ADVERSARIAL NETWORKS FOR SECURE WIRELESS COMMUNICATIONS

Thomas Marchioro[†]*, Nicola Laurenti*[†]*, Deniz Gündüz*[*]

[†] University of Padova, Italy
[*] Imperial College London, UK

ABSTRACT

We propose a data-driven secure wireless communication scheme, in which the goal is to transmit a signal to a legitimate receiver with minimal distortion, while keeping some information about the signal private from an eavesdropping adversary. When the data distribution is known, the optimal trade-off between the reconstruction quality at the legitimate receiver and the leakage to the adversary can be characterised in the information theoretic asymptotic limit. In this paper, we assume that we do not know the data distribution, but instead have access to a dataset, and we are interested in the finite blocklength regime rather than the asymptotic limits. We propose a data-driven adversarially trained deep joint source-channel coding architecture, and demonstrate through experiments with CIFAR-10 dataset that it is possible to transmit to the legitimate receiver with minimal end-to-end distortion while concealing information on the image class from the adversary.

***Index Terms*—** security, wiretap channel, convolutional neural networks, generative adversarial networks

1. INTRODUCTION

Physical layer secrecy achieves information confidentiality by exploiting an advantage for the legitimate channel with respect to some eavesdropper. This approach to security is particularly interesting, since it does not rely on cryptographic mechanisms, but only on physical characteristics of the channel, and provides security guarantees independent of the computational power of the eavesdropper. The limits of physical layer secrecy are characterized by the *secrecy capacity*, or the more general trade-off between the communication rate and the private message's equivocation (secrecy) rate [1, 2, 4–6].

Here, we consider the more general setting studied in [3] (see Fig. 1 for an illustration), where we consider the lossy delivery of an information source (U^k) to the legitimate receiver, while limiting the information leakage to an adversary. We will further generalize the model in [3], and assume that the transmitter only wants to keep a certain sensitive part of the information source (S) secure from the eavesdropper. By considering independent and identically distributed (i.i.d.) discrete memoryless source and channel distributions, the fundamental trade-off between the best achievable distortion at the legitimate receiver, and the leakage to the eavesdropper, measured by the equivocation rate, is characterized in [3] in the asymptotic information theoretic regime. It is shown that the optimal performance is achieved by a separation based scheme, where lossy compression of the source is followed by an optimal wiretap channel code to transmit the compressed bits. This formulation is attractive as it provides theoretical and quantifiable security guarantees; however, its application to practical systems is limited due to the idealistic and perfectly known source and channel distributions, and the result does not hold in practical finite blocklength regimes.

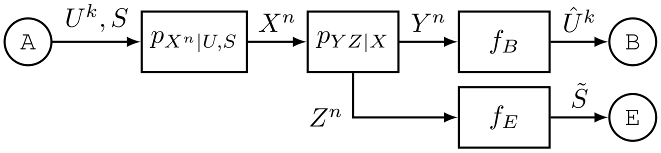

Fig. 1: The wiretap channel model.

In this work, our goal is to study the trade-off between the distortion achieved at the legitimate receiver and the leakage to an eavesdropper in a practical non-asymptotic regime. Moreover, we will not assume the knowledge of the distribution of the underlying source and the sensitive part, but instead follow a data-driven approach. Examples for the application of this framework are abundant: an activity/ health sensor can transmit user's vital measurements to an access point. Its goal would be to provide as accurate description of the underlying signals as possible, while keeping some private aspect of the data hidden from potential eavesdroppers (e.g., the identity of the user). Similarly, a surveillance drone may want to transmit back images without revealing the locations of critical infrastructures from eavesdroppers.

Data-driven approaches to wireless communications is receiving increasing attention [13], including autoencoder-based end-to-end design for channel coding [8], as well as for joint source-channel coding (JSCC) [9]. Yet, to the best our knowledge, the only prior work studying a similar data-

This work has been funded by the European Research Council (ERC) through project BEACON (No. 677854).

driven approach to wiretap channels is [10], where the authors propose clustering of constellation points in an autoencoder based communication scheme in order to achieve a trade-off between the reliability at the legitimate receiver and the eavesdropper. While [10] considers only the channel coding problem, we are interested in the end-to-end performance, and consider the sensitive information to be different from the underlying source (yet correlated with it).

We consider a fully convolutional autoencoder architecture to transmit U^k over the noisy channel. The autoencoder pair, in addition to optimizing the end-to-end reconstruction quality, also aims at preventing the leakage to the eavesdropper, modeled through an adversarial neural network. Due to the difficulty of estimating the mutual information, we use a variational approximation [11], and train the autoencoder with the combined objective of maximizing the reconstruction quality at the legitimate receiver while minimizing the adversarial loss. We apply our approach to secure image transmission, where the legitimate transmitter aims to share images with the legitimate receiver over a wireless channel, while the eavesdropper tries to classify them. Our results show that the adversarially trained communication scheme allows to achieve reasonable quality at the legitimate receiver, while confusing the eavesdropper.

2. PROBLEM FORMULATION

Consider the communication scenario illustrated in Fig. 1: (A)lice wants to reveal some information U^k to (B)ob over n uses of a noisy communication channel. (E)ve eavesdrops the channel, and receives a noisy version of the A's signal through another channel. The goal of A is to reveal U^k to B with minimum distortion under a given distortion measure $d(\cdot,\cdot)$, while preventing some sensitive information S, correlated with U^k with $p_{U^k,S}$, leaking to E. Information leakage to E is measured by the mutual information $I(S;Z^n)$. Source U^k is encoded by A into a codeword X^n according to a mechanism $p_{X^n|U^k}$. Note that we assume that A does not directly observe S. The codeword X^n is transmitted along the channel, which is characterized by the joint conditional distribution

$$p_{Y^nZ^n|X^n}(y^n,z^n|x^n)=\prod_{i=1}^{n}p_{YZ|X}(y_i,z_i|x_i). \tag{1}$$

Channel outputs Y^n and Z^n are received by B and E, respectively. In order to obtain an estimate $\hat{U}^k$ of U^k, B applies a function f_B.

We can formulate the optimization problem as

$$\begin{aligned} &\min_{p_{X^n|U^k},f_B} \mathbb{E}[d(U^k,\hat{U}^k)] \qquad (2)\\ &\text{s.t. } I(S;Z^n)\le c,\\ &S\rightarrow U^k\rightarrow X^n\rightarrow Y^n\rightarrow \hat{U}^k\\ &S\rightarrow U^k\rightarrow X^n\rightarrow Z^n\rightarrow \tilde{S}\end{aligned}$$

where $c>0$ is the secrecy constraint. Estimating the mutual information $I(S;Z^n)$ is known to be challenging; hence, we will use a variational lower bound commonly employed [11, 12], and write the optimization problem in (2) in the unconstrained form as follows:

$$\min_{p_{X^n|U^k},f_B}\max_{q_{S|Z^n}}\{\mathbb{E}[d(U^k,\hat{U}^k)]+\alpha\mathbb{E}[\log q_{S|Z^n}]\}, \tag{3}$$

where $\alpha\ge 0$ is the parameter regulating the privacy-distortion trade-off, and and $q_{S|Z^n}$ can be considered as the estimated distribution of S at the adversary based on its observation Z^n.

The problem in (3) is a minimax game between the A and B pair, and E. While E wants to maximize the leakage, measured by the negative log-loss term, by choosing the posterior distribution $q_{S|Z^n}$, A and B jointly decide on the encoding and decoding functions, $p_{X^n|U^k}$ and f_B respectively, to minimize a weighted sum of the distortion and the leakage.

To solve the optimization problem in 3, we will realize all three components to be optimized as neural networks. The autoencoder pair at A and B will be parametrized by θ_A and θ_B, respectively, while E's network will be parametrized by θ_E. Then the loss function to be minimized is

$$\mathcal{L}_M(\theta_A,\theta_B,\theta_E)=\mathbb{E}[d(U^k,\hat{U}^k)]+\alpha\mathbb{E}[\log q_{\theta_E}(S|Z^n)]. \tag{4}$$

Following the standard approach in GANs, these network parameters will be optimized by iteratively training them: each joint training step of the autoencoder pair (θ_A,θ_B) will be followed by a training step of θ_E by the eavesdropper.

3. IMPLEMENTATION

While the above formulation is generic, and can be applied to any type of information source and wiretap channel, and any distortion measure at the legitimate receiver, in the rest of the paper we will focus on secure transmission of images over an additive white Gaussian noise (AWGN) wiretap channel. We impose an average power constraint on the length n transmitted codewords. More specifically, we fix the power constraint to 1, i.e., $\frac{1}{n}\sum_{i=1}x_i^2\le 1$, but allow different noise variances, and hence SNR values, at the legitimate receiver and the eavesdropper, which will be denoted by Λ_B and Λ_E, respectively. We use the peak SNR (PSNR) as the distortion measure at the legitimate receiver, defined as PSNR $\triangleq \dfrac{1}{k}\sum_{i=1}^{k}10\log_{10}\left(\dfrac{255^2}{(u_i-\hat{u}_i)^2}\right)$.

For the autoencoder (θ_A,θ_B) that represent the legitimate JSCC encoder and decoder pair, we employed the network structure described in [9], consisting of five convolutional neural network layers. We will fix the bandwidth ratio between the available channel bandwidth n and the input image size k as $n/k=1/6$.

The adversary's network θ_E, as illustrated in Fig. 2, consists of a predictor that takes as input the vector Z^n of n real

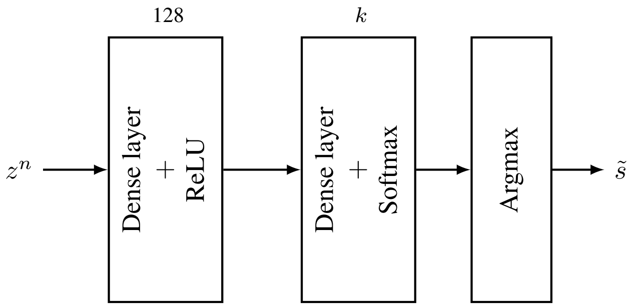

Fig. 2: The architecture of the network employed by the adversary. The number written above each layer is the number of components in the output.

valued channel outputs, and applies first a dense layer with rectified linear unit (ReLU) activation, followed by another dense layer with softmax activation. Finally, although that is not part of the network itself, we add a final stage where a guess is taken as the argmax of the distribution, so that we can assess the effectiveness of the adversary network by measuring its accuracy, i.e., that fraction of correct guesses.

Ideally, the termination condition of the iterative training procedure should be attaining a predetermined convergence margin, but we decided to fix the number of epochs in advance, and average the results across several trials.

4. LIKELIHOOD EQUALIZATION

Minimizing $\mathbb{E}[\log q_{S|Z^n}]$ means minimizing the value of the adversary's estimated likelihood corresponding to the correct value of S. When the iterative adversarial training approach is taken, the function $\mathbb{E}[\log q_{S|Z^n}]$ can be easily minimized by the legitimate autoencoder pair by performing a permutation of the encoding scheme. Suppose, for instance, we switch the codewords x_1^n and x_2^n corresponding to two different realizations of the input sequence: the adversary likelihood estimation can be easily brought down without any impact on the decoding distortion. Nonetheless, the permutation can be easily recovered by the adversary in its own training phase, again increasing the leakage. This leads to the saw-tooth behaviour which is shown in Fig. 3.

We therefore consider a different approach, which we call *likelihood equalization*. The main idea behind likelihood equalization is to get the likelihood estimation of the adversary as close as possible to a uniform distribution, rather than minimizing the likelihood related to the correct prediction, in order to make their prediction unreliable. We hence employed a new objective function for security, which consists of the cross-entropy between a uniform distribution $\bar{p}$ and the likelihood estimation $q_{S|Z^n}$, i.e., $H(\bar{p}, q_{S|Z^n})$. As can be seen in Fig. 3, the latter approach yields a more stable behaviour. The former likelihood minimization approach will be referred to as the *naive approach*.

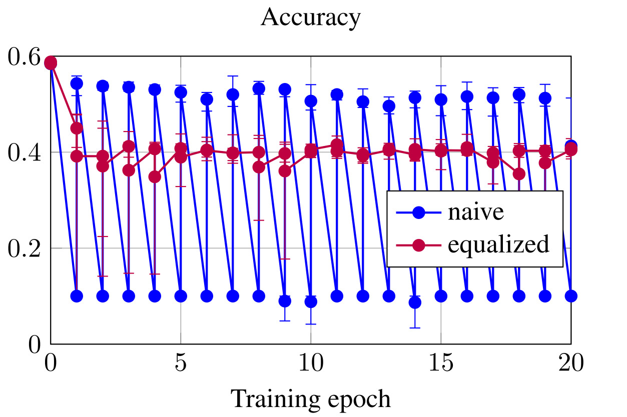

Fig. 3: Stability comparison between the naive approach and the likelihood equalization approach, with $\alpha = 1$, $\Lambda_B = 10$ dB and $\Lambda_E = 5$ dB.

Table 1: Parameters used for training

Parameter	Symbol	Value
Iterations in Phase 1	N_1	30000
Iterations in Phase 2	N_2	30000
Number of epochs	N_{epoch}	40
Main network iterations	N_M	500
Adversary network iterations	N_E	2000
Main receiver SNR	Λ_B	10 dB
Adversary SNR	Λ_E	5 dB
Learning rate	η	10^{-4}
Size of training batch	m_{batch}	32
Size of test set	m_{test}	10000

5. RESULTS

We applied our solution to the CIFAR-10 dataset, which consists of $N_{\text{i}} = 60000$ (50000 for training and 10000 for test) coloured images of size 32×32 pixels, divided into $k = 10$ classes. We first fixed the SNR of the channels and trained the adversarial network with different values of the trade-off parameter α. We measured the level of privacy using the accuracy of the adversary predictions, i.e., the fraction of images whose class was correctly identified, while measured the quality of the reconstructed images via PSNR.

The results in Fig. 4 show that the approach can provide either good quality in the transmission, when α is small, or high privacy, when α is large. Observe that $1/k = 0.1$ represents an ideal lower bound to the adversary accuracy as it corresponds to uniform guess, independent of the actual transmitted image. The PSNR and accuracy curves show a similar behaviour. Hence, we fixed $\alpha = 1$, which is a value that provides low accuracy without compromising the PSNR, and saved the weights of the network trained with $\Lambda_B = 10$ dB and $\Lambda_E = 5$ dB.

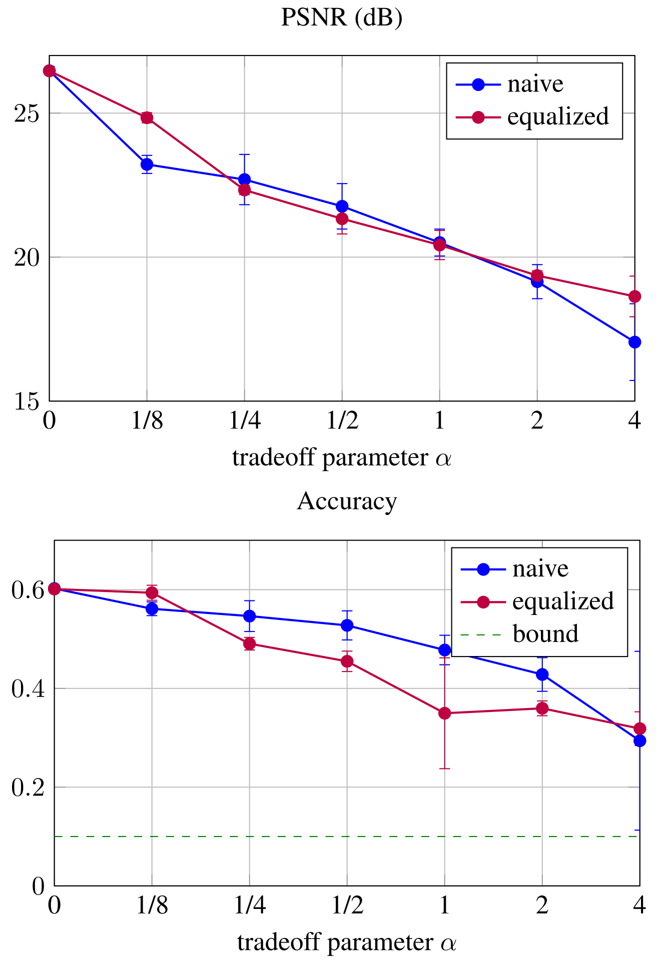

Fig. 4: Steady-state PSNR and adversarial accuracy vs the tradeoff parameter α, with $\Lambda_B = 10$ dB and $\Lambda_E = 5$ dB.

We then tested the adversarial network by varying the actual SNR of the adversary channel Λ_E, with respect to a fixed training value $\hat{\Lambda}_E$. Fig. 5 shows that the accuracy of the adversary predictor drops significantly when the SNR is brought below the training value, and is even moderately reduced when the SNR is higher. The parameters employed in the training phase are reported on Table 1.

6. CONCLUSIONS

We have developed a neural network-based framework to learn coding schemes to achieve security over a noisy wiretap channel. We have adopted an adversarial formulation that leads to the solution of a minimax game where a legitimate autoencoder network and an adversary network compete. We have tested our approach for secure transmission of images from the CIFAR-10 dataset.

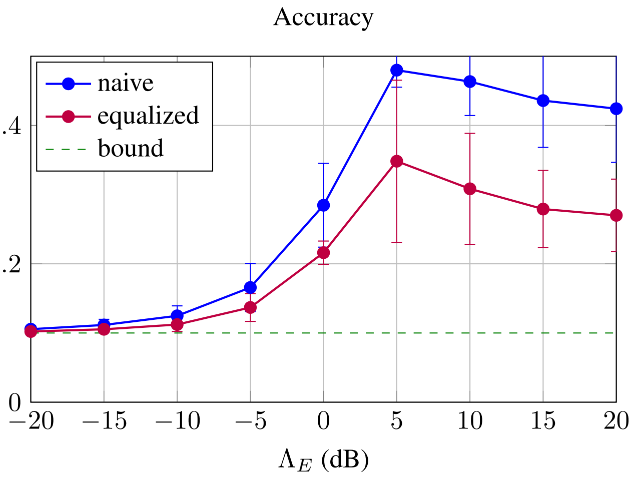

Fig. 5: Accuracy of the adversary with varying Λ_E performed after training with $\alpha = 1$, $\Lambda_B = 10$ dB and $\hat{\Lambda}_E = 5$ dB.

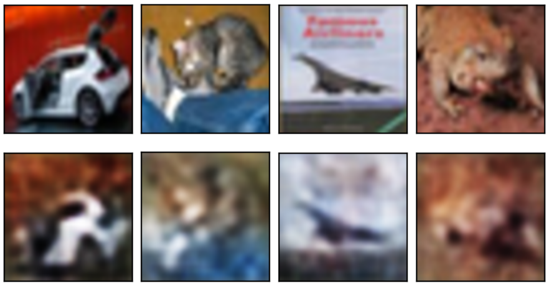

Fig. 6: Examples of images transmitted by $\mathtt{A}$ (above) and reconstructed by $\mathtt{B}$ (below) using the likelihood equalization approach, with $\alpha = 1$, $\Lambda_B = 10$ dB and $\Lambda_E = 5$ dB.

The network is able to guarantee a privacy-distortion trade-off, which becomes more advantageous when the disturbance in the adversary channel is increased. We have first adopted a naive approach, which aims at maximizing the adversary's cross-entropy, but also considered a more stable approach which aims to take the adversary softmax output close to a uniform distribution.

Future work will include random encoding functions as opposed to the deterministic approach used here. We will also consider other types of objective functions at the legitimate receiver which may allow further secrecy if this is not aligned with eavesdropper's objective.

7. REFERENCES

[1] A. D. Wyner, *The wire-tap channel*, Bell Syst. Tech. J., vol. 54, pp. 1355-1387, Oct. 1975

[2] S. Leung-Yan-Cheong, M. Hellman, *The Gaussian wire-tap channel*, IEEE Transactions on Information Theory, vol. 24, no. 4, pp. 451-456, 1978.

[3] H. Yamamoto, *Rate-distortion theory for the Shannon cipher system*, in IEEE Transactions on Information Theory, vol. 43, no. 3, pp. 827-835, May 1997.

[4] Y. Liang, H. V. Poor, S. Shamai (Shitz), *Secure communication over fading channels*, IEEE Transactions on Information Theory, vol. 54, no. 6, pp. 2470-2492, June 2008.

[5] M. R. Bloch, J. Barros, *Physical-Layer Security: From Information Theory to Security Engineering*, First edition, Cambridge University Press, 2011

[6] M. R. Bloch, J. N. Laneman, *Strong Secrecy From Channel Resolvability*, IEEE Transactions on Information Theory, 59(12), pp.8077-8098, 2013

[7] I. Goodfellow et al., *Generative adversarial nets*, In Proc. Int'l Conf. on Neural Information Processing Systems, 2014.

[8] T. O'Shea, J. Hoydis, *An Introduction to Deep Learning for the Physical Layer*, arXiv:1702.00832v2, July 2017

[9] E. Bourtsoulatze, D. B. Kurka, D. Gunduz, *Deep Joint Source-Channel Coding for Wireless Image Transmission*, IEEE Transactions on Cognitive Communications and Networking, vol. 5, no. 3, pp. 567–579, Sep. 2019

[10] R. Fritschek, R. F. Schaefer and G. Wunder, *Deep Learning for the Gaussian Wiretap Channel*, IEEE International Conference on Communications (ICC), Shanghai, China, 2019, pp. 1-6.

[11] D. Barber and F. Agakov, *The IM algorithm: A variational approach to information maximization*, In Proc. Int'l Conf. on Neural Information Processing Systems, 2003, pp. 201-208.

[12] A. Tripathy, Y. Wang, and P. Ishwar *Privacy-Preserving Adversarial Networks* arXiv:1712.07008v3, June 2019

[13] D. Gunduz, P. de Kerret, N. D. Sidiropoulos, D. Gesbert, C. R. Murthy and M. van der Schaar, *Machine Learning in the Air,* IEEE Journal on Selected Areas in Communications, vol. 37, no. 10, pp. 2184-2199, Oct. 2019.

[14] K. Besser et al., *Flexible Design of Finite Blocklength Wiretap Codes by Autoencoders*, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2512-2516, 2019.