SemBAT: Physical Layer Black-box Adversarial Attacks for Deep Learning-based Semantic Communication Systems

Zeju Li, Jinfei Zhou, Guoshun Nan*, Zhichun Li, Qimei Cui, Xiaofeng Tao
Beijing University of Posts and Telecommunications
lizeju@bupt.edu.cn, zhouchuluo@bupt.edu.cn, nanguo2021@bupt.edu.cn, monstry@bupt.edu.cn,
cuiqimei@bupt.edu.cn, taoxf@bupt.edu.cn

***Abstract*—Deep learning-based semantic communications (DLSC) replace the physical blocks in traditional communication systems as end-to-end neural networks. DLSC significantly boost communication efficiency by only transmitting the meaning of data, showing great potentials for applications like automatic driving, digital twin and smart health. However, DLSC are fragile to black-box adversarial attacks due to the openness of wireless channel and sensitivities of neural models. To this end, this paper proposes SemBAT, a novel approach for crafting physical layer black-box adversarial attacks for semantic communication systems. The key ingredients of our method include the training of surrogate encoder and generation of adversarial perturbations. Specifically, we train our surrogate encoder by directly estimating the gradients based on Jacobian-matrixs, and then generate the adversarial perturbations by the particle swarm optimizations. Extensive experiments on a public benchmark show the effectiveness of our proposed SemBAT. We observe that our SemBAT with black-box adversaries can sharply decrease the classification accuracy of the semantic communication system from 78.4% to 11.6%. Meanwhile, such attacks are also imperceptible in terms of image quality metrics measured by the Structural similarity index measure (SSIM) and Peak Signal to Noise Ratio(PSNR).**

***Index Terms*—Semantic communications, Black-box attacks, Particle swarm optimizations**

I. Introduction

Shannon and Weaver [1] discussed that communication can be divided into three levels: transmission of symbols; semantic exchange of transmitted symbols; the impact of semantic information exchange. When people conduct related research on communication systems based on Shannon's information theory, they mainly focus on issues at the grammatical level, and only aim to transmit bit data reliably and efficiently. Today, the problems related to the reliability and effectiveness of communication have been solved. With the increasingly close integration of artificial intelligence technology and communication technology, the problem of semantic level that was temporarily put on hold in the past has re-emphasized. Unlike traditional communication, deep-learning semantic communication systems (DLSC) [2]–[5] aim to transmit information related to the transmission target. Although promising, DLSC are fragile to adversarial attacks [6], [7] due to the openness of wireless channel and sensitivities of neural models.

*Corresponding author

Adversarial attacks [6] are mainly divided into white-box attacks and black-box attacks. For the former, the attacker knows all the information and parameters in the model, generates adversarial samples based on the gradient of the given model, and attacks the network. For the later black-box attack, the attacker does not have the knowledge of the parameters and structure information of the model. The attacker can only defeat the systems through the input and output of the model with adversarial samples. Hence, it is much more challenging to craft adversaries for a neural network system under the black-box setting. However, the underlying ideas of the two are consistent, and the gradient information is used to generate adversarial samples, so as to achieve the purpose of deceiving the network model. Taking advantage of this feature, we can train a surrogate model that can mimic the output of the target model, and then we are able to use such a surrogate encoder as a transmitter to communicate with the receiver in the semantic communication system. Traditional black-box attacks can use query-based methods to add noise to the input image directly by using various optimization algorithms to query the output to find adversarial samples.The other is training a surrogate model with a decision boundary similar to the original model and performing a white-box attack on the surrogate model, namely calculating the gradient information to add noise directly. Obviously, the first method requires a lot of queries to achieve the goal. Even if the optimization algorithm is efficient, there is an inevitable problem of low query efficiency. Inefficient queries can easily lead to the problem of being easily perceived as an attack and it is not concealed enough.Considering the openness of wireless channels in semantic communication system, we add noise to the channels. It is impossible to transmit information directly on the original transmitter, so we use the surrogate model to simulate transmitting and conduct a query-based black box attack on the channel by cooperating with output information, which can enhance the query efficiency greatly.

In this paper, we propose SemBAT, a novel method for generating physical layer black-box adversarial attacks for deep learning-based semantic communication systems. We train our surrogate encoder with gradient estimation through Jacobian-based Dataset Augmentation [8], and then we use the contextualized representations and labels from the receiver to

optimize a noise generation model through the Particle Swarm Optimization(PSO) algorithm [9]. We add noise disturbance in the process of passing the representations to the wireless channel, so that the accuracy of the classifier of the semantic communication model can be improved. As the number of training iterations increases, the classification accuracy of the model can be decreased to 11.6%. Meanwhile, we observe that such attacks are also imperceptible, as the image quality metrics measured by the Structural similarity index measure (SSIM) and Peak Signal to Noise Ratio(PSNR) are slightly dropped.

Specifically, we summarize our contributions as follows:

- We introduce SemBAT, a novel method that aims to generate physical layer black-box adversarial attacks for semantic communication systems.
- We train our surrogate encoder with gradient estimation and data augmentation based on Jacobian matrix, properly tackling parameter learning under the black-box scenario. We craft adversarial perturbations with the particle optimization algorithm that is adapted to our scenario.
- We conduct experiments on a public dataset to show the effectiveness of our SemBAT. Experimental results show that our black-box attacks can significantly decrease the classification accuracy, while slightly reduce the image qualities.

II. Model

In this section, we first introduce a semantic communication system, and then present our proposed SemBAT, which includes two key components surrogate encoder and noise generator. Figure 1 shows the architecture of the system.

A. Semantic Communication System

We select JSCC-OFDM semantic communication system [10] as our backbone and also introduce a classifier to obtain the output label. The system uses joint source channel coding (JSCC) for wireless image transmission over multipath fading channels. As shown in Figure 3, the semantic encoder directly maps the source images to complex-valued baseband samples for OFDM transmission with CSI feedback which is shown in Figure 2. We apply OFDM as the modulation technique to resist interference from multipath fading channels with frequency domain equalization for this JSCC framework. We concatenate deep neural networks with OFDM processing blocks by feeding the neural network encoded image as frequency domain OFDM baseband symbols. On the decoder side, the interference due to irrational characteristics of the channel is addressed by channel estimation, equalization and additional subnets. The previous semantic communication system generally ends up with image reconstruction. We introduce classifiers for joint training : image transmission to the opposite end generally has downstream tasks and classification is the most common scene. Compared with the separated communication and classification models, the semantic communication system with classifier has more advantages in speed and model size. Instead of naively treating a neural network as a black box, the neural network imposes a guided structure to introduce the traditional communication analysis model and signal processing module designed on both the encoder and decoder sides. We simply use convolutional neural networks (CNNs) as our surrogate encoder. Next, we show how we train such a surrogate encoder.

B. Surrogate Encoder

We train our surrogate encoder with gradient estimation [11] to facilitate an attacker for the generation of adversarial examples. The encoder can also ensure the privacy of data acquisition, as we don't have any knowledge of the original encoder of the semantic communication systems. The surrogate encoder generates the contextualized semantic information and then sends them to the receiver. The training of our proposed surrogate encoder consists of three steps:

1) In the first step, we design the structure of our surrogate encoder, which is shown in Figure 4. In this paper, we use full convolution CNNs. The convolution structure is suitable for the extraction and compression of semantic information. We make the output of the surrogate model has the same dimension as the high-dimensional vector encoded by the original encoder, to ensure it can have a similar decision boundary with the target model when connecting with the decoder.
2) In the second step, we simplify the process of obtaining the dataset during the experiment. The dataset is divided into two parts. We train the surrogate encoder with the data subset of 10,000 images. We use a data subset of 1000 images to augment the data with data augmentation, such as flips, translations, rotations and other minor changes, here we can use

$$S_{\rho+1} = \{\vec{x} + \mu sgn(J_F[\tilde{O}(\vec{x})]) : \vec{x} \in S_\rho\}$$

$\vec{x}$ is an image in S_ρ, $\tilde{O}(\vec{x})$ is the image label output by the target model, $J_F[\tilde{O}(\vec{x})]$ is the Jacobian-Matrix corresponding to this label, $sgn(J_F[\tilde{O}(\vec{x})])$ is the symbol of only the Jacobian matrix, namely the element value in the matrix becomes a regular 1, otherwise it becomes -1. It can be seen that the new $S_{\rho+1}$ is composed of the original S_ρ and the enhanced $\{\vec{x} + \mu sgn(J_F[\tilde{O}(\vec{x})])\}$, and the size becomes twice the original. This step can be combined with the third step.
3) In the third step, we adapt the zeroth order optimization [12] method for the gradient estimation. We estimate the parameter update on the basis of two samplings by differences, without relying on the first derivative information for gradient computation. Given an image x_0, let x denote the adversarial example of x_0, x_0 class label points to the misclassification. We find x by solving :

$$minimize_x \|x - x_0\|_2^2 + c * f(x)$$

$$subject\ to\ x \in [0,1]^p$$

$\|x - x_0\|_2^2$ denotes the Euclidean norm of $x - x_0$, $c > 0$ is a regularization parameter. Hence $x - x_0$

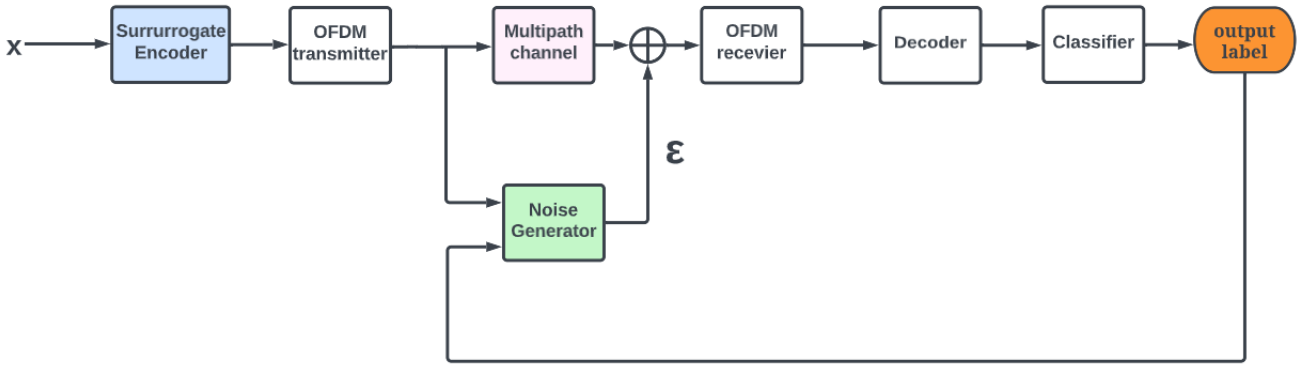

Fig. 1. System architecture. The key ingredients of our proposed SemBAT include a surrogate encoder and a noise generator, where the former is trained by gradient estimation and later is optimized with the particle algorithm.

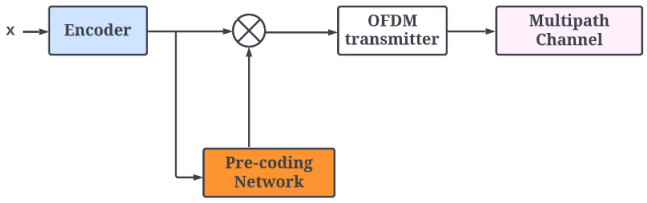

Fig. 2. Encoder structure with CSI feedback.CSI feedback can significantly improve the error rate because of the increased error exponent using dynamic power allocation algorithms and adaptive modulation.

is the adversarial image perturbation of x relative to x_0. $c * f(x)$ is the loss function that reflects the level of unsuccessful adversarial attacks. The output of our neural network $F(x)$ is determined by the softmax function:

$$[F(x)]_k = \frac{exp([Z(x)]_k)}{\sum_{i=1}^{K} exp([Z(x)]_i)} \quad \forall k \in (1, ..., K)$$

$Z(x) \in R^K$ is the logit layer representation in the neural network for x such that $[Z(x)]_k$ represents the predicted probability that x belongs to class k, We select untargeted attack, so we use the loss function:

$$f(x) = max\{log[F(x)]_{t_0} - \max_{i \neq t_0} log[F(x)]_i, -\kappa\}$$

t_0 is the original class label for x, and $\max_{i \neq t_0} log[F(x)]_i$ represents the most probable predicted class other than t_0. We use the symmetric difference quotient to estimate the gradient $\frac{\partial f(x)}{\partial x_i}$:

$$\frac{\partial f(x)}{\partial x_i} \approx \frac{f(x + he_i) - f(x - he_i)}{2h}$$

We set h=0.0001 and e_i is a standard basis vector. In this step, we train the surrogate encoder, and keep the parameters and structure of the decoder unchanged in Figure 5.

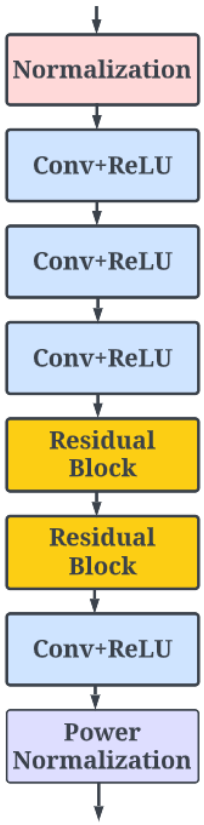

Fig. 3. Encoder structure

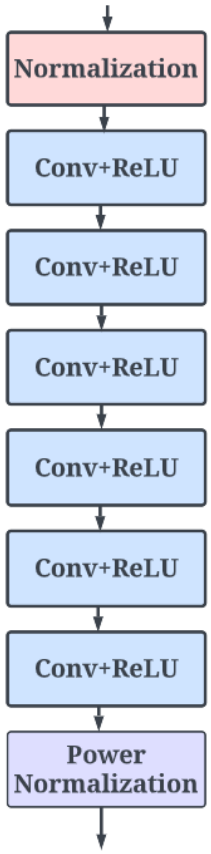

Fig. 4. Surrogate encoder structure

C. Noise Generator

As shown in Figure 1, we add adversarial noise to signals in the physical layer, and such a way of black-box adversarial attacks is quite different from the ones that inject attacks from the input. Our surrogate encoder is able to generate

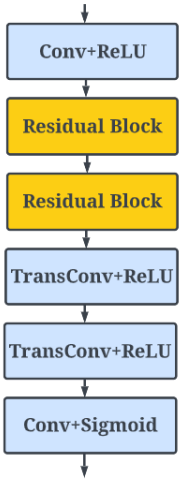

Fig. 5. Decoder structure

contextualized representations, and then we can get the labels from the classifier at the receiver side. To craft more practical black-box attacks, we add two constraints for the noise, i.e., SSIM [13] should be greater than a threshold and ACC should be less than a threshold. For example, we set the two thresholds as 80% 15% respectively. We optimize our noise generator based on our proposed particle swarm optimization algorithm.

We generate a random noise ϵ as a particle, w is the momentum factor, v_i is the velocity of particles, $rand()$ is a random number between 0 and 1, x_i is the current position of the particle, c_1 and c_2 is learning rate, $pbest_i$ is the best position searched by one particle, $gbest_i$ is the best position searched by the swarm so far

$$v_i = w*v_i+c_1*rand()*(pbest_i-x_i)+c_2*rand()*(gbest_i-x_i)$$

$$x_i = x_i + v_i$$

Every particle is iteratively updated by pbest and gbest. By doing so in Alg 1, we are able to obtain adversarial perturbations to the signals in wireless channel.

Algorithm 1: Training the noise generate

Input: input image x;
Output label of classifier y
Output: A high dimensional vector noise

Initialize noise ϵ ;
Sample batch of dataset $\mathcal{T}_i$;
while ACC¡15% SSIM¿80% **do**
  latent=surrogate_encoder(x);
  latent +=noise;
  foreach $\mathcal{T}_i$ *in dataset* **do**
    ϵ=PSO(latent,y);
    update(y)
    Compute ACC;
  Update ϵ ;

III. Experiment

We conduct experiments on CIFAR10, one of the most popular datasets that is used for image classifications. CIFAR10 consists of 60,000 32×32-pixel images in 10 classes, with 6,000 images per class. We use 50,000 images to train the targeting semantic communication system and the rest of the 10,000 ones are used to train the surrogate encoder. We employ several metrics to check the performance of our method, namely Peak Signal to Noise Ratio(PSNR) [14], structural similarity index measure (SSIM), average power ratio(PAPR) and accuracy of classification(ACC).

A. Implementation details

We use Pytorch to implement the neural network blocks and the OFDM communication model. We utilize Adam [15] to train our surrogate encoder, generator, and reconstruction net from the encoder, and retain the decoder structure for end-to-end parameter migration from the original communication system. The initial learning rates of all networks are set as 0.0005, they are gradually decreased to zero as the number of iterations increases. Our model is trained by one NVIDIA GeForce RTX3090 GPU. Each source image x is transmitted in a single OFDM packet that contains N_p pilot symbols and N_s information symbols. We select SNR =10dB , $N_p = 1$ and $N_s = 6$. The experiments show that the SSIM is above 0.9 and the PSNR is above 28 under this condition, which can ensure the reliability and effectiveness of communication.

B. Performance of surrogate encoder

TABLE I
Training effect of surrogate model

	Original Encoder	Surrogate Encoder
PSNR	25.23	24.27
SSIM	0.85	0.83
PAPR	11.58	11.91
ACC	0.83	0.78

Table I reports the performance comparisons between the surrogate encoder and the original encoder. We observe that our surrogate encoder, which is trained on a small part of dataset with gradient estimation, yields very similar performance to the original model in terms of PSNR, SSIM and PAPR. Hence, we can use this surrogate encoder to mimic the original encoder to facilitate the generation of black-box adversarial perturbations.

C. Black-box Adversarial Attacks

Table II shows the results under different iterations of physical layer black-box adversarial attacks. Through the experimental results, we can find that as the number of iterations increases, the attack effect of the noise generator gradually increases. We observed the accuracy of the classifier decreases to 11.6% after 50 iterations, while PSNR, PAPR and SSIM

TABLE II
EFFECT OF NOISE GENERATOR WITH DIFFERENT ITERATIONS

	No noise	10 iterations	30 iterations	50 iterations
PSNR	24.27	23.28	24.22	24.93
SSIM	0.83	0.81	0.82	0.81
PAPR	11.91	10.94	11.32	10.30
ACC	0.78	0.72	0.47	0.11

are slightly dropped. The results indicate that our black-box attacks are destructive and imperceptible.

IV. CONCLUSION

In this paper, we present SemBAT, a novel method that aims to generate physical layer black-box adversarial attacks for semantic communication systems. We train our surrogate encoder with gradient estimation and then optimize the noise generator to craft the adversarial perturbations. Experimental results show that our SemBAT is effective to significantly decrease the classification accuracy. Meanwhile, the attacks are imperceptible for humans as the image quality is slightly dropped. In the future, we will conduct more experiments to generate various physical layer black-box adversarial attacks, and evaluate their performance on more semantic communication systems.

V. ACKNOWLEDGMENTS

This work is supported by the Joint Funds for Regional Innovation and Development of the National Natural Science Foundation of China (No. U21A20449) , the National Youth Top-notch Talent Support Program, and Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning based semantic communications: An initial investigation," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

[3] ——, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[4] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li, "Robust semantic communications against semantic noise," in *2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*. IEEE, 2022, pp. 1–6.

[5] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech signals," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[8] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[9] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4. IEEE, 1995, pp. 1942–1948.

[10] M. Yang, C. Bian, and H.-S. Kim, "Ofdm-guided deep joint source channel coding for wireless multipath fading channels," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 584–599, 2022.

[11] S. Shen, Q. Liu, E. Chen, H. Wu, Z. Huang, W. Zhao, Y. Su, H. Ma, and S. Wang, "Convolutional knowledge tracing: Modeling individualization in student learning process," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1857–1860.

[12] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.

[13] L.-T. Wang, N. E. Hoover, E. H. Porter, and J. J. Zasio, "Ssim: A software levelized compiled-code simulator," in *Proceedings of the 24th ACM/IEEE Design Automation Conference*, 1987, pp. 2–8.

[14] U. Sara, M. Akter, and M. S. Uddin, "Image quality assessment through fsim, ssim, mse and psnr—a comparative study," *Journal of Computer and Communications*, vol. 7, no. 3, pp. 8–18, 2019.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.