A GAN-based Covert Channel over Autoencoder Wireless Communication Systems

Ali Mohammadi Teshnizi
Dept. Computer Science
University of Calgary
ali.mohammaditeshniz@ucalgary.ca

Majid Ghaderi
Dept. Computer Science
University of Calgary
mghaderi@ucalgary.ca

Dennis Goeckel
Dept. Electrical and Computer Engineering
University of Massachusetts Amherst
goeckel@ecs.umass.edu

***Abstract*—Use this template when writing a conference paper.**

I. Introduction

Due to the shared and broadcast nature of wireless channels, there is considerable attention on the security and privacy aspects of wireless communications. While traditional cryptography methods and physical-layer securities can protect the confidentiality of the content (i.e the information transmitted over the channel), there are occasions that hiding the very existence of the communication channel is more vital than securing the communicated message itself. Examples of such situations are military operations, cyber-espionage, social unrest, or communication parties' privacy. All above have motivated the study of hidden communication channels, namely, "covert channels" [1].

The preliminary attempt to obtain covertness started with the study of spread spectrum, which began almost a century ago with the main purpose of hiding military communications. The idea is to spread the transmit power into the noise so that the communication stays hidden. Many works continued to further examine different aspects of this idea, yet the fundamental performance limits of such work were unknown until recently when Bash et al. [2] established a square root limit on the number of covert bits that can be reliably sent over an additive white Gaussian noise (AWGN) channel while remaining covert to a channel's observer. Followed by this work, there has been a surge of interest in examining covert channels [3], [4] especially in point-to-point wireless communication models.

In the last decade, majority of the research works on covert communication focus on integration of covert communication techniques into the current and future wireless technologies, such as 6G wireless networks and IoT. With the advent of AI and recent advances in computer's computation power and algorithmic designs, machine learning (ML) is now becoming an integral part of many research topics. Correspondingly, wireless community started to adopt machine learning techniques and use them as solutions to the various network optimization problems, which were traditionally handled with statistical models. Deep neural networks (DNNs) in particular, a major force in machine learning, are now applied into many of these tasks including signal classification, channel estimation, transmitter identification, jamming and anti-jamming [?].

More recently, as a replacement for conventional modular-based designs, an end-to-end communication model based on deep learning methods has emerged [5]. In this new paradigm, the transmitter and receiver are jointly trained as the encoder and decoder of an autoencoder network. While the transmitter learns to encode the transmitted symbol to an embedding vector of wireless signals, the decoder, on the other side, learns to retrieve that symbol back from the transmitted signal after passing through the channel. One noticeable difference between end-to-end systems and conventional modular designs is that in end-to-end systems, the encoder learns coding and modulation tasks simultaneously as opposed to having separate modules to perform each task. This is also true for the decoder of the system that jointly learns demodulation and decoding. Given the fact that channel distortions usually have a non-linear behavior, a DNN-based communication model is expected to capture these dissimilitudes better compared to a statistical model. That means, these systems can learn more complex channel effects and can operate more robustly under different noise levels and channel imperfections [6].

II. Related Works

The main idea of our work stems from stenography techniques, which are commonly used in image steganography. Hence, we first talk about the history and current state of this field of research and then we continue reviewing some of the existing approaches for establishing covert communication at the physical layer in a wireless network.

Deep learning algorithms have proved their efficiency in many aspects. Steganography is one of these areas that has benefited tremendously from deep learning advancements in recent years. The earliest use cases were in steganalysis research. Convolution neural networks (CNNs) for instance, which are generally used in computer vision tasks, showed outstanding results in image steganalysis [7]–[9], replacing the traditional statistical methods. One of the earliest works of image steganography using deep neural networks is a work by [10]. In this work, Baluja proposes a hiding scheme in which the three networks of preparation, hiding, and reveal sort out the secret encoding and decoding task. The preparation network transforms the hidden message into features that are commonly used for compressing images. Then, the hiding network embeds it into the cover image and sends it to the reveal

network, where the secret message gets extracted from the container image. Following this work, it was discovered that the existence of preparation network is not necessary and the framework can be expressed in a simpler form by excluding this network [11]. The disadvantage of these schemes is that the encoding process is reliant on the cover image. To address this, Zhang et al. [12] propose a new architecture in which secret message can be encoded independent of the cover image. Beside having more flexibility on hiding the information, this approach has also become an effective method for image watermarking. To manifest robustness against steganalysis practices, researchers started to adopt generative adversarial network (GAN) architectures. A typical adversarial network consists of two neural networks, a generator network (G) and a discriminator network (D). These two networks are then trained against each other where the generator tries to deceive the discriminator by generating data similar to those of in the training set and the discriminator tries to correctly categorize them as fake and real [13]. Volkhonskiy et al. [14] propose a steganography technique based on GAN networks. The main idea of their work is to use a generative network to produce a new set of cover images that when carries the secret message using any of the available steganography techniques will be less exposed to be detected by a discriminator network (i.e. a steganalysis network). Similarly, Hayes et al. [15] introduce a GAN-based steganography technique with a different objective for the generator network. Instead of generating cover images, the generator learns to embed secret messages into the cover images so that the discriminator cannot distinguish cover images from stenographic images. Although this adversarial scenario was preliminary introduced for hiding data in images, researchers found it so versatile that it has been applied into the other forms of data such as video, audio and recently wireless signals.

Numerous works have studied the theoretical limits of covert communication over wireless channels in different scenarios [2], [4], [16], [17], but only few works have focused on a practical implementation of such channels. One real world example of a covert communication is the work by Dutta et al. [18]. They leverage the communication noise caused by the channel or the hardware imperfections to establish a covert channel. In their proposed method, messages are covertly encoded in the constellation error of the normal cover signals. Similarly, Cao et al. [19] further improve this method with the goal of reducing probability of detection. Hou et al. [20] propose an amplitude based covert channel over LoRa PHY. In their scheme, covert information is embedded with a modulation scheme orthogonal to chirp spread spectrum (CSS). Bonati et al. [21] introduce SteaLTE which is a full-stack wireless steganography method on softwarized cellular networks. To covertly modulate symbols, they employ three different approaches of dirty QPSK modulation, hierarchical amplitude shift keying (ASK) manipulation, and phase offset of the primary symbols modification. The disadvantage of all the above methods is that the distortions caused in the statistical properties of the system can be detected with high confidence by a careful observer. More recent works have explored the viability of deep neural networks in covert communication problem. Sankhe et al. [22] propose a method called Impairment Shift Keying that produces subtle variations in normal signals in a controlled way such that a CNN model can be trained to classify them as zeros or ones. To achieve the highest covert rate while minimizing the probability of detection, Liao et al. [23] employ a GAN model that can adaptively adjust the signal power at the covert sender. Motivated by the GAN-based steganography technique, Mohammed et al. [24] formulate the covert communication as a three-player game in which three networks compete against each other. In this setup, the encoder and decoder networks learn to covertly communicate through a form of noise and simultaneously try to confuse a detector network that tries to differentiate expected noise of the system from the covert added perturbations. While our proposed method shares the same idea as of this work, there are a couple of deficiencies in the previous work that out work aims to address. First, we study covert communication over Autoencoder DNN-based wireless systems which are believed to replace the traditional modular systems in future wireless communication technologies such as 6G [?]. Second, previous work assumes that the modulation type is known to the covert receiver; thus, the covert receiver knows what the cover signal is and by subtracting it from the received signal, it can recover the covert message. In our work, however, the covert receiver of our network does the decoding independent of this knowledge. Third, the previous work evaluates the performance of their proposed covert communication but gives no information on the impact of their added noise on the normal communication of the system. It is of utmost importance that the added covert noise signals do not interfere with the normal communication of the system, since any unexpected increase in communication error rate between normal users would indicate that a suspicious activity (i.e. in this case, a covert communication) is taking place. Finally yet most importantly, previous work assumes the channel between users to be AWGN for simplification, however, we consider a more realistic scenario wherein communication is also subject to fading.

III. Background On Autoencoder Wireless Systems

An autoencoder-based wireless communication system is an end-to-end learning paradigm that abstracts out the coding and modulation components of a traditional modular communication system by replacing the transmitter and receiver with DNNs. You can see a block diagram of such a system in figure ?. The encoder (transmitter) first uses a mapping to transform k bits of data into a message s where $s \in \{1, ..., M\}$ and $M = 2^k$. Then it takes this transformed message as an input and generates a signal $x = E(s) \in \mathbb{R}^{2n}$, which is a real valued vector. This $2 \times n$ dimensional real valued vector can be treated as an n dimensional complex vector where n is the number of channel uses the signal needs to be transmitted over. Then, the channel's noise effect z, which is usually considered to

be AWGN, is added to the signal vector. Thus, the received signal at the receiver has noise of the channel added and can be expressed as $y = x + z$. In case there are many objects in the environment, Rayleigh fading is a more reasonable model for the channel. In this channel model, received signal is given by $y = h \cdot x + z$, where $h \sim \mathcal{CN}(0, 1)$ is the fading channel gain. The decoder (receiver) applies the transformation $D : \mathbb{R}^{2n} \rightarrow M$ to outputs the reconstructed version of the message s, which is denoted as $\hat{s} = D(y)$.

IV. Our Covert Communication System Model

In this section, we begin with an overview of our covert communication model and will further discuss the details of our scheme in the subsequent sections.

A. Overview

The main idea of our work is to establish a covert channel on top of a normal communication between a sender (UserTX) and a receiver (UserRX) who are using an autoencoder-based wireless system to communicate. We consider both an AWGN and a Rayleigh fading channels between our communication parties. The objective of the covert sender (Alice) is to secretly communicate with the covert receiver (Bob) by embedding her messages in form of perturbations that have similar statistical properties as of the channel's noise. This is achieved by help of an observer (Willie), which acts as an observer of the channel, trying to rigorously classify transmitted signals as covert and non-covert after they pass through the channel.

B. System Architecture

As mentioned above, there are three main actors in our covert system which we call them by their placeholder names: Alice, Bob, and Willie from now on. All three are collaborating in our covert scheme and are represented by DNNs. Alice is using a generative model that embeds a confidential message m into a covert noise vector $\hat{z}$. This covert signal is then transmitted over the channel after being added to a normal signal x. Similar to every covert communication scheme, there is an observer or warden in the system that will be alerted when seeing any deviation in the statistical properties of the channel. To this end, we are implicitly incorporating a statistical undetectability constraint on the produced covert signals by having a discriminator network employed by Willie. The presence of this discriminator network helps us to ensure that these added covert noise signals are distributed as the real channel noise, making them undetectable. Mathematically, in the AWGN channel model, it means $\hat{z} \sim \mathcal{CN}(0, \sigma^2_{chl})$ where σ^2_{chl} will be the variance of the channel's noise.
For a given binary secret message m, Alice first one-hot encodes the message and then uses its generator model to produce a covert noise signal $\hat{z}$. This covert signal is then added to a vector of a normal signal, which is carrying messages between UserTX and UserRX. Therefore, the covert signals before being transmitted over the channel can be denoted as:

$$\hat{x} = x + \hat{z} \tag{1}$$

The signal is then transmitted over the channel. We mentioned that we assume the channel between sender and receiver of the system to be AWGN or Rayleigh. Therefore, there will be two different channel outputs for these two different channel models that can be represented as a mapping function $C(\cdot)$:

1) AWGN Channel Output: For the AWGN channel model, the signal received at the receiver carries within itself the channel noise effect $z \sim \mathcal{N}(0, \sigma^2_{chl})$. Thus, the channel function $C(\cdot)$ and final covert signal $\hat{y}$ can be represented as:

$$C(\hat{x}) \Rightarrow \hat{y} = \hat{x} + z \tag{2}$$

2) Rayleigh Fading Channel Output: For the Rayleigh fading channel model, we consider a flat block fading channel where each codeword is assumed to be faded independently. Let h be the fading coefficient when transmitting the codeword $\hat{x}$, then the channel function $C(.)$ and and the final covert signal $\hat{y}$ is given by:

$$C(\hat{x}) \Rightarrow \hat{y} = h \cdot \hat{x} + z \tag{3}$$

On the receiver side, Bob receives a transmitted signal $\hat{y}$ after the channel noise is added into it. He uses its decoder network to reconstruct the covert message $\hat{m}$. Meanwhile, the UserRX is using the same signal to extract the normal message $\hat{s}$, which is the reconstructed message of s sent from UserTX. The statistical properties of the signals on the channel are captured by Willie who observes the channel continuously. His objective is to classify sequences of normal y and covert signals $\hat{y}$ passed through the channel and provide useful feedback to Alice. This feedback helps Alice to modify the produced covert signals such that they are indistinguishable from the normal signals of the communication. In other words, it ensures that both normal and covert signals have the same statistical properties.

C. General Formulation

The very first objective of our covert model is to have a working covert channel. To this end, Bob has to have a plausible accuracy in decoding covert messages that Alice sends through the covert signals $\hat{y}$. As mentioned in previous section, Alice employs a generative model instead of an encoder model suggested by [24]. Using an encoder model to produce covert signal perturbations will map each covert message m to a single covert noise vector $\hat{z}$. Inevitably, these deterministic covert perturbations can be eliminated with ease by a careful observer or defender using one of the techniques proposed in the work of Bahramali et al. [25], who already studied a similar problem statement. Thus, we use an stochastic generative model for Alice so that each covert message maps to a set of different covert noise signals. Let $A(\cdot)$ be the underlying function of Alice's generative model that takes a random trigger $t \sim \mathcal{N}(0, 1)$ and a covert message m) and produce a covert signal $\hat{z}$) (we denote the

corresponding covert signal as $\hat{z}_{m,t} = A(m,t)$). Let also $B(\cdot)$ be the underlying function of the decoder network that Bob makes use of to reconstruct the covert message $\hat{m}$. Then the reliability of communication between Alice and Bob is achieved using the below loss function:

$$\begin{aligned} &\mathcal{L}_{Alice-Bob} = \mathbb{E}_m[CE(\hat{m}, m)] \\ &\Rightarrow \mathbb{E}_m[CE(B(C(A(m,t)+x)), m)] \end{aligned} \tag{4}$$

where $CE(\cdot)$ is the cross entropy between the reconstructed covert message $\hat{m}$ and the actual covert message m. This equation can be used to optimize both for Alice's and Bob's networks by freezing one or the other network's parameters iteratively. While (7) ensures the communication accuracy, we also need to mind that the generated perturbations should have no detrimental impact on the normal communication between UserTX and UserRX. This constraint is applied by optimizing for the below loss function during training Alice's generator network:

$$\begin{aligned} &\mathcal{L}_{Alice|UserRX} = \mathbb{E}_m[CE(\hat{s}, s)] \\ &\Rightarrow \mathbb{E}_m[CE(D(C(A(m,t)+E(s))), s)] \end{aligned} \tag{5}$$

where $D(\cdot)$ is UserRX's decoder network function, and $E(\cdot)$ is the underlying function of the UserTX's encoder network. Note that both UserTX's encoder and UserRX's decoder networks are frozen during this training and only Alice's parameters are updated.

In our model, the observer entity or Willie, act as the discriminator in GAN models [13]. The so-called real and fake data in GANs' discriminator training here is mapped to non-covert and covert signals and we define the loss function as:

$$\begin{aligned} &\mathcal{L}_{Willie} = \mathbb{E}_m[BCE(\hat{y}, y)] \\ &\Rightarrow \mathbb{E}_m[BCE(C(A(m,t)+x), C(x))] \end{aligned} \tag{6}$$

where $BCE(\cdot)$ is the binary cross entropy between the covert signal $\hat{y}$ and the normal signal y. This adversarial training against Alice network ensures that Willie will be adequately trained to tell covert and non-covert signals apart. On the other hand, we mentioned that the covert signals that Alice produces should not deviate from the statistical properties of the normal signals on the channel, otherwise the observer of the channel can detect and mitigate the covert communication. To achieve this undetectability property, we pose a new constraint on Alice's optimization function for maximizing Willie's uncertainty about the covert predictions. Having a regularizer as such helps Alice and Bob to form their covert communication in way that is indistinguishable from the actual channel's noise, yet understandable by both. Altogether, Alice's loss function can be expressed as:

$$\mathcal{L}_{Alice} = \mathcal{L}_{Alice-Bob} + \mathcal{L}_{Alice|UserRX} - \mathcal{L}_{Willie} \tag{7}$$

D. Neural Network Architecture

Before discussing the architecture of our neural network models, we need to state that the communication between UserTX and UserRX is taking place using an autoencoder communication model, which is utilized to encode and decode the normal messages between these two entities. Since the focus of this work is not to introduce an autoencoder wireless network, we briefly explains how this model works. Also, an overview of the network's architecture of implemented autoencoder model can be found in table (?). Detailed explanation of such a network or the training procedure can be found in the original paper [5]. The other three entities of our scheme, which are our covert actors, also employ neural networks to function. A detailed network architecture of each is likewise represented in table (?).

Similar to the original paper, our autoencoder model accepts a binary message s of size k bits. Afterwards, the encoder part first one-hot encodes the message and then maps it to a vector of signals of size $2 \times n$, where n is the number of channel uses. This transmitted signal is then given to a mapping function that applies the channel effects into it. On the receiver side, there will be two different structures given what the channel model is. In case of an AWGN channel model, the signal is simply passed through the decoder network and the intended message gets extracted. For the Rayleigh fading channel model, however, there will be a parameter estimation model that takes in the signal before getting passed through the decoder network in order to estimate the channel's fading coefficients from it. Followed by that, the extracted signal along with the estimated parameters are passed to a transformation function that is supposed to revert the channel effects. In our case, we are using a simple division transformation function that divides the received signal by the estimated channel fading coefficients. More complex transformation functions can be used and are described in [5]. Eventually, the transformed signal is fed to the decoder network and the original normal message is reconstructed. Similar to the encoder network, Alice takes a covert message m and transforms it to an one-hot encoding representation. Next, given a random trigger t, it uses its generator model to produce a covert noise signal $\hat{z}$ and adds it to the normal signal x that is being transmitted. Bob receives this covert signal after passing through the channel and feeds it through its decoder network regardless of what the channel is and extracts the secret message. Willie also receives the same covert signal $\hat{y}$ and the non-covert signal y and outputs a confidence probability P on how probable it is that the signal is normal.

V. Experiments and Evaluation

VI. Conclusion

References

[1] B. W. Lampson, "A note on the confinement problem," *Communications of the ACM*, vol. 16, no. 10, pp. 613–615, 1973.

[2] B. A. Bash, D. Goeckel, and D. Towsley, "Square root law for communication with low probability of detection on awgn channels," in *2012 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2012, pp. 448–452.

[3] T. V. Sobers, B. A. Bash, S. Guha, D. Towsley, and D. Goeckel, "Covert communication in the presence of an uninformed jammer," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 6193–6206, 2017.

[4] R. Soltani, D. Goeckel, D. Towsley, B. A. Bash, and S. Guha, "Covert wireless communication with artificial noise generation," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7252–7267, 2018.

[5] T. O'shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563–575, 2017.

[6] T. Wang, C.-K. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep learning for wireless physical layer: Opportunities and challenges," *China Communications*, vol. 14, no. 11, pp. 92–111, 2017.

[7] S. Tan and B. Li, "Stacked convolutional auto-encoders for steganalysis of digital images," in *Signal and information processing association annual summit and conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–4.

[8] Y. Qian, J. Dong, W. Wang, and T. Tan, "Deep learning for steganalysis via convolutional neural networks," in *Media Watermarking, Security, and Forensics 2015*, vol. 9409. SPIE, 2015, pp. 171–180.

[9] G. Xu, H.-Z. Wu, and Y.-Q. Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 708–712, 2016.

[10] S. Baluja, "Hiding images in plain sight: Deep steganography," *Advances in neural information processing systems*, vol. 30, 2017.

[11] C. Zhang, C. Lin, P. Benz, K. Chen, W. Zhang, and I. S. Kweon, "A brief survey on deep learning based data hiding, steganography and watermarking," *arXiv preprint arXiv:2103.01607*, 2021.

[12] C. Zhang, P. Benz, A. Karjauv, G. Sun, and I. S. Kweon, "Udh: Universal deep hiding for steganography, watermarking, and light field messaging," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 223–10 234, 2020.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[14] D. Volkhonskiy, I. Nazarov, and E. Burnaev, "Steganographic generative adversarial networks," in *Twelfth International Conference on Machine Vision (ICMV 2019)*, vol. 11433. International Society for Optics and Photonics, 2020, p. 114333M.

[15] J. Hayes and G. Danezis, "Generating steganographic images via adversarial training," *Advances in neural information processing systems*, vol. 30, 2017.

[16] A. Sheikholeslami, M. Ghaderi, D. Towsley, B. A. Bash, S. Guha, and D. Goeckel, "Multi-hop routing in covert wireless networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 3656–3669, 2018.

[17] K. Li, M. Ghaderi, and D. Goeckel, "Fundamental limits of activity-based covert channels," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.

[18] A. Dutta, D. Saha, D. Grunwald, and D. Sicker, "Secret agent radio: Covert communication through dirty constellations," in *International Workshop on Information Hiding*. Springer, 2012, pp. 160–175.

[19] P. Cao, W. Liu, G. Liu, X. Ji, J. Zhai, and Y. Dai, "A wireless covert channel based on constellation shaping modulation," *Security and Communication Networks*, vol. 2018, 2018.

[20] N. Hou and Y. Zheng, "Cloaklora: A covert channel over lora phy," in *2020 IEEE 28th International Conference on Network Protocols (ICNP)*. IEEE, 2020, pp. 1–11.

[21] L. Bonati, S. D'Oro, F. Restuccia, S. Basagni, and T. Melodia, "Stealte: Private 5g cellular connectivity as a service with full-stack wireless steganography," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[22] K. Sankhe, F. Restuccia, S. D'Oro, T. Jian, Z. Wang, A. Al-Shawabka, J. Dy, T. Melodia, S. Ioannidis, and K. Chowdhury, "Impairment shift keying: Covert signaling by deep learning of controlled radio imperfections," in *MILCOM 2019-2019 IEEE Military Communications Conference (MILCOM)*. IEEE, 2019, pp. 598–603.

[23] X. Liao, J. Si, J. Shi, Z. Li, and H. Ding, "Generative adversarial network assisted power allocation for cooperative cognitive covert communication system," *IEEE Communications Letters*, vol. 24, no. 7, pp. 1463–1467, 2020.

[24] H. Mohammed, X. Wei, and D. Saha, "Adversarial learning for hiding wireless signals," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 01–06.

[25] A. Bahramali, M. Nasr, A. Houmansadr, D. Goeckel, and D. Towsley, "Robust adversarial attacks against dnn-based wireless communication systems," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 126–140.