Unveiling Covert Semantics: Joint Source-Channel Coding Under a Covertness Constraint

Abdelaziz Bounhar*, Mireille Sarkiss§, Michèle Wigger*
*LTCI, Télécom Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France
{abdelaziz.bounhar, michele.wigger}@telecom-paris.fr
§SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France
{mireille.sarkiss}@telecom-sudparis.eu

***Abstract*—The fundamental limit of Semantic Communications (joint source-channel coding) is established when the transmission needs to be kept covert from an external warden. We derive information-theoretic achievability and matching converse results and we show that source and channel coding separation holds for this setup. Furthermore, we show through an experimental setup that one can train a deep neural network to achieve covert semantic communication for the classification task. Our numerical experiments confirm our theoretical findings, which indicate that for reliable joint source-channel coding the number of transmitted source symbols can only scale as the square-root of the number of channel uses.**

***Index Terms*—Semantic Communication, Joint Source-Channel Coding, Physical Layer Security, Covert Communication**

I. Introduction

Semantic Communication refers to the emerging communication paradigm where the transmitter sends only the semantics of a source but not the entire source itself [1]. Thanks to the significant gains in bandwidth efficiency, this communication paradigm has attracted great attention in the community and has the potential to be part of next generation communication networks. For traditional communication, Shannon's separation theorem [2] implies that without loss in optimality, one can establish semantic communication by simple concatenation of optimal source (compression) and channel codes. However, while above optimality only holds in the asymptotic regime of infinite blocklengths, for finite blocklengths a joint design of the source and channel codes can yield improved performances [3]. Indeed, various practical joint source-channel codes have been proposed in the literature, where recent advances particularly focus on implementations using Deep Neural Networks (DNNs) [4]–[8].

Moreover, privacy and security are becoming crucial for communication in many applications, see [9]–[11]. In this spirit, [5], [10] proposed to leverage a DNN-based architecture to simultaneously minimize the distortion of the reconstruction at the legitimate receiver, while also restricting information leakage to potential eavesdroppers. In this work, we propose a similar DNN-based implementation, which however respects the more stringent security constraint that potential attackers should not only be unable to learn about the transmitted source, but even stay agnostic of the mere fact that communication is going on. We are thus imposing the constraint that semantic communication be *Covert*, i.e., undetectable to external eavesdroppers. Our implementation shows that covert semantic communication for the classification task can be achieved through a DNN architecture, and that the number of extracted features should be in the order of the square-root of the number of channel uses used for communication. This reminds the well-known square-root law of covert data communication, which was established in [12]–[14]. Similar observations were noted in studies examining covert detection [15] and others involving both covert and non-covert communication [16], [17].

In this work, we endorse our numerical experiments with a rigorous information-theoretic analysis that establishes the fundamental limits of joint source-channel coding (JSCC) under a covertness constraint, i.e. the information theoretic limits of covert semantic communication. Our results provide necessary and sufficient conditions under which covert semantic communication is possible. These conditions in particular imply that separate source-channel coding is optimal for covert semantic communication, and that the number of source symbols should scale at most as the square-root of the number of channel uses, similarly as for covert data communication. Often this square-root scaling is not a problem in semantic communication as the extracted number of features typically occupies a small space. The main contribution of our information-theoretic results is the converse proof where we show that a separate source-channel coding architecture is optimal in the asymptotic regime of infinite blocklengths.

In brief, this paper makes the following contributions:

- We introduce and study the problem of joint source-channel coding under a covertness constraint.
- We show that source-channel separation is optimal in this setup by deriving matching information-theoretic achievability and converse proofs. This establishes necessary and sufficient conditions for the distortions that are achievable in covert joint source-channel coding.
- Our experimental setup showcases that a Deep Neural Network can achieve covertness when transmitting semantic information for a classification task. The experimental results confirm our theoretical findings and show that the classification task can only be achieved if the number of extracted features is in the order of the square-root of the number of channel uses.

II. Information-Theoretic Approach

A. Notation

We follow standard notations in [16], [18], [19]. In particular, we denote a random variable by X and its realization by x. We write X^n and x^n for the tuples $(X_1, \ldots, X_n)$ and $(x_1, \ldots, x_n)$, respectively, for any positive integer $n > 0$. For a distribution P on $\mathcal{X}$, we note its product distribution on $\mathcal{X}^n$ by $P^{\otimes n}(x^n) := \prod_{i=1}^n P(x_i)$. For two distributions P and Q on $\mathcal{X}$, $\mathbb{D}(P\|Q) := \sum_{x\in\mathcal{X}} P(x)\log(\frac{P(x)}{Q(x)})$ denotes the Kullback-Leibler (KL) divergence between P and Q, whilst the chi-squared test is denoted $\chi^2(P\|Q) := \sum_{x\in\mathcal{X}} \frac{(P(x)-Q(x))^2}{Q(x)}$. Finally, the logarithm function is understood in base 2, so the results are in bits.

B. Problem statement

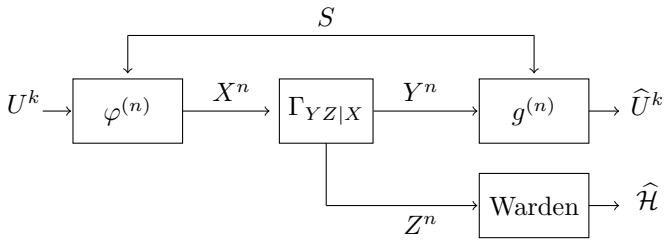

Fig. 1: Covert semantic communication system.

Consider the JSCC problem in Figure 1. The transmitter wishes to communicate a sequence $U^k \in \mathcal{U}^k$ that is drawn i.i.d. according to a given distribution P_U (for a given $k \geq 0$ and an arbitrary finite set $\mathcal{U}$) to a legitimate receiver in the presence of a warden, while tolerating a defined level of distortion $D \geq 0$ at the legitimate receiver. The transmitter and the legitimate receiver also share a secret-key S, which is uniformly distributed over a finite set $\mathcal{S}$ of sufficient large size.[1] Communication must remain covert, i.e., an external warden should not be able to detect the presence of communication.

Technically speaking we have two hypotheses: under $\mathcal{H} = 1$ the communication takes place as described above, while under $\mathcal{H} = 0$ the transmitter remains silent (sends the "off-symbol"). The transmitter and the legitimate receiver both know $\mathcal{H}$, which however has to remain undetectable to the warden. The receiver and the warden observe channel outputs produced by a Discrete and Memoryless Channel (DMC) with a known transition law $\Gamma_{YZ|X}$ and given finite input and output alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. That means, if the $X^n = x^n$ then the for any i the i-th output symbols Y_i and Z_i observed at the legitimate receiver and the warden are generated from the i-th input x_i according to the conditional laws $\Gamma_{Y|X}(\cdot|x_i)$ and $\Gamma_{Z|X}(\cdot|x_i)$.[2] For simplicity, we assume a binary input alphabet $\mathcal{X} = \{0, 1\}$ and consider that 0 is the "off-symbol". We can now describe the communication model and the constraints in full mathematical details.

Under $\mathcal{H} = 0$: the transmitter sends the all-zero sequence

$$X^n = 0^n. \tag{1}$$

Under $\mathcal{H} = 1$: the transmitter applies some encoding function $\varphi^{(n)}: \mathcal{U}^k \times \mathcal{S} \to \mathcal{X}^n$ to its sequence U^k and sends the resulting codeword

$$X^n = \varphi^{(n)}(U^k, S) \tag{2}$$

over the channel. For readability, we will also write $x^n(u^k, s)$ instead of $\varphi^{(n)}(u^k, s)$.

The legitimate receiver decodes the desired sequence U^k based on its observed output sequence Y^n and the secret key S. Thus, under $\mathcal{H} = 0$ it does nothing whereas under $\mathcal{H} = 1$ it uses a decoding function $g^{(n)}: \mathcal{Y}^n \times \mathcal{S} \to \widehat{\mathcal{U}}^k$ to produce the guess

$$\widehat{U}^k = g^{(n)}(Y^n, S), \tag{3}$$

over a given reconstruction alphabet $\widehat{\mathcal{U}}^k$, which can differ from $\mathcal{U}^k$. Allowing for a general reconstruction alphabet $\hat{\mathcal{U}}$ enables the consideration of more general reconstruction tasks, such as not reconstructing the entire source symbols but only a feature thereof. (In this case $\hat{\mathcal{U}}$ would be the feature space.)

Decoding performance under $\mathcal{H} = 1$ of a pair of encoding and decoding functions $(\varphi^{(n)}, g^{(n)})$ is measured by a bounded per-letter distortion measure $d(\cdot,\cdot)$.We require the average per-block distortion to be less or equal to a given positive threshold D,

$$\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} d(U_i, \widehat{U}_i)\right] \leq D, \tag{4}$$

where expectation is over the random source sequence U^k and the randomness in the channel. It is assumed that the distortion measure $d(\cdot,\cdot)$ is such that for any $u \in \mathcal{U}$, there exists a reconstruction symbol $\widehat{u} \in \widehat{\mathcal{U}}$ that has zero distortion, i.e. $d(u, \widehat{u}) = 0$.

Communication is subject to a covertness constraint at the warden, which observes the channel outputs Z^n. Under $\mathcal{H} = 1$, the warden's output distribution is thus

$$\widehat{Q}^n(z^n) \triangleq \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\sum_{u^k} P_U^{\otimes k}(u^k)\Gamma_{Z|X}^{\otimes n}(z^n|x^n(u^k, s)), \tag{5}$$

whereas under $\mathcal{H} = 0$ it is

$$\Gamma_{Z|X}^{\otimes n}(z^n|0^n). \tag{6}$$

Our covertness metric is the KL-divergence between these two output distributions $\mathbb{D}\left(\widehat{Q}^n \parallel \Gamma_{Z|X}^{\otimes n}(\cdot|0^n)\right)$. The choice of this measure is justified by the fact that any test satisfies [20] $\alpha + \beta \geq 1 - \mathbb{D}\left(\widehat{Q}^n \parallel \Gamma_{Z|X}^{\otimes n}(\cdot|0^n)\right)$, for α and β denoting

[1] It has been proved in [13, Section IV-c] that generally a secret-key size $\log|\mathcal{S}|$ that scales as $\sqrt{n}$ suffices to achieve covertness. When the channel the legitimate receiver is better than the channel to the warden, no secret-key is required at all. In a similar way, and using the convexity of the Kullback-Leibler divergence, one can show that the same also holds for the setup in this paper. Details are omitted due to lack of space.

[2] Notice the generality of this channel model that even allows the modeling of fast fading channels.

the probabilities of miss-detection and false alarm, respectively. Therefore, ensuring a negligible $\mathbb{D}\left(\widehat{Q}^n \parallel \Gamma_{Z|X}^{\otimes n}(\cdot|0^n)\right)$ is sufficient to achieve covertness.

Our problem is thus multi-objective in the sense that we not only wish to satisfy the distortion constraint (4), but also a vanishing detectability capability at the warden $\mathbb{D}\left(\widehat{Q}^n \parallel \Gamma_{Z|X}^{\otimes n}(\cdot|0^n)\right)$. These constraints are reflected in the following definition

Definition 1. *Let $k = f(n)$ for a given function $f(\cdot)$ on appropriate domains and $\{\delta_n\}_{n\geq 1}$ be a sequence tending to 0 as the blocklength $n \to \infty$. A source-channel pair $(P_U, \Gamma_{YZ|X})$ is (D, δ_n)-admissible under a covertness constraint if there exists a sequence of encoding and reconstruction functions $\{\varphi^{(n)}, g^{(n)}\}_n$ satisfying the two conditions*

$$\limsup_{k\to\infty} \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} d(U_i, \widehat{U}_i)\right] \leq D, \tag{7}$$

$$\mathbb{D}\left(\widehat{Q}^n \parallel \Gamma_{Z|X}^{\otimes n}(\cdot|0^n)\right) \leq \delta_n, \qquad \forall n. \tag{8}$$

C. Information-Theoretic Results

In this section, we characterize the fundamental limits of our covert JSCC setup. Our results show that if one wishes to attain a non-trivial distortion D and at the same time satisfy a covertness constraint δ_n (see (8)) then the number of source symbols k can scale at most proportional to $\sqrt{n\delta_n}$ (see Result 2). We say that a distortion is trivial if it can be achieved without communication by having the receiver produce a constant reconstruction symbol, i.e., if

$$D \geq D_{\text{trivial}} := \min_{\widehat{u}\in\widehat{\mathcal{U}}} \mathbb{E}[d(U, \widehat{u})]. \tag{9}$$

If the number of source symbols k scales as $\sqrt{n\delta_n}$, we show that a distortion is achievable if, and only if, it is achievable by a separate source-channel code, i.e., the number of symbols required for compressing the source is less than the channel capacity multiplied by the bandwidth mismatch factor (Result 3). Finally, (in Result 1) we also show that if the number of source symbols scales slower than $\sqrt{n\delta_n}$, then arbitrary small distortion levels $D \geq 0$ can be achieved.

Before stating our main result, recall the definition of the standard rate-distortion function [18]

$$R(D) \triangleq \min_{P_{\widehat{U}|U}(\widehat{u}|u):\mathbb{E}_{P_U P_{\widehat{U}|U}}\left[d(\widehat{U},U)\right]\leq D} \mathbb{I}(\widehat{U}, U), \tag{10}$$

and of the covert capacity [13], [14] for binary input alphabets

$$C_{\text{covert}} \triangleq \sqrt{2}\frac{\mathbb{D}\left(\Gamma_{Y|X}(\cdot \mid 1)\|\Gamma_{Y|X}(\cdot \mid 0)\right)}{\sqrt{\chi^2(\Gamma_{Z|X}(\cdot \mid 1)\|\Gamma_{Z|X}(\cdot \mid 0))}}. \tag{11}$$

Recall that the number of source symbols $k = f(n)$ for a given function $f(\cdot)$.

Theorem 1. *For any given function $f(\cdot)$ and vanishing sequence $\{\delta_n\}_{n\geq 1}$ the following holds.*

1) If

$$\lim_{n\to\infty} \frac{f(n)}{\sqrt{n\delta_n}} = 0, \tag{12}$$

then all nonnegative distortions $D \geq 0$ with finite $R(D)$ are (D, δ_n)-admissible.

2) If

$$\lim_{n\to\infty} \frac{f(n)}{\sqrt{n\delta_n}} = \infty, \tag{13}$$

then only trivial distortions $D \geq D_{\text{trivial}}$ are (D, δ_n)-admissible.

3) If

$$\lim_{n\to\infty} \frac{f(n)}{\sqrt{n\delta_n}} = \frac{1}{\gamma}, \tag{14}$$

for some $\gamma > 0$, then D is (D, δ_n)-admissible if, and only if,

$$R(D) \leq \gamma C_{\text{covert}}. \tag{15}$$

Notice that the parameter γ plays the same role as the bandwidth mismatch factor in traditional JSCC.

Proof: Result 1) only requires a proof of achievability and Result 2) only a proof of converse. Result 3) requires both proofs. The two converses are proved in Appendix A. We now prove the two achievability results based on the separate source-channel coding architecture in Figure 2.

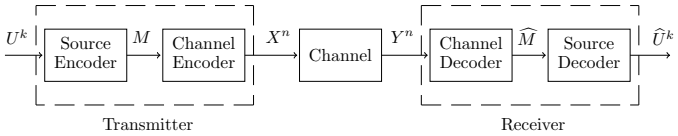

Fig. 2: Separate source and channel coding architecture.

Specifically, the transmitter initially compresses the source sequence U^k into an index $M \in \{1, \ldots, 2^{kR}\}$, which is subsequently encoded into a codeword X^n by a channel code, and then transmitted over the DMC. A channel decoder observe Y^n, a noisy version of X^n, and maps it to a guess $\widehat{M}$ of the index M. This index is then used by the source decoder to produce the reconstruction of the source $\widehat{U}^k$. It is possible to choose a good lossy compression scheme, such as the likelihood lossy compression scheme in [21], so that the reconstruction $\widehat{U}^k$ satisfies the distortion constraint (7) whenever $\widehat{M} = M$ and the rate $R > R(D)$. On the other hand, it can be shown that for any vanishing sequence δ_n there exists a good covert channel code [13] that conveys the message M with vanishing probability of error and at the same time respects the covertness constraint (8) whenever

$$\lim_{n\to\infty} \frac{kR}{\sqrt{n\delta}} < C_{\text{covert}}. \tag{16}$$

Recall that $k = f(n)$ and notice that under Condition (12), Inequality (16) is satisfied for all finite values of R. This establishes Result 1). Under Condition (14), it is possible to find a finite value of $R > R(D)$ satisfying (16) whenever

$$R(D) < \gamma C_{\text{covert}}, \tag{17}$$

thus proving achievability of Result 3). This concludes the desired proofs. ∎

III. Covert Semantic Extraction: Neural Network Experiment

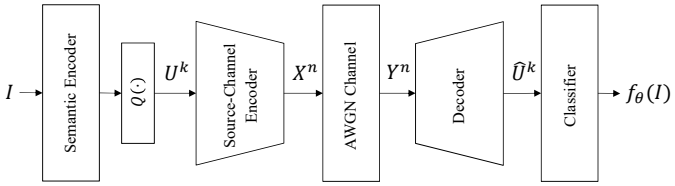

Fig. 3: Neural Network architecture for distributed classification under a covertness constraint.

To illustrate our theoretical findings in a practical context, we train two DNNs for the task of image classification over an Additive White Gaussian Noise (AWGN) channel, see Figure 3. Specifically, we train a first DNN based on a training dataset $\mathcal{D}_{\text{train}}$ of images and corresponding labels to implement the pair of Semantic Encoder (feature extractor) and Classifier, see Figure 3. We subsequently train a second DNN (on the same training dataset $\mathcal{D}_{\text{train}}$) to implement the Source-Channel Encoder and Decoder over the AWGN channel based on the output sequence U^k produced by the previously trained Semantic Encoder, so that communication remains undetectable to an external warden. In contrast to the standard JSCC model, in our experiment the sequence of extracted features U^k (also called semantic vector) is not i.i.d. but has a given distribution dictated by the Semantic Encoder.

Our datasets $\mathcal{D}_{\text{training}}$ and $\mathcal{D}_{\text{test}}$ both consist of image-label pairs (I, y), where $I \in \mathbb{R}^{h\times w\times c}$ (for h, w and c the height, width and the number of channels of the image respectively) and $y \in \{1, \dots, C\}$ (for C indicating the number of classes). If we denote by $f_\theta(I)$ the output of the combined two DNNs (see Figure 3) when the input image is I and the DNNs parameters fixed after the training are described by θ, then the accuracy on the test data, which measures the performance of our model on unseen images, is defined as

$$\text{Acc}(\theta, \mathcal{D}_{\text{test}}) \triangleq \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(I,y)\in\mathcal{D}_{\text{test}}} \mathbb{1}\{f_\theta(I) = y\}. \tag{18}$$

Notice that when combining the two DNNs, we can tweak the parameter k indicating the number of features to be extracted from the images or equivalently the number of source symbols that have to be sent over the AWGN channel. Based on the theoretical findings in the previous Section II-C, our goal is to show that in order to achieve a satisfactory accuracy, the length k of the semantic vector U^k that contains the semantic information used to classify an image must approximately be $\sqrt{n\delta_n}$ for a small δ_n and a large blocklength n. Failure to adhere to this scaling should result in a bad classification accuracy because we expect the channel coding to introduce too many errors.

A. DNN Architectures and Training

As aforementioned, our first DNN comprises a *Semantic Encoder* and a *Classifier*. The Semantic Encoder maps the input image I to a hidden representation (a vector in $\mathbb{R}^k$) to which we apply the binary quantization function $Q(x) \triangleq \mathbb{1}\{x > 0\}$ componentwise. The output of this quantization procedure then provides the binary semantic vector U^k with k a parameter that we can choose in our implementation. This first DNN is trained so as to minimize the cross-entropy[3], i.e., the loss

$$\mathcal{L}^{(1)} = -\lambda_{ce} \sum_{y=1}^{C} y \log(p_y), \tag{19}$$

where p_y represents the probability[4] that the DNN gives to the class y whereas λ_{ce} is a scaling constant that can be adjusted through tuning.

The second DNN implements the Source-Channel Encoder and the Decoder. It simultaneously seeks to minimize the Hamming distortion error for the reconstruction of the semantic vector U^k and aims to achieve covert communication. Instead of constraining the warden's divergence[5] $\mathbb{D}\left(\widehat{Q}^n \parallel \Gamma_{Z|X}^{\otimes n}(\cdot|0^n)\right)$, as we did in the previous section, here we attempt to constrain the power of the transmit signal X^n. In some sense, this can be viewed as a more universal approach as the covertness constraint does not rely on a specific model for the warden. Specifically, we add the loss term

$$\mathcal{L}_{\text{covert}} = \left| \left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i}{\sqrt{n\epsilon}} \right)^2 - 1 \right|, \tag{20}$$

which is motivated by the Central Limit Theorem which states that the noise uncertainty at a receiver suffering from Gaussian noise is in the order of $\sqrt{n}$. The $\sqrt{\epsilon}$ factor allows to adapt to the desired level of covertness. The objective when training the second DNN is thus to minimize the loss

$$\mathcal{L}^{(2)} = \lambda_d \cdot \frac{1}{k} \sum_{i=1}^{k} |\widehat{U}_i - U_i| + \lambda_{\text{covert}} \cdot \mathcal{L}_{\text{covert}}, \tag{21}$$

where λ_d and λ_{covert} are scaling constants that can be adjusted through tuning.

Notice that the quantizer $Q(\cdot)$ as defined earlier in this subsection is non-differentiable and is thus not trainable. To enable an end-to-end differentiable approach, we resort to the "Straight-Through Estimator" [23], which basically sets the gradients with respect to the quantizer to 1 in the backward pass. In fact, the gradients are "straight-through" from the loss function to the model parameters, despite the discontinuity introduced by the discrete sampling in the forward pass.

B. Numerical Results

We use the MNIST-digit dataset [24] with $|\mathcal{D}_{\text{training}}| = 50000$ training samples, $|\mathcal{D}_{\text{test}}| = 10000$ test samples, $C = 10$ classes,

[3] It can also be viewed as a minimization of the KL-divergence between the true distribution P and the DNN distribution Q, i.e. $\mathcal{L}^{(1)} = -\lambda_{ce} \cdot [\mathbb{H}(P) + \mathbb{D}(P\|Q)]$.

[4] Our DNN provides probabilities on the classes, which here we denote $\{p_y\}_{y=1}^{C}$ before deciding on its argmax class.

[5] One could consider the KL divergence of the histograms by resorting to soft histograms [22]. In practice, we find that this leads to poor performance.

and $c = 1$ channel (black and white images). We train the first DNN within 60 epochs and with a fixed learning rate of 0.01, while we use $\lambda_{ce} = 1$. The second DNN is trained with 60 epochs, a learning rate of 0.005, $\lambda_d = 10$, $\lambda_{\text{covert}} = 10$, $\epsilon = 0.01$, $\delta_n = 0.02$ and the noise power of the AWGN channel is fixed at 0.63. For the two DNNs, we set the batch size to 128 and we use the Adam optimizer [25] with $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$.

We consider two different values for the blocklength $n \in \{512, 2048\}$, and start by considering a model where the size of the semantic vector k is in the order of $\sqrt{n\delta_n}$, as indicated by Theorem 1.

1) Square-root covert model: For $n = 512$ we let k take value in $\mathcal{K}^{(512)}_{\text{square-root}} \triangleq \{1, 3, 4, 6, 7\}$ and for $n = 2048$ we let k in $\mathcal{K}^{(2048)}_{\text{square-root}} \triangleq \{2, 4, 5, 8, 10, 11, 12, 14\}$. For each value of n we then optimize the accuracy $\text{Acc}(\theta, \mathcal{D}_{\text{test}})$ in (18) over the value of k, and denote the optimal value by k^*. As indicated by Table I, the accuracy increases with larger blocklength n since a larger blocklength provides more room for error correction over the AWGN channel. Moreover, the number of classes in our experiment is $C = 10$ and thus smaller than 2^{k^*}, so it is beneficial to extract a larger semantic vector than available classes.

Blocklength	Accuracy	Optimal k^*
512	58.45	6
2048	87.44	11

TABLE I: Performance under the square-root covert model.

Our results indicate that the training of the DNNs was successful. In particular the joint source-channel transmission of the semantic vector seems to have been successful, when the size of the semantic vector k is close to $2{\cdot}\sqrt{n\delta_n}$, which for the chosen parameters evaluates to 6.4 and 12.8.

In the following, we further investigate above conclusions. To this end, we run two additional related models with sizes of the semantic vector that are in the order of n. In particular, for the Linear covert model 2) we keep the covertness constraint, which we then remove for the Linear non-covert model 3). The goal is to see whether extracting larger semantic vectors yields better classification performance (i.e. higher accuracy), and whether the conclusions depend on the imposed covertness constraint.

2) Linear covert model: Here, for $n = 512$ the parameter k takes value in $\mathcal{K}^{(512)}_{\text{linear}} \triangleq \{102, 409, 512\}$ and for $n = 2048$ it takes value in $\mathcal{K}^{(2048)}_{\text{linear}} \triangleq \{409, 1638, 2048\}$. We again optimize over the proposed set of k-values.

Blocklength	Accuracy	Optimal k^*
512	10.10	409
2048	11.34	1638

TABLE II: Performance under the linear covert model

3) Non-covert model: Same as the linear covert model, but the loss related to covertness is ignored, i.e., $\lambda_{\text{covert}} = 0$.

As Tables II and III show, the accuracy again increases with the blocklength n. However, while under the covertness constraint the achieved accuracies with the Linear covert model fall short compared to the ones with the Square-root covert model, without the covertness constraint the accuracy is high even at short blocklengths. This last finding indicates that large semantic vectors are beneficial to increase accuracy. In contrast, the low accuracy in the Linear covert model indicates that under a covertness constraint the probability of error over the communication channel is high when the feature vector is in the order of the blocklength, thus compromising the overall performance of the classifier.

Blocklength	Accuracy	Optimal k^*
512	98.88	102
2048	98.95	409

TABLE III: Performance under the linear non-covert model

This is perfectly in line with the theoretical findings of our previous Section II-C.

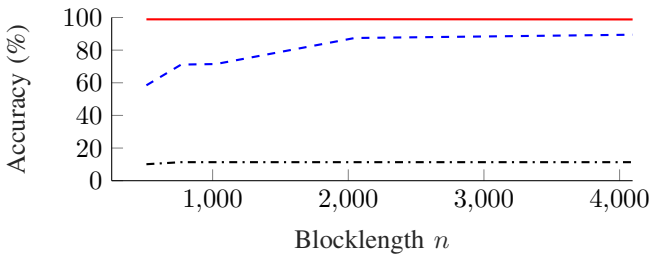

Fig. 4: Accuracy in (%) as a function of the blocklength n for the three models at SNR=1dB. The solid red curve denotes the Non-covert model, the dashed blue curve the Square-root covert model and the dash-dotted black curve the Linear covert model.

In Figure 4, we illustrate the comparison of the accuracies of the three models in function of the blocklengths n. Clearly, the linear covert model saturates and the high probability of communication error severely limits the classification task.

IV. Summary and Discussion

We established the fundamental limits of semantic communication under a covertness constraint by providing sufficient and necessary conditions for a source to be reconstructed with desired distortion at a distant receiver. In particular, we have demonstrated the optimality of source-channel separation for joint source-channel coding under a covertness constraint. Moreover, our experimental setup underscores the feasibility of training a deep neural network to accomplish covert semantic communication as long as it suffices to extract a feature vector of length approximately equal to the square-root of the communication blocklength. This confirms our theoretical findings showing the necessity of the described scaling, similarly to the case of covert data communication.

Future interesting research directions include extensions to setups with many users and with power and resource allocation strategies, as well as identifying the minimum secret-key that is required to ensure covertness.

APPENDIX A
INFORMATION-THEORETIC CONVERSE PROOF

Fix a function $f(\cdot)$ and a vanishing sequence $\{\delta_n\}_{n\geq 1}$. Consider then a sequence (one for each n) of JSCC schemes that satisfies both (7) and (8).

Following the same steps as in [13], it can be shown that:

$$\delta_n \geq n\frac{\bar{\alpha}_n^2}{2}\cdot\left[\chi^2(\Gamma_{Z|X}(\cdot|1)\|\Gamma_{Z|X}(\cdot|0)) + o(1)\right], \quad (22)$$

where $o(1)$ is a decreasing function in n and

$$\bar{\alpha}_n \triangleq \frac{1}{n}\sum_{i=1}^{n}\alpha_{n,i}. \quad (23)$$

Define T to be uniform over $\{1,\ldots,n\}$ independent of all inputs, outputs, source and reconstruction symbols. Then:

$$\mathbb{E}\left[d(U_T,\widehat{U}_T)\right] = \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}d(U_i,\widehat{U}_i)\right] \leq D, \quad (24)$$

where the last step holds by (7). Continue to bound:

$$\begin{aligned}
\mathbb{I}(U^k;\widehat{U}^k) &= \mathbb{H}(U^k) - \mathbb{H}(U^k \mid \widehat{U}^k) && (25)\\
&\overset{(a)}{\geq} \sum_{i=1}^{k}\mathbb{H}(U_i) - \mathbb{H}(U_i \mid \widehat{U}_i) = \sum_{i=1}^{k}\mathbb{I}(U_i;\widehat{U}_i) && (26)\\
&= k\mathbb{I}(U_T;\widehat{U}_T \mid T) \overset{(b)}{\geq} k\mathbb{I}(U_T;\widehat{U}_T) \overset{(c)}{\geq} kR(D), && (27)
\end{aligned}$$

where (a) holds because conditioning reduces entropy and U_i is independent of S; (b) holds by the independence of U_T and T; and (c) holds by the definition of $R(D)$ and because U_T is distributed according to P_U and $\widehat{U}_T$ satisfies (24).

Next, notice that the Markov Chain $U^k \leftrightarrow (X^n,S) \leftrightarrow (Y^n,S) \leftrightarrow \widehat{U}^k$ implies by the Data Processing Inequality:

$$\begin{aligned}
\mathbb{I}(U^k;\widehat{U}^k) &\leq \mathbb{I}(U^k;\widehat{U}^k,S) \overset{(a)}{=} \mathbb{I}(U^k;\widehat{U}^k \mid S) && (28)\\
&\leq \mathbb{I}(X^n;Y^n \mid S) && (29)\\
&= \sum_{i=1}^{n}\mathbb{H}(Y_i \mid Y^{i-1}) - \mathbb{H}(Y_i \mid Y^{i-1},X^n,S) && (30)\\
&\overset{(b)}{\leq} \sum_{i=1}^{n}\mathbb{H}(Y_i) - \mathbb{H}(Y_i \mid X_i) && (31)\\
&\overset{(c)}{\leq} n\bar{\alpha}_n\mathbb{D}\left(\Gamma_{Y|X}(\cdot \mid 1)\|\Gamma_{Y|X}(\cdot \mid 0)\right) && (32)\\
&\overset{(d)}{\leq} \sqrt{n\delta_n}C_{\text{covert}}, && (33)
\end{aligned}$$

where (a) holds because $\mathbb{I}(U^k;S)=0$; (b) by the memoryless channel; (c) by [13, Lemma 1]; and (d) by (22) and the definition of C_{covert} in (11). Combining (27) with (33) yields

$$R(D) \leq \varlimsup_{n\to\infty}\frac{\sqrt{n\delta_n}}{k}C_{\text{covert}}. \quad (34)$$

Recalling that $k = f(n)$, we can conclude that whenever $\lim_{n\to\infty}\frac{f(n)}{\sqrt{n\delta_n}} = 0$, then Inequality (34) implies that $R(D)=0$, allowing only for the trivial distortion D_{trivial}, thus establishing the converse to Result 2). On the other hand, if $\lim_{n\to\infty}\frac{f(n)}{\sqrt{n\delta_n}} = \frac{1}{\gamma}$, then Inequality (34) is only satisfied if $R(D) \leq \gamma C_{\text{covert}}$, thus establishing the converse to Result 3).

REFERENCES

[1] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE JSAC*, vol. 41, no. 1, pp. 5–41, 2023.

[2] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[3] V. Kostina and S. Verdú, "Lossy joint source-channel coding in the finite blocklength regime," *IEEE TIT*, vol. 59, no. 5, pp. 2545–2575, 2013.

[4] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE JSAC*, vol. 41, no. 8, pp. 2645–2657, 2023.

[5] T.-Y. Tung and D. Gündüz, "Deep joint source-channel and encryption coding: Secure semantic communications," in *IEEE ICC*, 2023, pp. 5620–5625.

[6] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Wireless image retrieval at the edge," *IEEE JSAC*, vol. 39, no. 1, pp. 89–100, 2021.

[7] N. Farsad, M. Rao, and A. Goldsmith, "Deep learning for joint source-channel coding of text," in *IEEE ICASSP*, 2018, pp. 2326–2330.

[8] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE TSP*, vol. 69, pp. 2663–2675, 2021.

[9] E. Erdemir, P. L. Dragotti, and D. Gündüz, "Privacy-aware communication over a wiretap channel with generative networks," in *IEEE ICASSP*, 2022, pp. 2989–2993.

[10] T. Marchioro, N. Laurenti, and D. Gündüz, "Adversarial networks for secure wireless communications," in *IEEE ICASSP*, 2020, pp. 8748–8752.

[11] G. Nan, Z. Li, J. Zhai, Q. Cui, G. Chen, X. Du, X. Zhang, X. Tao, Z. Han, and T. Q. S. Quek, "Physical-layer adversarial robustness for deep learning-based semantic communications," *IEEE JSAC*, vol. 41, no. 8, pp. 2592–2608, 2023.

[12] B. A. Bash, D. Goeckel, and D. Towsley, "Limits of reliable communication with low probability of detection on awgn channels," *IEEE JSAC*, vol. 31, no. 9, pp. 1921–1930, 2013.

[13] M. R. Bloch, "Covert communication over noisy channels: A resolvability perspective," *IEEE TIT*, vol. 62, no. 5, pp. 2334–2354, 2016.

[14] L. Wang, G. W. Wornell, and L. Zheng, "Fundamental limits of communication with low probability of detection," *IEEE TIT*, vol. 62, no. 6, pp. 3493–3503, 2016.

[15] A. Bounhar, M. Sarkiss, and M. Wigger, "Covert distributed detection over discrete memoryless channels," 2024. [Online]. Available: https://perso.telecom-paristech.fr/wigger/bounhar-ISIT24.pdf

[16] ——, "Mixing a covert and a non-covert user," in *IEEE ISIT*, 2023, pp. 2577–2582.

[17] ——, "Covert multi-access communication with a non-covert user," 2024. [Online]. Available: https://perso.telecom-paristech.fr/wigger/bounhar-ICC24.pdf

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Ed.* Wiley, 2006.

[19] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[20] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*. New York, NY, USA: Springer-Verlag, 2005.

[21] E. C. Song, P. Cuff, and H. V. Poor, "The likelihood encoder for lossy compression," *IEEE TIT*, vol. 62, no. 4, pp. 1836–1849, 2016.

[22] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," *Nature communications*, vol. 8, no. 1, p. 13890, 2017.

[23] Y. Bengio, N. Léonard, and A. C. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *CoRR*, vol. abs/1308.3432, 2013. [Online]. Available: http://arxiv.org/abs/1308.3432

[24] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

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