features and remove the redundant content. On the receiver's side, the semantic features can be reconstructed by a deep learning model into the original message or directly applied for downstream tasks such as image classification and speech recognition.

Although many works have been proposed for semantic communication considering different aspects, few studies have taken into account the security problems [?], [?], [?]. Tung et al. [?] proposed to encrypt the transmitted signal in semantic communication, but the encryption algorithm incurs a large computation overhead. Security is crucial in semantic communication for two main reasons. Firstly, semantic communication is more prone to privacy leakage compared to traditional communication. In traditional communication systems, the bit sequences being transmitted contain redundant bits to ensure reliable transmission, which can be used to provide a certain level of privacy protection. However, the semantic communication systems transmit compact and more semantic-related symbols which may reveal more private information. Secondly, deep-learning-based semantic communication may be vulnerable to attacks targeting DL models. Extensive studies have been conducted on attacks on the DL model, a review of which can be referred to [?]. If the semantic features being transmitted are eavesdropped by a malicious attacker, the attacker can reconstruct the raw message by utilizing the DL-based attack techniques. The attacker can also add perturbation to the transmitted data, causing the semantic communication system to make incorrect decisions on downstream tasks. For example, Sagduyu et al. [?] proposed a multi-domain evasion attack to cause the semantic communication system to make incorrect classifications, which is achieved by introducing noises to input images or the semantic features. Du et al. [?] proposed a semantic data poisoning attack, which causes the receiver to receive irrelevant messages from the transmitter. For example, the receiver wants to receive an image with a pear but gets an image with an apple instead. This attack is performed by minimizing the difference between the semantic features of the targeted message and the irrelevant message.

In this paper, we consider the security issue in semantic communication systems and introduce the model inversion eavesdropping attack (MIEA) for semantic communication, where an attacker eavesdrops the transmitted symbols and attempts to reconstruct the original message from them by inverting the DL model used at the transmitter. We perform MIEA under both the white-box and the black-box settings. The attacker has knowledge of the DL model in the white-box setting while not in the black-box setting. To defend against MIEA, we also propose a defense method based on random permutation and substitution. Evaluations demonstrate that the MIEA attack works under different channel conditions, i.e., different values of the signal-to-noise ratio (SNR), which reveals the risk of privacy leaks in semantic transmission. Numerical results also validate the effectiveness of our proposed

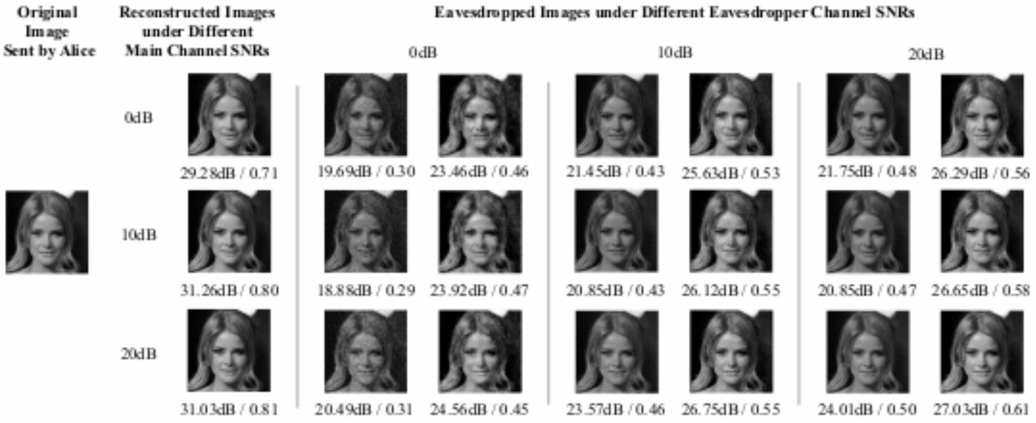

Fig. 3. Visualization of MIEA for the white-box attack and the black-box attack under different channel conditions. For each channel condition, the eavesdropped image by the white-box attack is displayed on the left and the one obtained by the black-box attack is on the right.

of the images received by Bob, the eavesdropped images are visually recognizable and their privacy is compromised, which confirms the effectiveness of MIEA and reveals the risk of privacy leaks in current semantic communication.

TABLE I
The average SSIM and PSNR of the eavesdropped images under different channel conditions

	Main Channel SNR 0dB	 10dB	 20dB
Reconstructed Images by Bob	30.02dB / 0.70	32.28dB / 0.79	33.28dB / 0.80
EC[1]SNR 0dB	17.44dB / 0.31 21.65dB / 0.46	16.69dB / 0.30 22.24dB / 0.46	18.58dB / 0.32 22.85dB / 0.46
EC SNR 10dB	18.58dB / 0.40 23.50dB / 0.53	17.80dB / 0.40 23.76dB / 0.54	20.56dB / 0.43 24.33dB / 0.56
EC SNR 20dB	18.74dB / 0.43 23.88dB / 0.57	17.94dB / 0.42 23.98dB / 0.59	20.84dB / 0.46 24.41dB / 0.61

[1] EC refers to the eavesdropper channel.

C. Evaluation of the Proposed Defense Method

Next, we evaluate the proposed defense method by repeating the evaluation of MIEA in section ?? with the defense method implemented on $\boldsymbol{y}_{\mathrm{f}}$.

Fig. ?? visualizes the eavesdropped images by the two attack types after applying the proposed defense method, using the same individual as in Fig. ??. It can be observed that the eavesdropped images are visually unrecognizable, demonstrating the effectiveness of the proposed defense method in preventing Eve from eavesdropping on raw images. We can also see that the contour of the female in the white-box attack is less obvious than that in the black-box attack, suggesting that the defense against the white-box attack is superior to that against the black-box attack. This is because that Eve has no prior knowledge of the defense method when performing the white-box attack, whereas $f^{-1}(\cdot)$ used in the black-box attack has learned some knowledge of the defense method from the training samples.

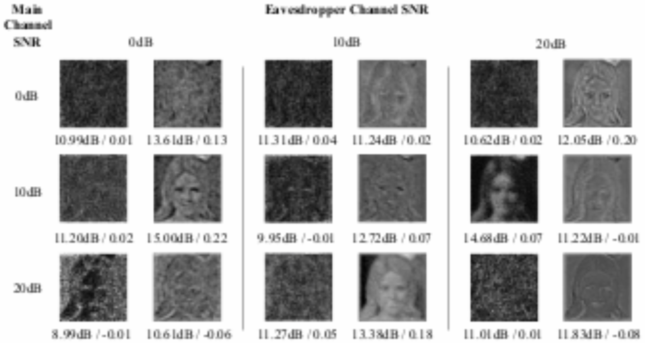

Fig. 4. Visualization of MIEA for both attacks under different channel conditions after applying the proposed method.

TABLE II
The average SSIM and PSNR of the eavesdropped images by MIEA after defense

MC[1] EC	0dB	10dB	20dB
0dB	8.03dB / 0.02 11.36dB / 0.11	8.74dB / 0.02 11.41dB / 0.07	6.94dB / 0.00 12.51dB / 0.07
10dB	8.55dB / 0.04 11.72dB / 0.16	7.70dB / -0.01 11.34dB / 0.11	9.07dB / 0.05 13.22dB / 0.13
20dB	8.02dB / 0.03 12.59dB / 0.21	13.31dB / 0.09 11.55dB / 0.10	8.39dB / 0.02 11.54dB / 0.07

[1] MC refers to the main channel.

In addition, we provide the average SSIM and PSNR of the eavesdropped images for both attacks in Table. ??. For a given SNR of the main channel, the SSIM and PSNR do not increase as the SNR of the eavesdropper channel increases because different P and S are used for different transmitted features. The average SSIM and PSNR of the eavesdropped images by the black-box attack are larger than those by the white-box attack, which is consistent with the observation from Fig. ??. Overall, the average SSIM and PSNR are relatively small, which indicates the effectiveness of the proposed defense method in preventing Eve from obtaining meaningful information from the eavesdropped signal.

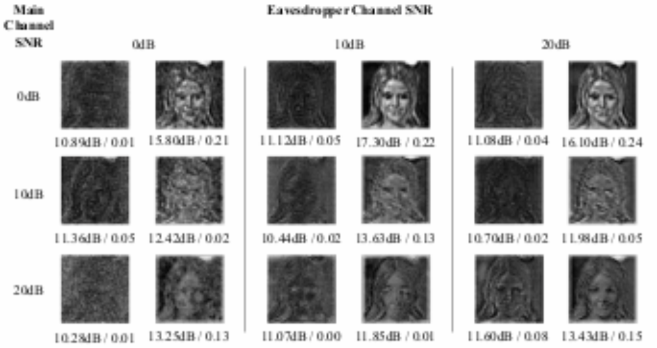

Fig. 5. Visualization of MIEA for both attacks under different channel conditions after applying only the random permutation.

TABLE III
The average SSIM and PSNR of the eavesdropped images when applying only random permutation

MC EC	0dB	10dB	20dB
0dB	8.05dB / 0.02 14.00dB / 0.22	8.73dB / 0.06 11.43dB / 0.10	8.50dB / 0.02 13.25dB / 0.13
10dB	8.32dB / 0.06 14.55dB / 0.26	8.02dB / 0.03 12.97dB / 0.19	8.87dB / 0.01 13.19dB / 0.12
20dB	8.36dB / 0.04 14.97dB / 0.28	8.05dB / 0.02 12.83dB / 0.18	8.77dB / 0.07 12.83dB / 0.13

Next, we conduct an ablation study to further validate our proposed method. Fig. ?? and Table. ?? demonstrate the eavesdropped images and the average SSIM and PSNR for both attacks by applying only the random permutation. As shown in Fig. ??, when only the random substitution is applied, the white-box attack can be effectively defended, while the black-box attack can still reconstruct visually recognizable images for some P and S. Moreover, the average SSIM and PSNR for the black-box attack in Table. ?? are larger than those in Table. ??, which demonstrates that only random permutation is insufficient for defending against MIEA.

Fig. ?? and Table. ?? show the related results for both attacks by applying only the random substitution. For both attacks, most of the eavesdropped images are visually recognizable, indicating that the attacker can still obtain sensitive information from the transmitted symbols, even though some of the semantic features have been substituted. The average SSIM and PSNR in Table. ?? are larger than those in Table. ??, which means that the random permutation is more effective than random substitution in defending against MIEA. From the ablation study, we can observe that the proposed defense method outperforms both the random-permutation-based and random-substitution-based defense methods, demonstrating that both permutation and substitution are essential for the effectiveness of the proposed defense method.

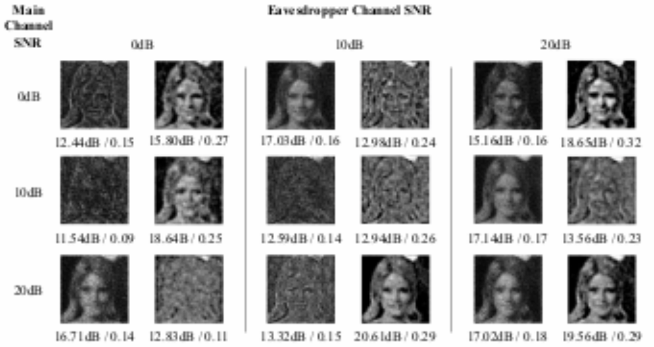

Fig. 6. Visualization of MIEA for both attacks under different channel conditions after applying only the random substitution.

TABLE IV
The average SSIM and PSNR of the eavesdropped images when applying only random substitution

MC EC	0dB	10dB	20dB
0dB	8.99dB / 0.14 16.06dB / 0.28	8.91dB / 0.10 15.00dB / 0.25	15.80dB / 0.16 14.70dB / 0.20
10dB	15.62dB / 0.19 14.68dB / 0.26	9.16dB / 0.14 14.49dB / 0.26	10.43dB / 0.16 15.80dB / 0.27
20dB	12.68dB / 0.18 15.78dB / 0.30	15.35dB / 0.18 14.53dB / 0.27	16.13dB / 0.21 17.29dB / 0.28

V. Conclusion

In this paper, we propose MIEA to expose privacy risks in semantic communication. MIEA enables an attacker to eavesdrop on the transmitted symbols through an eavesdropper channel and reconstruct the raw message by inverting the DL model employed in the semantic communication system. We consider MIEA under the white-box attack and the black-box attack and propose a novel defense method based on random permutation and substitution to defend against both types of attack. In our evaluation, we first examine MIEA for both attacks under various channel conditions. We then conduct experiments and an ablation study to demonstrate the effectiveness of our proposed defense method.