

Article

Semantic Information Recovery in Wireless Networks

Edgar Beck , Carsten Bockelmann  and Armin Dekorsy 

Department of Communications Engineering, University of Bremen, 28359 Bremen, Germany; bockelmann@ant.uni-bremen.de (C.B.); dekorsy@ant.uni-bremen.de (A.D.)

* Correspondence: beck@ant.uni-bremen.de

Abstract: Motivated by the recent success of Machine Learning (ML) tools in wireless communications, the idea of semantic communication by Weaver from 1949 has gained attention. It breaks with Shannon's classic design paradigm by aiming to transmit the meaning of a message, i.e., semantics, rather than its exact version and, thus, enables savings in information rate. In this work, we extend the fundamental approach from Basu et al. for modeling semantics to the complete communications Markov chain. Thus, we model semantics by means of hidden random variables and define the semantic communication task as the data-reduced and reliable transmission of messages over a communication channel such that semantics is best preserved. We consider this task as an end-to-end Information Bottleneck problem, enabling compression while preserving relevant information. As a solution approach, we propose the ML-based semantic communication system SINFONY and use it for a distributed multipoint scenario; SINFONY communicates the meaning behind multiple messages that are observed at different senders to a single receiver for semantic recovery. We analyze SINFONY by processing images as message examples. Numerical results reveal a tremendous rate-normalized SNR shift up to 20 dB compared to classically designed communication systems.

Keywords: semantic communication; wireless communications; wireless networks; infomax; information bottleneck; machine learning; task-oriented communication; goal-oriented communication



Citation: Beck, E.; Bockelmann, C.; Dekorsy, A. Semantic Information Recovery in Wireless Networks. *Sensors* **2023**, *23*, 6347. <https://doi.org/10.3390/s23146347>

Academic Editor: Changchuan Yin

Received: 12 June 2023

Revised: 1 July 2023

Accepted: 7 July 2023

Published: 12 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

When Shannon laid the theoretical foundation of the research area of communications engineering back in 1948, he deliberately excluded semantic aspects from the system design [1,2]. In fact, the idea of addressing semantics in communications arose shortly after Shannon's work in [2], but it remained largely unexplored. Since then, the design focus of communication systems has been on digital error-free point-to-point symbol transmission.

Today, the systems already operate close to the Shannon limit calling for a paradigm shift towards including semantic content of messages in the system design. For example, the data traffic growth still continues with the emergence of the Internet-of-Everything including, e.g., autonomous driving and virtual reality, and cannot be managed by semantics-agnostic communication as it limits the achievable efficiency in terms of bandwidth, power, latency, and complexity trade-offs [3]. Other notable examples include wireless sensor networks, broadcast scenarios, and non-ergodic channels where separation of source and channel coding according to Shannon's digital design paradigm is generally suboptimal [4,5].

Owing to the great success of Artificial Intelligence (AI) and, in particular, its sub-domain Machine Learning (ML), ML tools have been recently investigated for wireless communications and has shown promising application for improving the performance complexity trade-off [6–8]. Now, ML with its ability to extract features appears to be a proper means to realize a semantic design. Further, we note that the latter design is supported and possibly enabled by the 6G vision of integrating AI and ML on all layers of the communications system design, i.e., by an ML-native air interface.

Motivated by these new ML tools, and driven by the unprecedented needs of the next wireless communication standard, 6G, in terms of data rate, latency, and power, the idea

of semantic communication has received considerable attention [2,9–13]. It breaks with the existing classic design paradigms by including semantics in the design of the wireless transmission. The goal of such a transmission is, therefore, to deliver the required data from which the highest levels of quality of information may be derived, as perceived by the application and/or the user. More precisely, semantic communication aims to transmit the meaning of a message rather than its exact version and hence enables compression and coding to the actual semantic content. Thus, savings in bandwidth, power, and complexity are expected.

In the following, we first summarize in Section 2 related work on semantic communication and justify our main contributions in Section 3. In Section 4.1, we reinterpret Weaver's philosophical considerations paving the way for our proposed theoretical framework in Section 4. Finally, in Sections 5 and 6, we provide one numerical example of semantic communication, i.e., SINFONY, and summarize the main results, respectively.

2. Related Work

The notion of semantic communication traces back to Weaver [2] who reviewed Shannon's information theory [1] in 1949 and amended considerations with regard to semantic content of messages. Often quoted is his statement that "*there seem to be [communication] problems at three levels*" [2]:

- A. How accurately can the symbols of communication be transmitted? (The technical problem.)
- B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)
- C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)

Since then semantic communication was mainly investigated from a philosophical point of view, see, e.g., [14,15].

The generic model of Weaver was revisited by Bao, Basu et al. in [16,17] where the authors define semantic information source and semantic channel. In particular, the authors consider a semantic source that "*observes the world and generates meaningful messages characterizing these observations*" [17]. The source is equivalent to conclusions, i.e., "models" of the world, that are unequivocally drawn following a set of known inference rules based on observation of messages. In [16], the authors consider joint semantic compression and channel coding at Level B with the classic transmission system, i.e., Level A, as the (semantic) channel. In contrast, [17] only deals with semantic compression and uses a different definition of the semantic channel (which we will make use of in this article): It is equal to the entailment relations between "models" and "messages". By this means, the authors are able to derive semantic counterparts of the source and channel coding theorems. However, as the authors admit, these theorems do not tell how to develop optimal coding algorithms and the assumption of a logic-based model-theoretical description leads to "*many non-trivial simplifications*" [16].

In [18], the authors follow a different approach in the context of Natural Language Processing (NLP). They define semantic similarity as a semantic error measure using taxonomies, i.e., human knowledge graphs, to quantify the distance between the meanings of two words. Based on this metric, communication of a finite set of words is modeled as a Bayesian game from game theory and optimized for improved semantic transmission over a binary symmetric channel.

Recently, drawing inspiration from Weaver, Bao, Basu et al. [2,16,17] and enabled by the rise of ML in communications research, Deep Neural Network (DNN)-based NLP techniques, i.e., transformer networks, were introduced in AutoEncoders (AEs) for the task of text transmission [19–21]. The aim of these techniques is to learn compressed hidden representations of the semantic content of sentences to improve communication efficiency, but the exact recovery of the source (text) is the main objective. The approach improves performance in semantic metrics, especially at low SNR compared to classical digital transmissions. It has been adapted to numerous other problems, e.g., speech transmission [22,23]

and multi-user transmission with multi-modal data [24]. Even knowledge graphs, i.e., a prior knowledge base, were incorporated into the transformer-based AE design to improve inference at the receiver side and, thus, text recovery [25].

Not considering Weaver's idea of semantic communication in particular, the authors in [26] show, for the first time, that task-oriented communications (Level C) for edge cloud transmission can be mathematically formulated as an Information Bottleneck (IB) optimization problem. Moreover, for solving the IB problem, they introduce a DNN-based approximation and show its applicability for the specific task of edge cloud transmission. The terminus "*semantic information*" is only mentioned once in [26] referring to Joint Source-Channel Coding (JSCC) of text from [19] using recurrent neural networks. In [19], the authors observe that sentences that express the same idea have embeddings that are close together in Hamming distance. But they use cross entropy between words and estimated words as the loss function and use the word error rate as the performance measure, which both do not reflect if two sentences have the same meaning but rather that both are exactly the same.

As a result, semantic communication is still a nascent field; it still remains unclear what this term exactly means [27] and, in particular, its distinction from JSCC [19,28]. As a result, many survey papers aim to provide an interpretation, see, e.g., [9–13]. We will revisit this issue in Section 4.

3. Main Contributions

The main contributions of this article are:

- Motivated by the approach of Bao, Basu et al. [16,17], we adopt the terminus of a semantic source. Inspired by Weaver's notion, we bring it to the context of communications by considering the complete Markov chain, including semantic source, communications source, transmit signal, communication channel, and received signal in contrast to both [16,17]. Further, we also extend beyond the example of deterministic entailment relations between "models" and "messages" based on propositional logic in [16,17] to probabilistic semantic channels.
- We define the task of semantic communication in the sense that we perform data compression, coding, and transmission of messages observed such that the semantic Random Variable (RV) at a recipient is best preserved. Basically, we implement joint source-channel coding of messages conveying the semantic RV, but not differentiating between Levels A and B. We formulate the semantic communication design either as an Information Maximization or as an Information Bottleneck (IB) optimization problem [29–31].
 - Although the approach pursued here again leads to an IB problem as in [26], our article introduces a new classification and perspective of semantic communication and different ML-based solution approaches. Different from [26], we solve the IB problem maximizing the mutual information for a fixed encoder output dimension that bounds the information rate.
 - The publication presented here differs also both in the interpretation of what is meant by semantic information and in the objective of recovering this semantic information from approaches to semantic communication presented in the literature like, e.g., [21,32].
- Finally, we propose the ML-based semantic communication system SINFONY for a distributed multipoint scenario in contrast to [26]: SINFONY communicates the meaning behind multiple messages that are observed at different senders to a single receiver for semantic recovery. Compared to the distributed scenario in [33,34], we include the communication channel.
- We analyze SINFONY by processing images as an example of messages. Notably, numerical results reveal a tremendous rate-normalized SNR shift up to 20 dB compared to classically designed communication systems.

4. A Framework for Semantics

4.1. Philosophical Considerations

Despite the much-renewed interest, research on semantic communication is still in its infancy and recent work reveals a differing understanding of the word *semantics*. In this work, we contribute our interpretation. To motivate it, we shortly revisit the research birth hour of communications from a philosophical point of view; its theoretical foundation was laid by Shannon in his landmark paper [1] in 1948.

He stated that *“Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem”*. In fact, this viewpoint abstracts all kinds of information one may transmit, e.g., oral and written speech, sensor data, etc., and also lays the foundation for the research area of Shannon information theory. Thus, it found its way into many other research areas where data or information are processed, including Artificial Intelligence (AI) and especially its subdomain Machine Learning (ML).

Weaver saw this broad applicability of Shannon’s theory back in 1949. In his comprehensive review of [1], he first states that *“there seem to be [communication] problems at three levels”* [2] already mentioned in Section 2. These three levels are quoted in recent works, where Level C is oftentimes referred to as goal-oriented communication instead [10].

But we note that, in his concluding section, he then questions this segmentation. He argues for the generality of the theory at Level A for all levels and *“that the interrelation of the three levels is so considerable that one’s final conclusion may be that the separation into the three levels is really artificial and undesirable”*.

It is important to emphasize that the separation is rather arbitrary. We agree with Weaver’s statement because the most important point that is also the focus herein is the definition of the term semantics, e.g., by Basu et al. [16,17]. Note that the entropy of the semantics is less than or equal to the entropy of the messages. Consequently, we can save information rate by introducing meaning or context. In fact, we are able to add arbitrarily many levels of semantic details to the communication problem and optimize communications for a specific semantic background, e.g., an application or human.

4.2. Semantic System Model

4.2.1. Semantic Source and Channel

Now, we will define our information-theoretic system model of semantic communication. Figure 1 shows the schematic of our model. We assume the existence of a semantic source, described as a hidden target multivariate Random Variable (RV) $\mathbf{z} \in \mathcal{M}_z^{N_z \times 1}$ from a domain \mathcal{M}_z of dimension N_z distributed according to a probability density or mass function (pdf/pmf) $p(\mathbf{z})$. To simplify the discussion, we assume it to be discrete and memoryless. For the remainder of the article, note that the domain of all RVs \mathcal{M} may be either discrete or continuous. Further, we note that the definition of entropy for discrete and continuous RVs differs. For example, the differential entropy of continuous RVs may be negative whereas the entropy of discrete RVs is always positive [35]. Without loss of generality, we will thus assume all RVs either to be discrete or to be continuous. In this work, we avoid notational clutter by using the expected value operator, replacing the integral by summation over discrete RVs, the equations are also valid for discrete RVs and vice versa.

Our approach is similar to that of [16,17]. In [16,17], the semantic source is described by “models of the world”. (Note that, in [17], the semantic information source is defined as a tuple $(\mathbf{z}, \mathbf{s}, p(\mathbf{z}, \mathbf{s}), L)$. In this original notation, \mathbf{z} is the model, \mathbf{s} the message, $p(\mathbf{z}, \mathbf{s})$ the joint distribution of \mathbf{z} and \mathbf{s} , and L is the deterministic formal language.) In [17], a semantic channel then generates messages through entailment relations between “models” and “messages”. We will call these “messages” source signal and define it to be a RV $\mathbf{s} \in \mathcal{M}_s^{N_s \times 1}$ as it is usually observed and enters the communication system. In the classic Shannon design, the aim is to reconstruct the source \mathbf{s} as accurately as possible at the receiver side. Further, we note that the authors in [17] considered the example of a semantic channel with deterministic entailment relations between \mathbf{z} and \mathbf{s} based on propositional logic. In

this article, we go beyond this assumption and consider probabilistic semantic channels modeled by distribution $p(\mathbf{s}|\mathbf{z})$ that include the entailment in [17] as special cases, i.e., $p(\mathbf{s}|\mathbf{z}) = \delta(\mathbf{s} - f(\mathbf{z}))$ where $\delta(\cdot)$ is the Dirac delta function and $f(\cdot)$ is any generic function. Our viewpoint is motivated by the recent success of pattern recognition tools that advanced the field of AI in the 2010s and may be used to extract semantics [7].

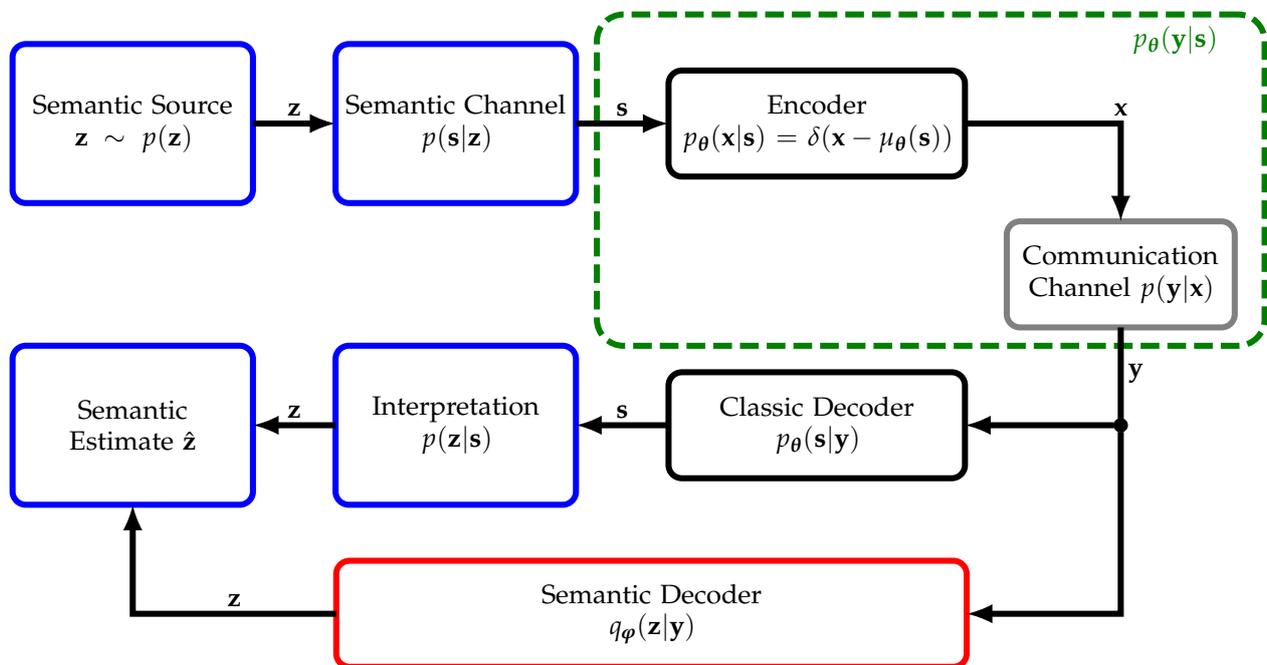


Figure 1. Block diagram of the considered semantic system model.

Our approach also extends models as in [21]. There, the authors design a semantic communication system for the transmission of written language/text similar to [19] using transformer networks. In contrast to our work, [21] does not define meaning as RV \mathbf{z} . The objective in [21] is to reconstruct \mathbf{s} (sentences) as well as possible, rather than the meaning (RV \mathbf{z}) conveyed in \mathbf{s} . Optimization is completed with regard to a loss function consisting of two parts, cross entropy between language input \mathbf{s} and output estimate $\hat{\mathbf{s}}$, as well as a scaled mutual information term between transmit signal \mathbf{x} and receive signal \mathbf{y} . After optimization, the authors measure semantic performance by some semantic metric $\mathcal{L}(\mathbf{s}, \hat{\mathbf{s}})$.

We now provide an example to explain what we understand under a semantic source \mathbf{z} and channel $p(\mathbf{s}|\mathbf{z})$. Let us imagine a biologist who has an image of a tree. The biologist wants to know what kind of tree it is by interpreting the observed data (image). In this case, the semantic source \mathbf{z} is a multivariate RV composed of a categorical RV with M tree classes. For any realization (sample value) \mathbf{z}_i of the semantic source, the semantic channel $p(\mathbf{s}|\mathbf{z})$ then outputs with some probability one image \mathbf{s}_i of a tree conveying characteristics of \mathbf{z} , i.e., its meaning. Note that the underlying meaning of the same sensed data (message) can be different for other recipients, e.g., humans or tasks/applications, i.e., in other semantic contexts. Imagine a child, i.e., a person with different characteristics (personality, expertise, knowledge, goals, and intentions) than the biologist, who is only interested if he/she can climb up this tree or whether the tree provides shade. Thus, we include the characteristics of the sender and receiver in RV \mathbf{z} and consider it directly in compression and encoding.

Compared to [16], we, therefore, argue that we also include level C by semantic source and channel since context can be included on increasing layers of complexity. First, a RV \mathbf{z}_1 might capture the interpretation, like the classification of images or sensor data. Moving beyond the first semantic layer, then a RV \mathbf{z}_2 might expand this towards a more general goal, like keeping a constant temperature in power plant control. In fact, we can add or remove context, i.e., semantics and goals, arbitrarily often according to the human or

application behind, and we can optimize the overall (communication) system with regard to $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_i$, respectively.

As a last remark, we note that we basically defined probabilistic semantic relationships, and it remains the question of how exactly they might look. In our example, the meaning of the images needs to be labeled into real-world data pairs $\{\mathbf{s}_i, \mathbf{z}_i\}$ by experts/humans, since image recognition lacks precise mathematical models. This is also true for NLP [21]; how can we measure if two sentences have the same meaning, i.e., how does the semantic space look like? In contrast, in [17], the authors are able to solve their well-defined technical problem (motion detection) by a model-driven approach. We can thus distinguish between model and data-driven semantics, which both can be handled within Shannon's information theory.

4.2.2. Semantic Channel Encoding

After the semantic source and channel in Figure 1, we extend upon [16] by differentiating between "message"/source signal \mathbf{s} and transmit signal $\mathbf{x} \in \mathcal{M}_x^{N_{Tx} \times 1}$. Our challenge is to encode the source signal \mathbf{s} onto the transmit signal vector \mathbf{x} for reliable semantic communication through the physical communication channel $p(\mathbf{y}|\mathbf{x})$, where $\mathbf{y} \in \mathcal{M}_y^{N_{Rx} \times 1}$ is the received signal vector. We assume the encoder $p_\theta(\mathbf{x}|\mathbf{s})$ to be parametrized by a parameter vector $\theta \in \mathbb{R}^{N_\theta \times 1}$. Note that $p_\theta(\mathbf{x}|\mathbf{s})$ is probabilistic here, but assumed to be deterministic in communications with $p_\theta(\mathbf{x}|\mathbf{s}) = \delta(\mathbf{x} - \mu_\theta(\mathbf{s}))$ and encoder function $\mu_\theta(\mathbf{s})$.

In summary, in contrast to both [16,17], we consider the complete Markov chain $\mathbf{z} \leftrightarrow \mathbf{s} \leftrightarrow \mathbf{x} \leftrightarrow \mathbf{y}$ including semantic source \mathbf{z} , communications source \mathbf{s} , transmit signal \mathbf{x} and receive signal \mathbf{y} . By this means, we distinguish from [17] which only deals with semantic compression, and [16] which is about joint semantic compression and channel coding (Level B). In [16], the authors consider the classic transmission system (Level A) as the (semantic) channel (not to be confused with the definition of the semantic channel in [17] which we make use of in this publication).

At the receiver side, one approach is maximum a posteriori decoding with regard to RV \mathbf{s} that uses the posterior $p_\theta(\mathbf{s}|\mathbf{y})$, being deduced from prior $p(\mathbf{s})$ and likelihood $p_\theta(\mathbf{y}|\mathbf{s})$ by application of Bayes law. Based on the estimate of \mathbf{s} , then the receiver interprets the actual semantic content \mathbf{z} by $p(\mathbf{z}|\mathbf{s})$.

Another approach we propose is to include the semantic hidden target RV \mathbf{z} into the design by processing $p_\theta(\mathbf{z}|\mathbf{y})$. If the calculation of the posterior is intractable, we can replace $p_\theta(\mathbf{z}|\mathbf{y})$ by the approximation $q_\varphi(\mathbf{z}|\mathbf{y})$, i.e., the semantic decoder, with parameters $\varphi \in \mathbb{R}^{N_\varphi \times 1}$. We expect the following benefit: We assume the entropy $\mathcal{H}(\mathbf{z}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}[-\ln p(\mathbf{z})]$ of the semantic RV \mathbf{z} , i.e., the actual semantic uncertainty or information content, to be less or equal to the entropy $\mathcal{H}(\mathbf{s})$ of the source \mathbf{s} , i.e., $\mathcal{H}(\mathbf{z}) \leq \mathcal{H}(\mathbf{s})$. There, $\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})}[f(\mathbf{x})]$ denotes the expected value of $f(\mathbf{x})$ with regard to both discrete or continuous RVs \mathbf{x} . Consequently, since we would like to preserve the relevant, i.e., semantic, RV \mathbf{z} rather than \mathbf{s} , we can compress more s.t. preserving \mathbf{z} conveyed in \mathbf{s} . Note that in semantic communication the relevant variable is \mathbf{z} , not \mathbf{s} . Thus, processing $p_\theta(\mathbf{s}|\mathbf{y})$ without taking \mathbf{z} into consideration resembles the classical approach. Instead of using (and transmitting) \mathbf{s} for inference of \mathbf{z} , we now want to find a compressed representation \mathbf{y} of \mathbf{s} containing the relevant information about \mathbf{z} .

4.3. Semantic Communication Design via InfoMax Principle

After explaining the system model and the basic components, we are able to approach a semantic communication system design. We first define an optimization problem to obtain the encoder $p_\theta(\mathbf{x}|\mathbf{s})$ following the Information Maximization (InfoMax) principle from an information theoretic perspective [35]. Thus, we like to find the distribution $p_\theta(\mathbf{x}|\mathbf{s})$

that maps \mathbf{s} to a representation \mathbf{x} such that most information of the relevant RV \mathbf{z} is included in \mathbf{y} , i.e., we maximize the Mutual Information (MI) $I(\mathbf{z}; \mathbf{y})$ with regard to $p_{\theta}(\mathbf{x}|\mathbf{s})$ [36]:

$$\arg \max_{p_{\theta}(\mathbf{x}|\mathbf{s})} I_{\theta}(\mathbf{z}; \mathbf{y}) \quad (1)$$

$$= \arg \max_{p_{\theta}(\mathbf{x}|\mathbf{s})} \mathbb{E}_{\mathbf{z}, \mathbf{y} \sim p_{\theta}(\mathbf{z}, \mathbf{y})} \left[\ln \frac{p_{\theta}(\mathbf{z}, \mathbf{y})}{p(\mathbf{z})p_{\theta}(\mathbf{y})} \right] \quad (2)$$

$$= \arg \max_{p_{\theta}(\mathbf{x}|\mathbf{s})} \mathcal{H}(\mathbf{z}) - \mathcal{H}(p_{\theta}(\mathbf{z}, \mathbf{y}), p_{\theta}(\mathbf{z}|\mathbf{y})) \quad (3)$$

$$= \arg \max_{p_{\theta}(\mathbf{x}|\mathbf{s})} \mathbb{E}_{\mathbf{z}, \mathbf{y} \sim p_{\theta}(\mathbf{z}, \mathbf{y})} [\ln p_{\theta}(\mathbf{z}|\mathbf{y})]. \quad (4)$$

There, $\mathcal{H}(p(\mathbf{x}), q(\mathbf{x})) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [-\ln q(\mathbf{x})]$ is the cross entropy between two pdfs/pmfs $p(\mathbf{x})$ and $q(\mathbf{x})$. Note independence from θ in $\mathcal{H}(\mathbf{z})$ and dependence in $p_{\theta}(\mathbf{z}|\mathbf{y})$ and $p_{\theta}(\mathbf{z}, \mathbf{y})$ through the Markov chain $\mathbf{z} \rightarrow \mathbf{s} \rightarrow \mathbf{y}$. Problem (1) is concave with regard to the encoder $p_{\theta}(\mathbf{x}|\mathbf{s})$ for fixed $p(\mathbf{s})$ [37], but not necessarily concave with regard to the encoder parameters θ . For example, it is non-concave if the encoder function is non-convex with regard to its parameters being typically the case with DNN encoders. It is worth mentioning that we so far have not set any constraint on the variables we deal with. Hence, the form of $p_{\theta}(\mathbf{y}|\mathbf{s})$ has to be constrained to avoid learning a trivial identity mapping $\mathbf{y} = \mathbf{s}$. We indeed constrain the optimization by our communication channel $p(\mathbf{y}|\mathbf{x})$ we assume to be given.

If the calculation of the posterior $p_{\theta}(\mathbf{z}|\mathbf{y})$ in (4) is intractable, we are able to replace it by a variational distribution $q_{\varphi}(\mathbf{z}|\mathbf{y})$ with parameters φ . Similar to the transmitter, DNNs are usually proposed [21,38] for the design of the approximate posterior $q_{\varphi}(\mathbf{z}|\mathbf{y})$ at the receiver. To improve the performance complexity trade-off, the application of *deep unfolding* can be considered, a model-driven learning approach that introduces model knowledge of $p_{\theta}(\mathbf{s}, \mathbf{x}, \mathbf{y}, \mathbf{z})$ to create $q_{\varphi}(\mathbf{z}|\mathbf{y})$ [8,39]. With $q_{\varphi}(\mathbf{z}|\mathbf{y})$, we are able to define a Mutual Information Lower Bound (MILBO) [36] similar to the well-known Evidence Lower Bound (ELBO) [7]:

$$I_{\theta}(\mathbf{z}; \mathbf{y}) \geq \mathbb{E}_{\mathbf{z}, \mathbf{y} \sim p_{\theta}(\mathbf{z}, \mathbf{y})} [\ln q_{\varphi}(\mathbf{z}|\mathbf{y})] \quad (5)$$

$$= \mathbb{E}_{\mathbf{y} \sim p_{\theta}(\mathbf{y})} [\mathbb{E}_{\mathbf{z} \sim p_{\theta}(\mathbf{z}|\mathbf{y})} [\ln q_{\varphi}(\mathbf{z}|\mathbf{y})]] \quad (6)$$

$$= -\mathbb{E}_{\mathbf{y} \sim p_{\theta}(\mathbf{y})} [\mathcal{H}(p_{\theta}(\mathbf{z}|\mathbf{y}), q_{\varphi}(\mathbf{z}|\mathbf{y}))] \quad (7)$$

$$= -\mathcal{L}_{\theta, \varphi}^{\text{CE}}. \quad (8)$$

The lower bound holds since $-\mathcal{H}(p_{\theta}(\mathbf{z}, \mathbf{y}), p_{\theta}(\mathbf{z}|\mathbf{y}))$ itself is a lower bound of the expression in (3) and $\mathbb{E}_{\mathbf{z}, \mathbf{y} \sim p_{\theta}(\mathbf{z}, \mathbf{y})} [\ln p_{\theta}(\mathbf{z}|\mathbf{y}) / \ln q_{\varphi}(\mathbf{z}|\mathbf{y})] \geq 0$. Now, we can calculate optimal values of θ and φ of our semantic communication design by minimizing the amortized cross entropy $\mathcal{L}_{\theta, \varphi}^{\text{CE}}$ in (7), i.e., marginalized across observations \mathbf{y} [8].

Thus, the idea is to learn parametrizations of the transmitter discriminative model and of the variational receiver posterior, e.g., by AEs or reinforcement learning. Note that, in our semantic problem (1), we do not auto-encode the hidden \mathbf{z} itself, but encode \mathbf{s} to obtain \mathbf{z} by decoding. This can be seen from Figure 1 and by rewriting the amortized cross entropy (7) and (8):

$$\mathcal{L}_{\theta, \varphi}^{\text{CE}} = \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} [\mathcal{H}(p_{\theta}(\mathbf{z}|\mathbf{y}), q_{\varphi}(\mathbf{z}|\mathbf{y}))] \quad (9)$$

$$= \mathbb{E}_{\mathbf{s}, \mathbf{x}, \mathbf{y}, \mathbf{z} \sim p_{\theta}(\mathbf{s}, \mathbf{x}, \mathbf{y}, \mathbf{z})} [-\ln q_{\varphi}(\mathbf{z}|\mathbf{y})] \quad (10)$$

$$= \mathbb{E}_{\mathbf{s}, \mathbf{z} \sim p(\mathbf{s}, \mathbf{z})} \left[\mathbb{E}_{\mathbf{x} \sim p_{\theta}(\mathbf{x}|\mathbf{s})} \left[\mathbb{E}_{\mathbf{y} \sim p(\mathbf{y}|\mathbf{x})} [-\ln q_{\varphi}(\mathbf{z}|\mathbf{y})] \right] \right].$$

We can further prove the amortized cross entropy to be decomposable into

$$\mathcal{L}_{\theta, \varphi}^{\text{CE}} = \mathbb{E}_{\mathbf{y} \sim p_{\theta}(\mathbf{y})} [\mathbb{E}_{\mathbf{z} \sim p_{\theta}(\mathbf{z}|\mathbf{y})} [-\ln q_{\varphi}(\mathbf{z}|\mathbf{y}) + \ln p_{\theta}(\mathbf{z}|\mathbf{y}) - \ln p_{\theta}(\mathbf{z}|\mathbf{y})]] \quad (11)$$

$$= \mathbb{E}_{\mathbf{y} \sim p_{\theta}(\mathbf{y})} [D_{\text{KL}}(p_{\theta}(\mathbf{z}|\mathbf{y}) \parallel q_{\varphi}(\mathbf{z}|\mathbf{y}))] + \underbrace{\mathcal{H}(\mathbf{z}|\mathbf{y})}_{=-I_{\theta}(\mathbf{z};\mathbf{y}) + \mathcal{H}(\mathbf{z})} \quad (12)$$

$$= \mathcal{H}(\mathbf{z}) - \underbrace{I_{\theta}(\mathbf{z};\mathbf{y})}_{\text{enc. objective}} + \underbrace{\mathbb{E}_{\mathbf{y} \sim p_{\theta}(\mathbf{y})} [D_{\text{KL}}(p_{\theta}(\mathbf{z}|\mathbf{y}) \parallel q_{\varphi}(\mathbf{z}|\mathbf{y}))]}_{\text{dec. objective}}. \quad (13)$$

In the end, maximization of the MILBO with regard to θ and φ balances maximization of the mutual information $I_{\theta}(\mathbf{z};\mathbf{y})$ and minimization of the Kullback–Leibler (KL) divergence $D_{\text{KL}}(p_{\theta}(\mathbf{z}|\mathbf{y}) \parallel q_{\varphi}(\mathbf{z}|\mathbf{y}))$. The former objective can be seen as a regularization term that favors encoders with high mutual information, for which decoders can be learned that are close to the true posterior.

4.4. Classical Design Approach

If we consider classical communication design approaches, we would solve the problem

$$\arg \max_{p_{\theta}(\mathbf{x}|\mathbf{s})} I(\mathbf{s};\mathbf{y}) \quad (14)$$

which relates to Joint Source–Channel Coding (JSCC). There, the aim is to find a representation \mathbf{x} that retains a significant amount of information about the source signal \mathbf{s} in \mathbf{y} . Again, we can apply the lower bound (8). In fact, bounding (14) by (8) shows that approximate maximization of the mutual information justifies the minimization of the cross entropy in the AutoEncoder (AE) approach [6], often seen in recent wireless communication literature [6,19,28].

4.5. Information Bottleneck View

It should be stressed that we have not set any constraints on the variables in the InfoMax problem so far. However, in many applications, compression is needed because of the limited information rate. Therefore, we can formulate an optimization problem where we like to maximize the relevant information $I_{\theta}(\mathbf{z};\mathbf{y})$ subject to the constraint to limit the compression rate $I_{\theta}(\mathbf{s};\mathbf{y})$ to a maximum information rate I_C :

$$\arg \max_{p_{\theta}(\mathbf{x}|\mathbf{s})} I_{\theta}(\mathbf{z};\mathbf{y}) \quad \text{s.t.} \quad I_{\theta}(\mathbf{s};\mathbf{y}) \leq I_C. \quad (15)$$

Problem (15) is an important variation of the InfoMax principle and called the Information Bottleneck (IB) problem [10,29,40,41]. The IB method introduced by Tishby et al. [29] has been the subject of intensive research for years and has proven to be a suitable mathematical/information-theoretical framework for solving numerous problems—as well as in wireless communications [30,31,42,43]. Note that we aim for an encoder that compresses \mathbf{s} into a compact representation \mathbf{x} for discrete RVs by clustering and for continuous RVs by dimensionality reduction.

To solve the constrained optimization problem (15), we can use Lagrangian optimization and obtain

$$\arg \max_{p_{\theta}(\mathbf{x}|\mathbf{s})} I_{\theta}(\mathbf{z};\mathbf{y}) - \beta I_{\theta}(\mathbf{s};\mathbf{y}) \quad (16)$$

with Lagrange multiplier $\beta \geq 0$. The Lagrange multiplier β allows the defining of a trade-off between the relevant information $I_{\theta}(\mathbf{z};\mathbf{y})$ and compression rate $I_{\theta}(\mathbf{s};\mathbf{y})$, which indicates the relation to rate distortion theory [30]. With $\beta = 0$, we have the InfoMax problem (1) whereas for $\beta \rightarrow \infty$ we minimize compression rate. Calculation of the mutual information terms may be computationally intractable, as in the InfoMax problem (1). Approximation

approaches can be found in [44,45]. Notable exceptions include if the RVs are all discrete or Gaussian distributed.

We note that in [10,26] the authors already introduced the IB problem to task-oriented communications. But [10,26] do not address our viewpoint or classification. We compress and channel encode the messages/communications source \mathbf{s} for given entailment $p(\mathbf{s}|\mathbf{z})$, in the sense of a data-reduced and reliable communication of the semantic RV \mathbf{z} . Basically, we implement joint source-channel coding of \mathbf{s} s.t. preserving the semantic RV \mathbf{z} , and we do not differentiate between Levels A and B, as indicated by Weaver’s notion outlined in Section 2. Indeed, we draw a direct connection to IB compared to related semantic communication literature [19,21,38] that, so far, only included optimization with terms reminiscent of the IB problem.

4.5.1. Semantic Information Bottleneck

This article does not only distinct itself on a conceptual, but also on a technical level from [26,34]. We follow a different strategy to solve (15).

First, using the data processing inequality [46], we see that the compression rate is upper bounded by the mutual information of the encoder $I_\theta(\mathbf{s};\mathbf{x})$ and that of the channel $I(\mathbf{x};\mathbf{y})$:

$$I_\theta(\mathbf{s};\mathbf{y}) \leq \min\{I_\theta(\mathbf{s};\mathbf{x}), I(\mathbf{x};\mathbf{y})\}. \tag{17}$$

In case of negligible encoder compression $I_\theta(\mathbf{s};\mathbf{x}) > I(\mathbf{x};\mathbf{y})$, the channel becomes the limiting factor of information rate. For example, with a deterministic continuous mapping $\mathbf{x} = \mu_\theta(\mathbf{s})$, this is true since $I_\theta(\mathbf{s};\mathbf{x}) \rightarrow \infty$. Using the chain rule of mutual information [46], we see that this upper bound on compression rate grows with the dimension of \mathbf{x} , i.e., the number of channel uses N_{Tx} :

$$I_\theta(\mathbf{s};\mathbf{y}) \leq I(\mathbf{x};\mathbf{y}) = \sum_{n=1}^{N_{Tx}} \underbrace{I(x_n;\mathbf{y}|x_{n-1}, \dots, x_1)}_{\geq 0}. \tag{18}$$

Assuming \mathbf{y} to be conditional dependent on x_n given x_{n-1}, \dots, x_1 , i.e., $p(\mathbf{y}|x_n, \dots, x_1) \neq p(\mathbf{y}|x_{n-1}, \dots, x_1)$ being, e.g., true for an AWGN channel, it is $I(x_n;\mathbf{y}|x_{n-1}, \dots, x_1) > 0$ [46] and the sum in (18) indeed strictly increases. Replacing \mathbf{y} in $I(\mathbf{x};\mathbf{y})$ of (18) by \mathbf{s} , the result also holds for encoder compression $I_\theta(\mathbf{s};\mathbf{x})$, respectively. Hence, increasing the encoder output dimension N_{Tx} , we can increase the possible compression rate $I_\theta(\mathbf{s};\mathbf{y})$. Interchanging \mathbf{x} and \mathbf{y} in (18), we see that the same holds for the receiver input dimension N_{Rx} .

Furthermore, the mutual information of the channel and, thus, the compression rate are upper bounded by channel capacity:

$$I_\theta(\mathbf{s};\mathbf{y}) \leq I(\mathbf{x};\mathbf{y}) \leq \max_{p(\mathbf{x});E[|x_n|^2] \leq 1} I(\mathbf{x};\mathbf{y}) = C. \tag{19}$$

For example, with an AWGN channel with noise standard deviation σ_n , we have $C = N_{Tx}/2 \cdot \ln(1 + 1/\sigma_n^2)$ again increasing with N_{Tx} .

Now, let us assume the RVs to be discrete so that $\mathcal{H}(\mathbf{x}|\mathbf{s}) \geq 0$. Indeed, this is true if the RVs are processed discretely with finite resolution on digital signal processors, as in the numerical example of Section 5. As long as $I_\theta(\mathbf{s};\mathbf{x}) < C$, all information of the discrete RVs can be transmitted through the channel with arbitrary low error probability according to Shannon’s channel coding theorem [1]. Then, we can upper bound encoder compression $I_\theta(\mathbf{s};\mathbf{x})$ and thus compression rate $I_\theta(\mathbf{s};\mathbf{y})$ by the sum of entropies of any output x_n [46] of the encoder $p_\theta(\mathbf{x}|\mathbf{s})$ —each with cardinality $|\mathcal{M}_x|$:

$$I_\theta(\mathbf{s};\mathbf{x}) = \mathcal{H}(\mathbf{x}) - \underbrace{\mathcal{H}(\mathbf{x}|\mathbf{s})}_{\geq 0} \leq \mathcal{H}(\mathbf{x}) \leq \sum_{n=1}^{N_{Tx}} \mathcal{H}(x_n) \leq N_{Tx} \cdot \log_2(|\mathcal{M}_x|). \tag{20}$$

Note that the entropy sum in (20) grows again with N_{Tx} for discrete RVs since $0 \leq \mathcal{H}(x_n) \leq \log_2(|\mathcal{M}_x|)$. Moreover, we can define an encoder capacity C_θ analogous to channel capacity C in (19) that upper bounds encoder compression $I_\theta(\mathbf{s}; \mathbf{x})$. It may be restricted by the chosen (DNN) model $p_\theta(\mathbf{x}|\mathbf{s})$ and optimization procedure with regard to θ , i.e., the hypothesis class [7].

In summary, we have proven by (19) and (20) that there is an information bottleneck when maximizing the relevant information $I_\theta(\mathbf{z}; \mathbf{y})$ either due to the channel distortion $I(\mathbf{x}; \mathbf{y})$ or encoder compression $I_\theta(\mathbf{s}; \mathbf{y})$.

To fully exploit the available resources, we set constraint I_C to be equal to the upper bound, i.e., channel capacity C or the upper bound on encoder compression rate $N_{\text{Tx}} \cdot \log_2(|\mathcal{M}_x|)$. In both cases, the upper bound grows (linearly) with the encoder output dimension N_{Tx} , and, thus, we can set the constraint I_C higher or lower by choosing N_{Tx} .

With fixed constraint I_C , we maximize the relevant information $I_\theta(\mathbf{z}; \mathbf{y})$. By doing so, we derive an exact solution to (15) that maximizes $I_\theta(\mathbf{z}; \mathbf{y})$ for a fixed encoder output dimension that bounds the compression rate. As in the InfoMax problem, we can exploit the MILBO to use the amortized cross entropy $\mathcal{L}_{\theta, \varphi}^{\text{CE}}$ in (9) as the optimization criterion.

4.5.2. Variational Information Bottleneck

In [26], however, the authors solve the variational IB problem of (16) and require tuning of β . Albeit also using the MILBO as a variational approximation to the first term in (16), they introduce a KL divergence term as an upper bound to compression rate $I_\theta(\mathbf{s}; \mathbf{y})$ derived by $D_{\text{KL}}(p_\theta(\mathbf{y}) \parallel q_\theta(\mathbf{y})) \geq 0$ with some variational distribution $q_\theta(\mathbf{y})$ with parameters θ [44]. Then, the variational IB objective function reads [44]:

$$I_\theta(\mathbf{z}; \mathbf{y}) - \beta I_\theta(\mathbf{s}; \mathbf{y}) \geq \mathbb{E}_{\mathbf{z}, \mathbf{y} \sim p_\theta(\mathbf{z}, \mathbf{y})}[\ln q_\theta(\mathbf{z}|\mathbf{y})] - \beta \cdot \mathbb{E}_{\mathbf{s} \sim p(\mathbf{s})}[D_{\text{KL}}(p_\theta(\mathbf{y}|\mathbf{s}) \parallel q_\theta(\mathbf{y}))]. \quad (21)$$

Moreover, the authors use a log-uniform distribution as the variational prior $q_\theta(\mathbf{y})$ in [26] to induce sparsity on \mathbf{y} so that the number of outputs is dynamically determined based on the channel condition or SNR, i.e., $p_\theta(\mathbf{y}|\mathbf{s}, \sigma_n^2)$. The approach additionally necessitates approximation of the KL divergence term in (21) and estimation of the noise variance σ_n^2 .

With our approach we avoid the additional approximations and tuning of the hyperparameter β in (21) possibly enabling better semantic performance as well as reduced inference and training complexity at the cost of full usage of N_{Tx} channels even when the channel capacity C enables its reduction. We leave a numerical comparison to [26] for future research as this is out of the scope of this paper.

4.6. Implementation Considerations

Now, we will provide important implementation considerations for optimization of (8)/(10) and (15). We note that computation of the MILBO leads to similar problems as for the ELBO [35]; if calculating the expected value in (10) cannot be solved analytically or is computationally intractable—as typically the case with DNNs—we can approximate it using Monte Carlo sampling techniques with N samples $\{\mathbf{z}_i, \mathbf{s}_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$.

For Stochastic Gradient Descent (SGD)-based optimization like, e.g., in the AE approach, the gradient with regard to φ can then be calculated by

$$\frac{\partial \mathcal{L}_{\theta, \varphi}^{\text{CE}}}{\partial \varphi} = \frac{\partial}{\partial \varphi} \mathbb{E}_{\mathbf{z}, \mathbf{s}, \mathbf{y} \sim p_\theta(\mathbf{y}|\mathbf{s})p(\mathbf{s}|\mathbf{z})p(\mathbf{z})}[-\ln q_\varphi(\mathbf{z}|\mathbf{y})] \quad (22)$$

$$= -\mathbb{E}_{\mathbf{z}, \mathbf{s}, \mathbf{y} \sim p_\theta(\mathbf{y}|\mathbf{s})p(\mathbf{s}|\mathbf{z})p(\mathbf{z})} \left[\frac{\partial \ln q_\varphi(\mathbf{z}|\mathbf{y})}{\partial \varphi} \right] \quad (23)$$

$$\approx -\frac{1}{N} \sum_{i=1}^N \frac{\partial \ln q_\varphi(\mathbf{z}_i|\mathbf{y}_i)}{\partial \varphi} \quad (24)$$

with N being equal to the batch size N_b and by application of the backpropagation algorithm to $\frac{\partial}{\partial \varphi} \ln q_{\varphi}(\mathbf{z}_i | \mathbf{y}_i) = \frac{\partial}{\partial \varphi} q_{\varphi}(\mathbf{z}_i | \mathbf{y}_i) / q_{\varphi}(\mathbf{z}_i | \mathbf{y}_i)$ in Automatic Differentiation Frameworks (ADF), e.g., TensorFlow and PyTorch. Computation of the so-called REINFORCE gradient with regard to θ leads to a high variance of the gradient estimate since we sample with regard to the distribution $p_{\theta}(\mathbf{y} | \mathbf{s})$ dependent on θ [35].

Reparametrization Trick

Leveraging the direct relationship between θ and \mathbf{y} in $\ln q_{\varphi}(\mathbf{z} | \mathbf{y})$ can help reduce the estimator's high variance. Typically, e.g., in the Variational AE (VAE) approach, the *reparametrization trick* is used to achieve this [35]. Here, we can apply it if we can decompose the latent variable $\mathbf{y} \sim p_{\theta}(\mathbf{y} | \mathbf{s})$ into a differentiable function $\mathbf{y} = f_{\theta}(\mathbf{s}, \mathbf{n})$ and a RV $\mathbf{n} \sim p(\mathbf{n})$ independent of θ . Fortunately, the typical forward model of a communication system $p_{\theta}(\mathbf{y} | \mathbf{s})$ fulfills this criterion. Assuming a deterministic DNN encoder $\mathbf{x} = \mu_{\theta}(\mathbf{s})$ and additive noise \mathbf{n} with covariance Σ , we can thus rewrite \mathbf{y} into $f_{\theta}(\mathbf{s}, \mathbf{n}) = \mu_{\theta}(\mathbf{s}) + \Sigma^{1/2} \cdot \mathbf{n}$ and, accordingly, the amortized cross entropy gradient into:

$$\frac{\partial \mathcal{L}_{\theta, \varphi}^{\text{CE}}}{\partial \theta} = - \frac{\partial}{\partial \theta} \mathbb{E}_{\mathbf{z}, \mathbf{s}, \mathbf{y} \sim p_{\theta}(\mathbf{y} | \mathbf{s}) p(\mathbf{s}, \mathbf{z})} [\ln q_{\varphi}(\mathbf{z} | \mathbf{y})] \quad (25)$$

$$= - \mathbb{E}_{\mathbf{z}, \mathbf{s}, \mathbf{n} \sim p(\mathbf{n}) p(\mathbf{s} | \mathbf{z}) p(\mathbf{z})} \left[\frac{\partial f_{\theta}(\mathbf{s}, \mathbf{n})}{\partial \theta} \cdot \frac{\partial \ln q_{\varphi}(\mathbf{z} | \mathbf{y})}{\partial \mathbf{y}} \right] \quad (26)$$

$$\approx - \frac{1}{N} \sum_{i=1}^N \frac{\partial f_{\theta}(\mathbf{s}_i, \mathbf{n}_i)}{\partial \theta} \cdot \frac{\partial \ln q_{\varphi}(\mathbf{z}_i | \mathbf{y}_i)}{\partial \mathbf{y}} \Bigg|_{\mathbf{y} = f_{\theta}(\mathbf{s}_i, \mathbf{n}_i)}. \quad (27)$$

The reparametrization trick can be easily implemented in ADFs by adding a noise layer—typically used for regularization in ML literature—after (DNN) function $\mathbf{x} = \mu_{\theta}(\mathbf{s})$. Then, our loss function (10) amounts to

$$\mathcal{L}_{\theta, \varphi}^{\text{CE}} \approx - \frac{1}{N} \sum_{i=1}^N \ln q_{\varphi}(\mathbf{z}_i | \mathbf{y}_i = f_{\theta}(\mathbf{s}_i, \mathbf{n}_i)). \quad (28)$$

This enables the joint optimization of both θ and φ , as demonstrated in recent works [6], treating unsupervised optimization of AEs as a supervised learning problem.

5. Example of Semantic Information Recovery

In this section, we provide one numerical example of data-driven semantics to explain what we understand under a semantic communication design and to show its benefits: It is the task of image classification. In fact, we consider our example of the biologist from Section 4.2 who wants to know what type of tree it is.

For the remainder of this article, we will thus assume the hidden semantic RV to be a one-hot vector $\mathbf{z} \in \{0, 1\}^{M \times 1}$ where all elements are zero except for one element representing one of the M image classes. Then, the semantic channel $p(\mathbf{s} | \mathbf{z})$ (see Figure 1) generates images belonging to this class, i.e., the source signal \mathbf{s} .

Note that for point-to-point transmission, as in [26], we could first classify the image based on the posterior $q_{\varphi}(\mathbf{z} | \mathbf{s})$, as shown in Figure 2 and transmit the estimate $\hat{\mathbf{z}}$ (encoded into \mathbf{x}) through the physical channel since this would be most rate or bandwidth efficient.

But if the image information is distributed across multiple agents, all (sub) images may contribute useful information for classification. We could thus lose information when making hard decisions on each transmitter's side. In the distributed setting, transmission and combination of features, i.e., soft information, is crucial to obtain high classification accuracy.

Further, we note that transmission of full information, i.e., raw image data \mathbf{s} , through a wireless channel from each agent to a central unit for full image classification would consume a lot of bandwidth. This case is also shown in Figure 2 assuming perfect com-

munication links between the output of the semantic channel and the input of the ResNet Feature Extractor.

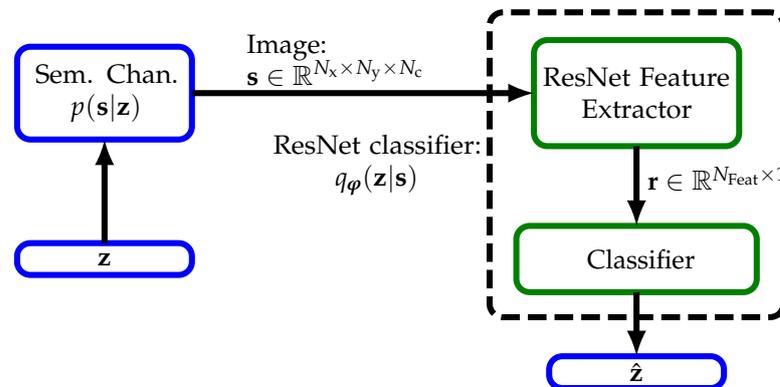


Figure 2. Central image processing: Based on the images, ResNet extracts semantics by classification.

Therefore, we investigate a distributed setting shown in Figure 3. There, each of four agents sees its own image $\mathbf{s}_1, \dots, \mathbf{s}_4 \sim p(\mathbf{s}_i|\mathbf{z})$ being generated by the same semantic RV \mathbf{z} . Based on these images, a central unit shall extract semantics, i.e., perform classification. We propose to optimize the four encoders $p_{\theta_i}(x_i|\mathbf{s}_i)$ with $i = 1, \dots, 4$, each consisting of a bandwidth efficient feature extractor (ResNet Feature Extractor i) and transmitter (Tx i) **jointly** with a decoder $q_{\varphi}(\mathbf{z}|\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4]^T)$, consisting of a receiver (Rx) and concluding classifier (Classifier), with regard to cross entropy (10) of the semantic labels (see Figure 3). Hence, we maximize the system's overall semantic measure, i.e., classification accuracy. Note that this scenario is different from both [33,34]; we include a physical communication channel (Comm. Channel i) since we aim to transmit and not only compress. For the sake of simplicity, we assume orthogonal channel access. The IB is addressed by limiting the number of channel uses, which defines the constraint I_C in (15).

As a first demonstration example, we use the grayscale MNIST and colored CIFAR10 datasets with $M = 10$ image classes [47]. We assume that the semantic channel generates an image that we divide into four equally sized quadrants and each agent observes one quadrant $\mathbf{s}_1, \dots, \mathbf{s}_4 \in \mathbb{R}^{N_x \times N_y \times N_c}$, where N_x and N_y is the number of image pixels in the x- and y-dimension, respectively, and N_c is the number of color channels. Albeit this does not resemble a realistic scenario, note that we can still show the basic working principle and ease implementation.

5.1. ResNet

For the design of the overall system, we rely on a famous DNN approach for feature extraction, breaking records at the time of invention: ResNet [47,48]. The key idea of ResNet is that it consists of multiple residual units. Each unit's input is fed directly to its output and if the dimensions do not match, a convolutional layer is used. This structure enables fast training and convergence of DNNs since the training error can be backpropagated to early layers through these skip connections. From a mathematical point of view, usual DNNs have the design flaw that using a larger function class, i.e., more DNN layers, does not necessarily increase the expressive power. However, this holds for nested functions like ResNet which contain the smaller classes of early layers.

Each residual unit itself consists of two Convolutional NNs (CNNs) with subsequent batch normalization and ReLU activation function, i.e., $\rho_1(\cdot) = \max(\cdot, 0)$, to extract translation invariant and local features across two spatial dimensions N_x and N_y . Color channels, like in CIFAR10, add a third dimension $N_c = 3$ and additional information. The idea behind stacking multiple layers of CNNs is that features tend to become more abstract from early layers (e.g., edges and circles) to final layers (e.g., beaks or tires).

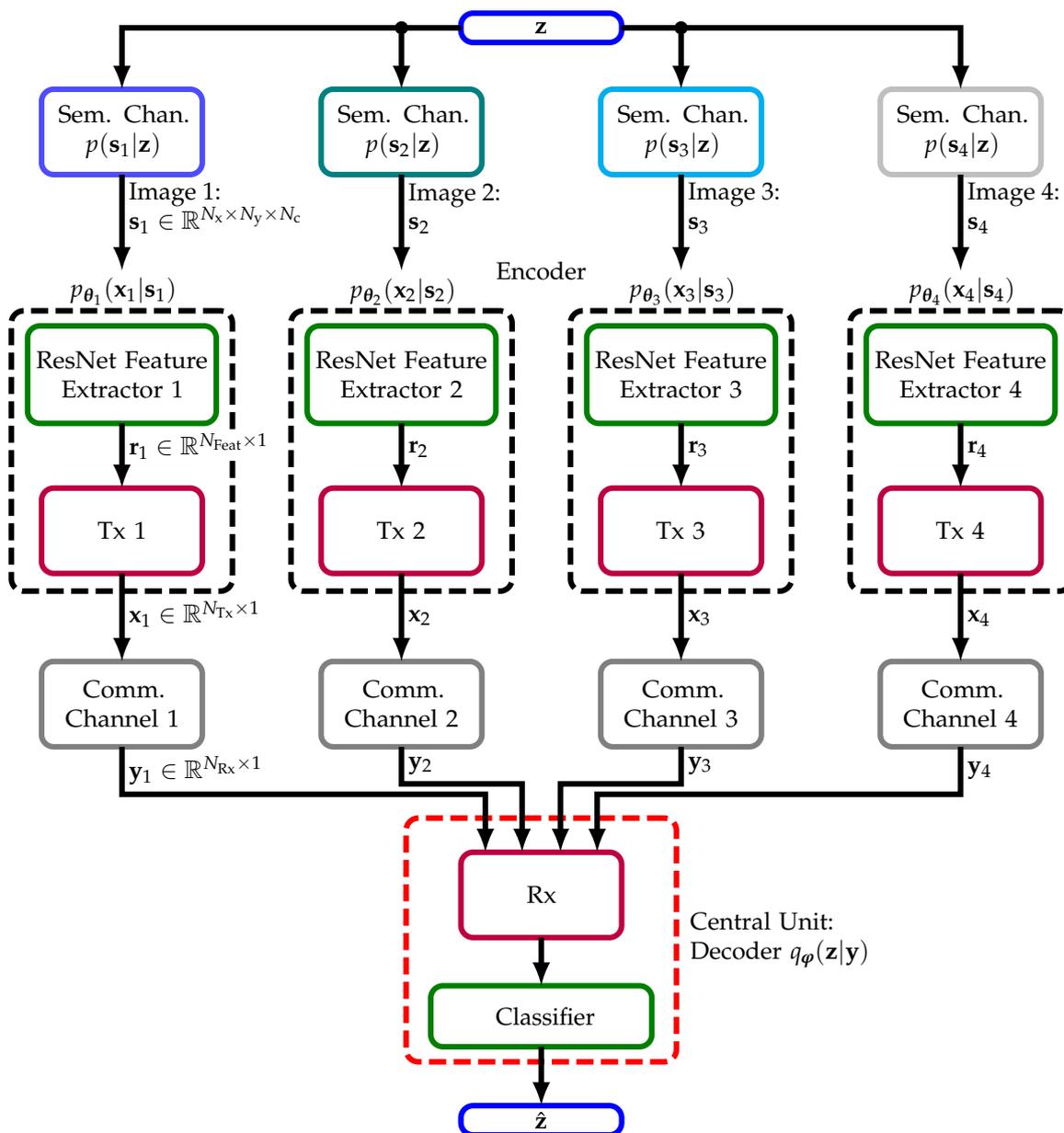


Figure 3. Semantic INFORMATION transmission and recoverY (SINFONY) for distributed agents. Each agent extracts features for bandwidth-efficient transmission. Based on the received signal, the central unit extracts semantics by classification.

In this work, we use the preactivation version of ResNet without bottlenecks from [47,48] implemented for classification on the dataset CIFAR10. In Table 1, we show its structure for the distributed scenario from Figure 3. There, ResNetBlock is the basic building block of the ResNet architecture. Each block consists of multiple residual units (res. un.) and we use 2 for the MNIST dataset and 3 for the CIFAR10 dataset, which means we use ResNet14 and ResNet20, respectively. We arrive at the architecture of central image processing from Figure 2 by removing the components Tx, (physical) Channel, and Rx and increasing each spatial dimension by 2 to contain all quadrants of the original image. For further implementation details, we refer the reader to the original work [48].

Table 1. Semantic INFOrmation traNsmission and recoverY (SINFONY)—DNN architecture for distributed image classification.

Component	Layer	Dimension
Input	Image (MNIST, CIFAR10)	(14, 14, 1), (16, 16, 3)
4×	Conv2D	(14, 14, 14), (16, 16, 16)
Feature	ResNetBlock (2/3 res. un.)	(14, 14, 14), (16, 16, 16)
Extractor	ResNetBlock (2/3 res. un.)	(7, 7, 28), (8, 8, 32)
	ResNetBlock (2/3 res. un.)	(4, 4, 56), (4, 4, 64)
	Batch Normalization	(4, 4, 56), (4, 4, 64)
	ReLU activation	(4, 4, 56), (4, 4, 64)
	GlobalAvgPool2D	(56), (64)
4× Tx	ReLU	N_{Tx}
	Linear	N_{Tx}
	Normalization (dim.)	N_{Tx}
4× Channel	AWGN	N_{Tx}
Rx	ReLU (4× shared)	(2, 2, N_w)
	GlobalAvgPool2D	N_w
Classifier	Softmax	$M = 10$

5.2. Distributed Semantic Communication Design Approach

Our key idea here is to modify ResNet with regard to the communication task by splitting it at a suitable point where a representation $\mathbf{r} \in \mathbb{R}^{N_{Feat} \times 1}$ of semantic information with low-bandwidth is present (see Figures 2 and 3). ResNet and CNNs in general can be interpreted to extract features; with full images, we obtain a feature map of size $8 \times 8 \times N_{Feat}$ after the last ReLU activation (see Table 1). These local features are aggregated by the global average pooling layers across the 2 spatial dimensions into \mathbf{r} . Based on these N_{Feat} global features in \mathbf{r} , the softmax layer finally classifies the image. We note that the features contain the relevant information with regard to the semantic RV \mathbf{z} and are of low dimension compared to the original image or even its sub-images, i.e., 64 compared to $16 \times 16 \times 3 = 768$ for CIFAR10.

Therefore, we aim to transmit each agent's local features $\mathbf{r}_i \in \mathbb{R}^{N_{Feat} \times 1}$ ($i = 1, \dots, 4$) instead of all sub-images \mathbf{s}_i and add the component Tx in Table 1 to encode the features \mathbf{r}_i into $\mathbf{x}_i \in \mathbb{R}^{N_{Tx} \times 1}$ for transmission through the wireless channel (see Figure 3). We note that $\mathbf{x}_i \in \mathbb{R}^{N_{Tx} \times 1}$ is analog and that the output dimension N_{Tx} of \mathbf{x}_i defines the number of channel uses per agent/image. Note that the less often we use the wireless channel (N_{Tx}), the less information we transmit but the less bandwidth we consume, and vice versa. Hence, the number of channel uses defines the IB in (15). We implement the Tx module by DNN layers. To limit the transmission power to one, we constrain the Tx output by the norm along the training batch or the encoding vector dimension (dim.), i.e., $x_n = \tilde{x}_n / \sqrt{E[\tilde{x}_n^2]}$ or $\mathbf{x}_i = \sqrt{N_{Tx}} \cdot \tilde{\mathbf{x}}_i / \|\tilde{\mathbf{x}}_i\|_2$ where $\tilde{\mathbf{x}}_i \in \mathbb{R}^{N_{Tx} \times 1}$ is the output of the layer Linear from Table 1. For numerical simulations, we choose all Tx layers to have width N_{Tx} .

At the receiver side, we use a single Rx module only with shared DNN layers and parameters φ_{Rx} for all inputs \mathbf{y}_i . This setting would be optimal if any feature is reflected in any sub-image and if the statistics of the physical channels are the same. Exploiting the prior knowledge of location-invariant features and assuming Additive White Gaussian Noise (AWGN) channels, this design choice seems reasonable. In our experiments, all layers of the Rx module have width N_w . A larger layer width N_w is equivalent to more computing power.

The output of the Rx module can be interpreted as a representation of the image features \mathbf{r}_i with index i indicating the spatial location. Thus, we have a representation of a feature map of size $(2, 2, N_w)$ that we aggregate across the spatial dimension according to the ResNet structure. Based on this semantic representation, a softmax layer with 10 units finally computes class probabilities $q_\varphi(\mathbf{z}|\mathbf{y})$ whose maximum is the maximum a posteriori estimate \hat{z} . In the following, we name our proposed approach Semantic INFORMATION traNsmission and recoverY (SINFONY).

5.3. Optimization Details

We evaluate SINFONY in TensorFlow 2 [49] on the MNIST and CIFAR10 datasets. The source code is available in [50] and the default simulation and training parameters are summarized in Table 2. We split the dataset into $N_{\text{train}} = 60$ k/50 k training data and 10 k validation data samples, respectively. For preprocessing, we normalize the pixel inputs to range $[0, 1]$, but we do not use data augmentation, in contrast to [47,48], yielding slightly worse accuracy. The ReLU layers are initialized with uniform distribution according to He and all other layers according to Glorot [51].

Table 2. Default simulation and training parameters.

Parameter Name	Variable	Value (MNIST, CIFAR10)
Batch size	N_b	64
Epoch number	N_e	20, 200
Learning rate	ϵ	Schedule: $\epsilon = \{0.1, 0.01, 0.001\}$ with $N_e = \{3, 6\}, \{100, 150\}$
Optimizer		SGD with momentum= 0.9
Preprocessing		Input normalization to $[0, 1]$
Training SNR range	$\text{SNR}_{\text{train}}$	$[-4, 6]$ dB
Training dataset size	N_{train}	60 k, 50 k
Validation dataset size		10 k
Weight decay		0.0001
Weight initialization		Glorot uniform, ReLU: He uniform
Encoder normalization	dim.	Batch dimension
Rx layer width	N_w	56, 64

In the case of CIFAR10 classification with central image processing and original ResNet, we need to train $N_\theta + N_\varphi = 273,066$ parameters. We like to stress that although we divided the image input into four smaller pieces, this number grows more than four times to $4N_\theta + N_\varphi = 1,127,754$ with $N_{\text{Tx}} = N_{\text{Feat}} = 64$ for SINFONY. The reason lies in the ResNet structure with minor dependence on the input image size and that we process at four agents with an additional Tx module. Only $N_\varphi = 4810$ parameters amount to the Rx module and classification, i.e., the central unit. We note that the number of added Tx and Rx parameters of 33,560 and 3192 is relatively small. Since the number of parameters only weakly grows with Rx layer width N_w in our design, we choose $N_w = N_{\text{Feat}}$ as the default.

For optimization of the cross entropy (10) or the loss function (28), we use the reparametrization trick from Section 4.6 and SGD with a momentum of 0.9 and a batch size of $N_b = 64$. We add l_2 -regularization with a weight decay of 0.0001 as in [47,48]. The learning rate of $\epsilon = 0.1$ is reduced to 0.01 and 0.001 after $N_e = 100$ and 150 epochs for CIFAR10 and after 3 and 6 epochs for MNIST. In total, we train for $N_e = 200$ epochs with CIFAR10 and for 20 with MNIST. In order to optimize the transceiver for a wider SNR range, we choose the training SNR to be uniformly distributed within $\text{SNR}_{\text{train}} \in [-4, 6]$ dB where $\text{SNR} = 1/\sigma_n^2$ with noise variance σ_n^2 .

5.4. Numerical Results and Discussion

In the following, we will investigate the influence of specific design choices on our semantic approach SINFONY. Then, we compare a semantic with a classical Shannon-based transmission approach. The design choices are as follows:

- **Central:** Central and joint processing of full image information by the ResNet classifier, see Figure 2. It indicates the maximum achievable accuracy.
- **SINFONY - Perfect com.:** The proposed distributed design SINFONY trained with perfect communication links and without channel encoding, i.e., Tx and Rx module, but with Tx normalization layer. Thus, the plain and power-constrained features are transmitted with $N_{Tx} = N_{Feat}$ channel uses. It serves as the benchmark since it indicates the maximum performance of the distributed design.
- **SINFONY - AWGN:** SINFONY Perfect com. evaluated with AWGN channel.
- **SINFONY - AWGN + training:** SINFONY Perfect com. trained with AWGN channel.
- **SINFONY - Tx/Rx ($N_{Tx} = N_{Feat}$):** SINFONY trained with channel encoding, i.e., Tx and Rx module, and $N_{Tx} = N_{Feat}$ channel uses.
- **SINFONY - Tx/Rx ($N_{Tx} < N_{Feat}$):** SINFONY trained with channel encoding and $N_{Tx} < N_{Feat}$ channel uses for feature compression.
- **SINFONY - Classic digital com.:** SINFONY - Perfect com. with classic digital communications (Huffman coding, LDPC coding with belief propagation decoding, and digital modulation) as additional Tx and Rx modules. For details, see Section 5.4.4.
- **SINFONY - Analog semantic AE:** SINFONY - Perfect com. with ML-based analog communications (AE with regard to \mathbf{r}) as additional Tx and Rx modules. It is basically the semantic communication approach from [19,21,28,32]. For details, see Section 5.4.5.

Since meaning is expressed by the RV \mathbf{z} , we use classification accuracy to measure semantic transmission quality. For illustration in logarithmic scale, we show the opposite of accuracy in all plots, i.e., classification error rate.

5.4.1. MNIST Dataset

The numerical results of our proposed approach SINFONY on the MNIST validation dataset are shown in Figure 4 for $N_w = 56$. To obtain a fair comparison between transmit signals $\mathbf{x}_i \in \mathbb{R}^{N_{Tx} \times 1}$ of different length N_{Tx} , we normalize the SNR by the spectral efficiency or rate $\eta = N_{Feat}/N_{Tx}$. First, we observe that the classification error rate of 0.5% of the central ResNet unit with full image information (Central) is smaller than that of 0.9% of SINFONY - Perfect com. Note that we assume ideal communication links. However, the difference seems negligible considering that the local agents only see a quarter of the full images and learn features independently based on it.

With noisy communication links (SINFONY - AWGN), the performance degrades especially for $\text{SNR} < 10$ dB, and we can avoid degradation just partly by training with noise (SINFONY - AWGN + training). Introducing the Tx module (SINFONY - Tx/Rx $N_{Tx} = 56$), we further improve classification accuracy at low SNR. If we encode the features from $N_{Feat} = 56$ to only $N_{Tx} = 14$ in the Tx module (SINFONY - Tx/Rx $N_{Tx} = 14$) to have less channel uses/bandwidth (stronger bottleneck), the error rate is lowest compared to other SINFONY examples with non-ideal links for low normalized SNR. At high SNR, we observe a small error offset, which indicates lossy compression. In fact, our system SINFONY learns a reliable semantic encoding to improve the classification performance of the overall system with non-ideal links. Every design choice in Table 1 is well-motivated.

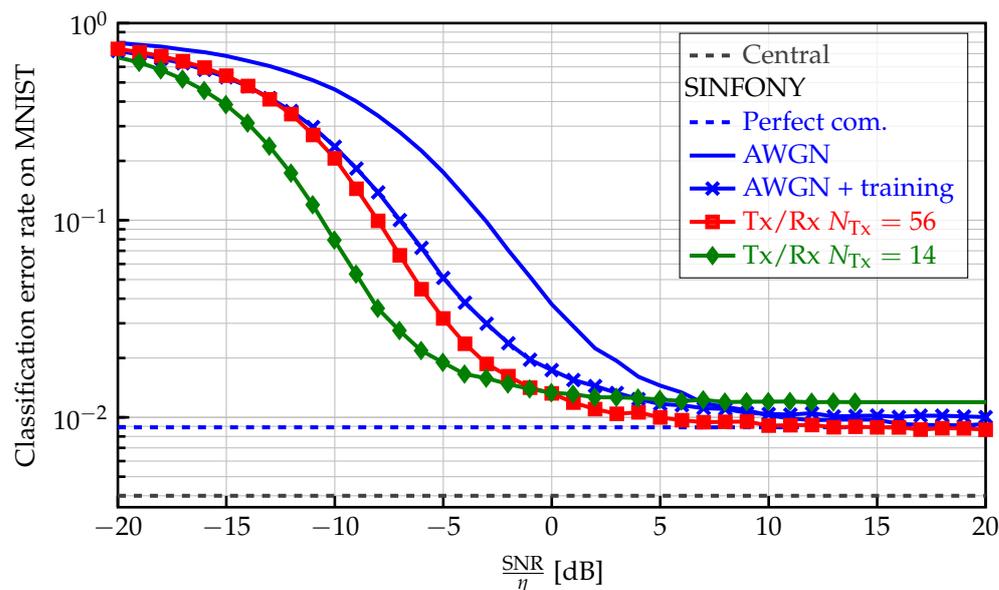


Figure 4. Classification error rate of different SINFONY examples (distributed setting) and central image processing on the MNIST validation dataset as a function of normalized SNR.

5.4.2. CIFAR10 Dataset

Comparing these results to the classification accuracy on CIFAR10 shown in Figure 5, we observe a similar behavior. But a few main differences become apparent. Central performs much better with a 12% error rate than SINFONY - Perfect com. with 20%. We expect the reason to lie in the more challenging dataset with more color channels. Further, SINFONY - AWGN + training with $N_{Tx} = N_{Feat} = 64$ channel uses runs into a rather high error floor. Notably, even SINFONY - Tx/Rx ($N_{Tx} = 16$) with fewer channel uses performs better than both SINFONY - AWGN and SINFONY - AWGN + training over the whole SNR range and achieves channel encoding with negligible loss. This means adding more flexible channel encoding, i.e., Tx/Rx module, is crucial for CIFAR10.

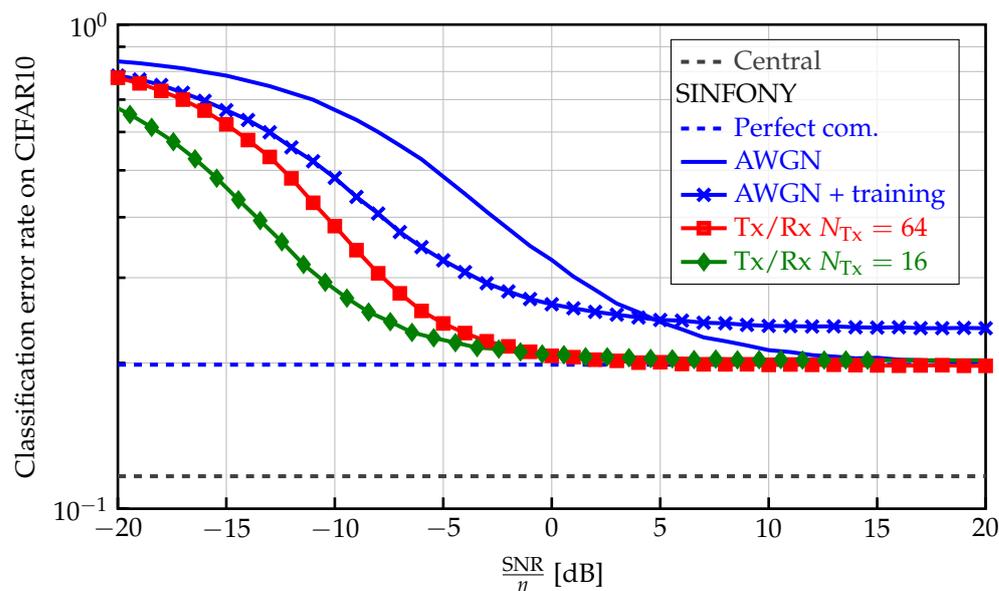


Figure 5. Classification error rate of different SINFONY examples (distributed setting) and central image processing on CIFAR10 as a function of normalized SNR.

5.4.3. Channel Uses Constraint

Since one of the main advantages of semantic communication lies in savings of information rate, we also investigate the influence of the number of channel uses N_{Tx} on MNIST classification error rate shown in Figure 6. From a practical point of view, we fix the information bottleneck by the output dimension N_{Tx} and maximize the mutual information $I_{\theta}(\mathbf{z}; \mathbf{y})$. Decreasing the number of channel uses from $N_{Tx} = 14$ to 2 and accordingly the upper bound I_C on the mutual information $I_{\theta}(\mathbf{s}; \mathbf{y})$, i.e., compression rate, from (19) or (20), we observe that the error floor at high SNR increases. We assume that, since the channel capacity decreases with SNR and N_{Tx} , higher compression is required for reliable transmission through the channel in the training SNR interval. For $N_{Tx} = 56$, almost no error floor occurs at the cost of a smaller channel encoding gain. This means compression and channel coding are balanced based on the channel condition, i.e., training SNR region, to find the optimal trade-off to maximize $I_{\theta}(\mathbf{z}; \mathbf{y})$, which we can also observe in unshown simulations.

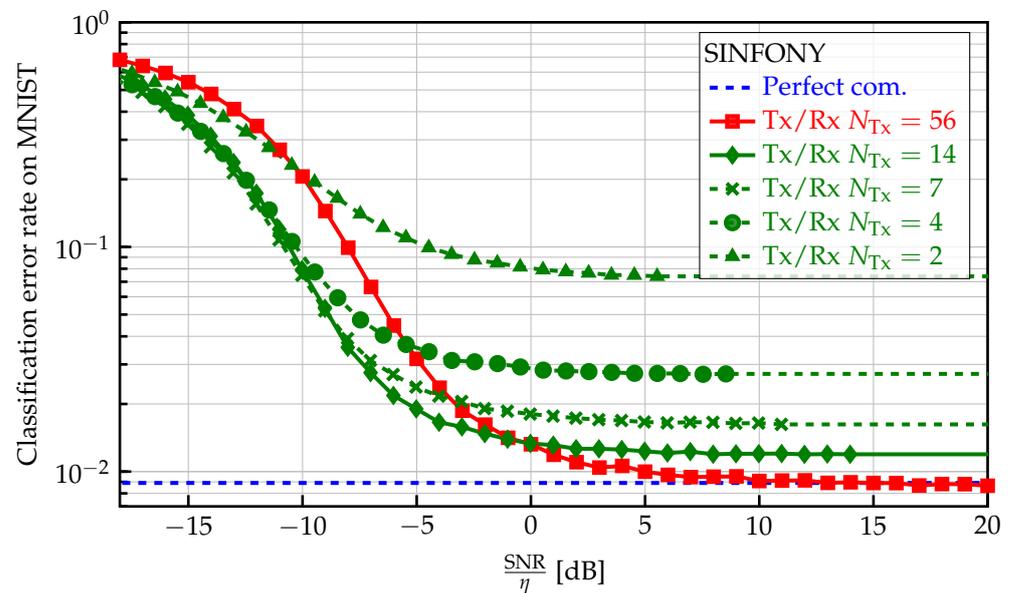


Figure 6. Classification error rate of SINFONY on the MNIST validation dataset for different rate/channel uses constraints as a function of normalized SNR.

5.4.4. Semantic vs. Classic Design

Finally, we compare semantic and classic communication system designs. For the classic digital design, we first assume that the images are compressed lossless and protected by a channel code for transmission and reliable overall image classification by the central unit based on $q_{\varphi}(\mathbf{z}|\mathbf{s})$ (Central image com.). We apply Huffman encoding to a block containing 100 images \mathbf{s}_i where each RGB color entry contains 8 bits.

For fairness, we also compare to a SINFONY version where Tx and Rx modules of Table 1 are replaced by a classic design (SINFONY - Classic digital com.). We first compress each element of the feature vector \mathbf{r}_i that is computed in 32-bit floating-point precision in the distributed setting SINFONY - AWGN to 16-bit. Then, we apply Huffman encoding to a block containing 100 feature vectors of length N_{Feat} .

Further, we use a 5G LDPC channel code implementation from [52] with interleaver, rate $R_C = \{0.25, 0.5, 0.25\}$ and long block length of $\{15360, 16000, 15360\}$, and modulate the code bits with {BPSK, BPSK, 16-QAM} such that we have, e.g., parameter set $\{0.25, 15360, \text{BPSK}\}$ in one simulation. For digital image transmission, we use a rate of $R_C = 0.25$ with a block length of 15360 and BPSK modulation. At the receiver, we assume belief propagation decoding, where the noise variance is perfectly known for LLR computation.

The results in Figure 7 reveal tremendous information rate savings for the semantic design with SINFONY. We observe an enormous SNR shift of roughly 20 dB compared to the classic digital design with regard to both image (Central image com.) and feature transmission (SINFONY - Classic digital com.). Note that the classic design is already near the Shannon limit and even if we improve it by ML we are only able to shift its curve by a few dB. The reason may lie in overall system optimization with SINFONY with regard to semantics and analog encoding of x .

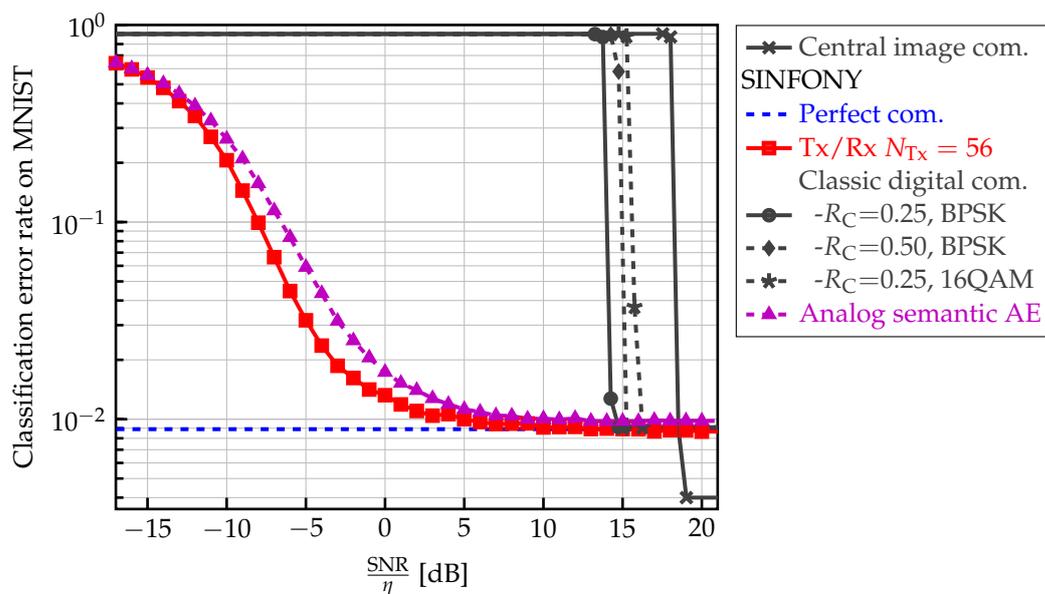


Figure 7. Classification error rate of SINFONY with different kinds of optimized Tx/Rx modules and central image processing with digital image transmission on the MNIST validation dataset as a function of normalized SNR.

5.4.5. SINFONY vs. Analog “Semantic” Autoencoder

To distinguish both influences, we also implemented the approach of (14) according to Shannon by analog AEs. The analog AE has been introduced by O’Shea and Hoydis in [6]. From the viewpoint of semantic communication, it resembles the semantic approach from [19,21,28,32] without differentiating between semantic and channel coding, and the mutual information constraint $I(x; y)$ like in [21]. We trained the AE matching the Tx and Rx module in Table 1 with mean square error criterion for reliable transmission of the feature vector \mathbf{r} with SINFONY training settings. The Rx module consists of one ReLU layer of width $N_w = N_{Tx}$ providing the estimate of \mathbf{r} . We provide results (SINFONY - Analog semantic AE) in Figure 7. Indeed, most of the shift is due to analog encoding. By this means, we further avoid the typical thresholding behavior of a classic digital system seen at 14 dB.

In conclusion, this surprisingly clear result justifies an analog “semantic” communications design and shows its huge potential to provide bandwidth savings. However, introducing the semantic RV \mathbf{z} by SINFONY, we can further shift the curve by 2 dB and avoid a slightly higher error floor compared to the analog “semantic” AE. With expect a larger performance gap with more challenging image datasets, such as CIFAR10. More importantly, the main benefit of SINFONY lies in lower training complexity. We avoid separate and possibly iterative semantic and communication training procedures where in the first step we need to train SINFONY with ideal links hard to achieve in practice.

6. Conclusions

Motivated by the approach of Bao, Basu et al. [16,17] and inspired by Weaver's notion of semantic communication [2], we brought the terminus of a semantic source to the context of communications by considering its complete Markov chain. We defined the task of semantic communication in the sense of a data-reduced and reliable transmission of communications sources/messages over a communication channel such that the semantic Random Variable (RV) at a recipient is best preserved. We formulated its design either as an Information Maximization or as an Information Bottleneck optimization problem covering important implementations aspects like the reparametrization trick and solved the problems approximately by minimizing the cross entropy that upper bounds the negative mutual information. With this article, we distinguish from related literature [16,17,21,26,32] in both classification and perspective of semantic communication and a different ML-based solution approach.

Finally, we proposed the ML-based semantic communication system SINFONY for a distributed multipoint scenario: SINFONY communicates the meaning behind multiple messages that are observed at different senders to a single receiver for semantic recovery. We analyzed SINFONY by processing images as an example of messages. Notably, numerical results reveal a tremendous rate-normalized SNR shift up to 20 dB compared to classically designed communication systems.

Outlook

In this work, we contributed to the theoretical problem description of semantic communication and data-based ML solution approaches with DNNs. There remain open research questions such as:

- **Numerical Comparison to Variational IB:** It remains unclear if solving the variational IB problem (21) holds benefits compared to our proposed approach.
- **Implementation:** Optimization with the reparametrization trick requires a known differential channel model and training at one location with dedicated hardware such as graphics processing units [53]. In addition, large amounts of labeled data are required with data-driven ML techniques, which can be expensive and time-consuming to acquire and process. Hence, further research is required to clarify how a semantic design can be implemented efficiently in practice.
- **Semantic Modeling:** Developing effective models of semantics is crucial, and thus we proposed the usage of probabilistic models. If the underlying problem can be described by a well-known model, e.g., a physical process to be measured and processed by a sensor network [32], a promising idea is to apply model-based approaches based on Bayesian inference for encoding and decoding—potentially combined with the technique of deep unfolding. In the context of NLP, design of knowledge graphs such as ontologies or taxonomies is a promising modeling approach for human language.
- **Inconsistent Knowledge Bases:** We assumed that sender and recipient share the same background knowledge base: How does performance deteriorate if there is a mismatch and how to deal with this problem [27]?

Author Contributions: Conceptualization, E.B., C.B. and A.D.; methodology, E.B., C.B. and A.D.; software, E.B.; validation, E.B., C.B. and A.D.; formal analysis, E.B.; investigation, E.B.; resources, E.B.; data curation, E.B.; writing—original draft preparation, E.B.; writing—review and editing, E.B., C.B. and A.D.; visualization, E.B.; supervision, C.B. and A.D.; project administration, C.B. and A.D.; funding acquisition, E.B., C.B. and A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly funded by the Federal State of Bremen and the University of Bremen as part of the Human on Mars Initiative, and by the German Ministry of Education and Research (BMBF) under grant 16KISK016 (Open6GHub).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in [50].

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADF	Automatic Differentiation Framework
AE	Autoencoder
AI	Artificial Intelligence
AWGN	Additive White Gaussian Noise
CNN	Convolutional Neural Network
dim.	Dimension
DNN	Deep Neural Network
ELBO	Evidence Lower Bound
IB	Information Bottleneck
InfoMax	Information Maximization
JSCC	Joint Source-Channel Coding
KL	Kullback–Leibler
MILBO	Mutual Information Lower Bound
MI	Mutual Information
ML	Machine Learning
NLP	Natural Language Processing
pdf	Probability Density Function
pmf	Probability Mass Function
ReLU	Rectified Linear Unit
ResNet	Residual Network
res. un.	Residual Unit
RV	Random Variable
Rx	Receiver
SGD	Stochastic Gradient Descent
SINFONY	Semantic INFORMATION traNsmission and recoverY
Tx	Transmitter
VAE	Variational Autoencoder

References

- Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
- Weaver, W. Recent Contributions to the Mathematical Theory of Communication. In *The Mathematical Theory of Communication*; The University of Illinois Press: Urbana, IL, USA, 1949; Volume 10, pp. 261–281.
- Xu, W.; Yang, Z.; Ng, D.W.K.; Levorato, M.; Eldar, Y.C.; Debbah, M. Edge Learning for B5G Networks with Distributed Signal Processing: Semantic Communication, Edge Computing, and Wireless Sensing. *IEEE J. Sel. Top. Signal Process.* **2023**, *17*, 9–39. [[CrossRef](#)]
- Gastpar, M.; Rimoldi, B.; Vetterli, M. To code, or not to code: Lossy source-channel communication revisited. *IEEE Trans. Inf. Theory* **2003**, *49*, 1147–1158. [[CrossRef](#)]
- Gastpar, M.; Vetterli, M. Source-Channel Communication in Sensor Networks. In *Information Processing in Sensor Networks*; Goos, G., Hartmanis, J., van Leeuwen, J., Zhao, F., Guibas, L., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; Volume 2634, pp. 162–177. [[CrossRef](#)]
- O’Shea, T.; Hoydis, J. An Introduction to Deep Learning for the Physical Layer. *IEEE Trans. Cogn. Commun. Netw.* **2017**, *3*, 563–575. [[CrossRef](#)]
- Simeone, O. A Very Brief Introduction to Machine Learning with Applications to Communication Systems. *IEEE Trans. Cogn. Commun. Netw.* **2018**, *4*, 648–664. [[CrossRef](#)]
- Beck, E.; Bockelmann, C.; Dekorsy, A. CMDNet: Learning a Probabilistic Relaxation of Discrete Variables for Soft Detection with Low Complexity. *IEEE Trans. Commun.* **2021**, *69*, 8214–8227. [[CrossRef](#)]
- Popovski, P.; Simeone, O.; Boccardi, F.; Gündüz, D.; Sahin, O. Semantic-Effectiveness Filtering and Control for Post-5G Wireless Connectivity. *J. Indian Inst. Sci.* **2020**, *100*, 435–443. [[CrossRef](#)]
- Strinati, E.C.; Barbarossa, S. 6G networks: Beyond Shannon towards semantic and goal-oriented communications. *Comput. Netw.* **2021**, *190*, 107930. [[CrossRef](#)]

11. Lan, Q.; Wen, D.; Zhang, Z.; Zeng, Q.; Chen, X.; Popovski, P.; Huang, K. What is Semantic Communication? A View on Conveying Meaning in the Era of Machine Intelligence. *J. Commun. Inf. Netw.* **2021**, *6*, 336–371. [[CrossRef](#)]
12. Uysal, E.; Kaya, O.; Ephremides, A.; Gross, J.; Codreanu, M.; Popovski, P.; Assaad, M.; Liva, G.; Munari, A.; Soret, B.; et al. Semantic Communications in Networked Systems: A Data Significance Perspective. *IEEE/ACM Trans. Netw.* **2022**, *36*, 233–240. [[CrossRef](#)]
13. Gündüz, D.; Qin, Z.; Aguerri, I.E.; Dhillon, H.S.; Yang, Z.; Yener, A.; Wong, K.K.; Chae, C.B. Beyond Transmitting Bits: Context, Semantics, and Task-Oriented Communications. *IEEE J. Sel. Areas Commun.* **2023**, *41*, 5–41. [[CrossRef](#)]
14. Floridi, L. Philosophical Conceptions of Information. In *Formal Theories of Information: From Shannon to Semantic Information Theory and General Concepts of Information*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2009; pp. 13–53. [[CrossRef](#)]
15. Hofkirchner, W. *Emergent Information: A Unified Theory of Information Framework*; World Scientific: Singapore, 2013; Volume 3.
16. Bao, J.; Basu, P.; Dean, M.; Partridge, C.; Swami, A.; Leland, W.; Hendler, J.A. Towards a theory of semantic communication. In Proceedings of the 2011 IEEE Network Science Workshop (NSW), West Point, NY, USA, 22–24 June 2011; pp. 110–117. [[CrossRef](#)]
17. Basu, P.; Bao, J.; Dean, M.; Hendler, J. Preserving quality of information by using semantic relationships. *Pervasive Mob. Comput.* **2014**, *11*, 188–202. [[CrossRef](#)]
18. Güler, B.; Yener, A.; Swami, A. The Semantic Communication Game. *IEEE Trans. Cogn. Commun. Netw.* **2018**, *4*, 787–802. [[CrossRef](#)]
19. Farsad, N.; Rao, M.; Goldsmith, A. Deep Learning for Joint Source-Channel Coding of Text. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2326–2330. [[CrossRef](#)]
20. Xie, H.; Qin, Z.; Li, G.Y.; Juang, B.H. Deep Learning based Semantic Communications: An Initial Investigation. In Proceedings of the 2020 IEEE Global Communications Conference (GLOBECOM), Taipei, Taiwan, 7–11 December 2020; pp. 1–6. [[CrossRef](#)]
21. Xie, H.; Qin, Z.; Li, G.Y.; Juang, B.H. Deep Learning Enabled Semantic Communication Systems. *IEEE Trans. Signal Process.* **2021**, *69*, 2663–2675. [[CrossRef](#)]
22. Weng, Z.; Qin, Z.; Li, G.Y. Semantic Communications for Speech Signals. In Proceedings of the 2021 IEEE International Conference on Communications (ICC), Virtual Conference, 14–23 June 2021; pp. 1–6. [[CrossRef](#)]
23. Weng, Z.; Qin, Z.; Tao, X.; Pan, C.; Liu, G.; Li, G.Y. Deep Learning Enabled Semantic Communications with Speech Recognition and Synthesis. *IEEE Trans. Wirel. Commun.* **2023**, *1*. [[CrossRef](#)]
24. Xie, H.; Qin, Z.; Tao, X.; Letaief, K.B. Task-Oriented Multi-User Semantic Communications. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 2584–2597. [[CrossRef](#)]
25. Wang, B.; Li, R.; Zhu, J.; Zhao, Z.; Zhang, H. Knowledge Enhanced Semantic Communication Receiver. *IEEE Commun. Lett.* **2023**, *1*. [[CrossRef](#)]
26. Shao, J.; Mao, Y.; Zhang, J. Learning Task-Oriented Communication for Edge Inference: An Information Bottleneck Approach. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 197–211. [[CrossRef](#)]
27. Luo, X.; Chen, H.H.; Guo, Q. Semantic Communications: Overview, Open Issues, and Future Research Directions. *IEEE Trans. Wirel. Commun.* **2022**, *29*, 210–219. [[CrossRef](#)]
28. Bourtsoulatzé, E.; Kurka, D.B.; Gündüz, D. Deep Joint Source-Channel Coding for Wireless Image Transmission. *IEEE Trans. Cogn. Commun. Netw.* **2019**, *5*, 567–579. [[CrossRef](#)]
29. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, Monticello, IL, USA, 22–24 September 1999, pp. 368–377. [[CrossRef](#)]
30. Hassanpour, S.; Monsees, T.; Wübben, D.; Dekorsy, A. Forward-Aware Information Bottleneck-Based Vector Quantization for Noisy Channels. *IEEE Trans. Commun.* **2020**, *68*, 7911–7926. [[CrossRef](#)]
31. Hassanpour, S.; Wübben, D.; Dekorsy, A. Forward-Aware Information Bottleneck-Based Vector Quantization: Multiterminal Extensions for Parallel and Successive Retrieval. *IEEE Trans. Commun.* **2021**, *69*, 6633–6646. [[CrossRef](#)]
32. Beck, E.; Shin, B.S.; Wang, S.; Wiedemann, T.; Shutin, D.; Dekorsy, A. Swarm Exploration and Communications: A First Step towards Mutually-Aware Integration by Probabilistic Learning. *Electronics* **2023**, *12*, 1908. [[CrossRef](#)]
33. Aguerri, I.E.; Zaidi, A. Distributed Variational Representation Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 120–138. [[CrossRef](#)] [[PubMed](#)]
34. Shao, J.; Mao, Y.; Zhang, J. Task-Oriented Communication for Multidevice Cooperative Edge Inference. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 73–87. [[CrossRef](#)]
35. Simeone, O. A Brief Introduction to Machine Learning for Engineers. *Found. Trends[®] Signal Process.* **2018**, *12*, 200–431. [[CrossRef](#)]
36. Vincent, P.; Larochelle, H.; Jaoje, I.; Bengio, Y.; Manzagol, P.A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
37. Hassanpour, S. Source & Joint Source-Channel Coding Schemes Based on the Information Bottleneck Framework. Ph.D. Thesis, University of Bremen, Bremen, Germany, 2022.
38. Sana, M.; Strinati, E.C. Learning Semantics: An Opportunity for Effective 6G Communications. In Proceedings of the 2022 IEEE 19th Annual Consumer Communications Networking Conference (CCNC), Las Vegas, NV, USA, 8–11 January 2022; pp. 631–636. [[CrossRef](#)]

39. Farsad, N.; Shlezinger, N.; Goldsmith, A.J.; Eldar, Y.C. Data-Driven Symbol Detection Via Model-Based Machine Learning. In Proceedings of the 2021 IEEE Statistical Signal Processing Workshop (SSP), Rio de Janeiro, Brazil, 11–14 July 2021; pp. 571–575. [CrossRef]
40. Goldfeld, Z.; Polyanskiy, Y. The Information Bottleneck Problem and its Applications in Machine Learning. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 19–38. [CrossRef]
41. Zaidi, A.; Estella-Aguerri, I.; Shamai (Shitz), S. On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views. *Entropy* **2020**, *22*, 151. [CrossRef]
42. Kurkoski, B.M.; Yagi, H. Quantization of Binary-Input Discrete Memoryless Channels. *IEEE Trans. Inf. Theory* **2014**, *60*, 4544–4552. [CrossRef]
43. Lewandowsky, J.; Bauch, G. Information-Optimum LDPC Decoders Based on the Information Bottleneck Method. *IEEE Access* **2018**, *6*, 4054–4071. [CrossRef]
44. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep Variational Information Bottleneck. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017; pp. 1–19.
45. Belghazi, M.I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, R.D. MINE: Mutual Information Neural Estimation. In Proceedings of the 35th International Conference on Machine Learning (PMLR), Stockholm, Sweden, 10–15 June 2018.
46. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In Proceedings of the European Conference on Computer Vision 2016 (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Lecture Notes in Computer Science, pp. 630–645. [CrossRef]
49. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. *arXiv* **2016**, arXiv:1603.04467
50. Beck, E. Semantic Information Transmission and Recovery (SINFONY) Software. Zenodo. 2023. Available online: <https://doi.org/10.5281/zenodo.8006567> (accessed on 1 July 2023).
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
52. Hoydis, J.; Cammerer, S.; Aoudia, F.A.; Vem, A.; Binder, N.; Marcus, G.; Keller, A. Sionna: An Open-Source Library for Next-Generation Physical Layer Research. *arXiv* **2022**, arXiv:2203.11854. [CrossRef].
53. Beck, E.; Bockelmann, C.; Dekorsy, A. Model-free Reinforcement Learning of Semantic Communication by Stochastic Policy Gradient. *arXiv* **2023**, arXiv:2305.03571. [CrossRef].

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.