Semantic Communication: An Information Bottleneck View

Edgar Beck, *Graduate Student Member, IEEE,* Carsten Bockelmann, *Member, IEEE,* and Armin Dekorsy, *Senior Member, IEEE*

***Abstract*—Motivated by recent success of machine learning tools at the PHY layer and driven by high bandwidth demands of the next wireless communication standard 6G, the old idea of semantic communication by Weaver from 1949 has received considerable attention. It breaks with the classic design paradigm according to Shannon by aiming to transmit the meaning of a message rather than its exact copy and thus potentially allows for savings in bandwidth.**

In this work, inspired by Weaver, we propose an information-theoretic framework where the semantic context is explicitly introduced into probabilistic models. In particular, for bandwidth efficient transmission, we define semantic communication system design as an Information Bottleneck optimization problem and consider important implementation aspects. Further, we uncover the restrictions of the classic 5G communication system design w.r.t. semantic context. Notably, based on the example of distributed image classification, we reveal the huge potential of a semantic communication system design. Numerical results show a tremendous saving in bandwidth of 20 dB with our proposed approach ISCNet compared to a classic PHY layer design.

***Index Terms*—Application-aware, context-aware, communications, infomax, information bottleneck, information theory, machine learning, semantic, task-oriented.**

I. 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]. Since then the design focus of communication systems shifted towards digital error-free and application-agnostic transmission. Owing to the great success of Artificial Intelligence (AI) and in particular its subdomain Machine Learning (ML) in pattern recognition in the 2010s [2], new ML tools were recently introduced to the PHY layer for further enhancements [3]. Indeed, today the systems already operate close to the Shannon limit. In recent years, it has become clear that agnostic communication limits the achievable efficiency in terms of bandwidth, power and complexity trade-offs. Notable examples include sensor networks and broadcast scenarios.

Motivated by these new ML tools and driven by high bandwidth demands of the next wireless communication standard 6G, thus the old idea of semantic communication has received considerable attention. It breaks with the existing classic design paradigms of human-centric and application agnostic communication according to Shannon by including the application, i.e., the semantic context, into the design. More precisely, semantic communication aims to transmit the meaning of a message rather than its exact copy and hence allows for savings in bandwidth. In fact, the idea arose shortly after Shannon's work [1] in [4] but it remained largely unexplored. Now, ML with is ability to extract features appears to be a proper means to realize a semantic design. Further, we note that the latter approach 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 AI-native air interface (connecting intelligence).

The authors are with the Department of Communications Engineering, University of Bremen, 28359 Bremen, Germany (e-mail: {beck, bockelmann, dekorsy}@ant.uni-bremen.de).

A. Related Work

The notion of semantic communication traces back to Weaver [4] who reviewed Shannon's information theory [1] in 1949 and amended considerations w.r.t. semantic content of messages. We will elaborate on this in Sec. II. Since then semantic communication was mainly investigated from a philosophical point of view, see, e.g., [5], [6].

Notable exceptions include the works [7], [8] where the authors extend the propositional logic-based approach from one of the earliest works [9] into a model-theoretical framework to quantify semantic information in information sources and communication channels. It is mainly based on Shannon's information theory and Weaver's comments and avoids pitfalls of earlier works, e.g., [5]. In particular, the authors consider a semantic source that *"observes the world and generates meaningful messages characterizing these observations"* [8]. The models of the world are equal to conclusions that are unequivocally drawn following a set of known deduction rules based on observation of certain facts, i.e., the messages. Hence, the semantic relationships are deterministic. 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 model-theoretical description leads to "many non-trivial simplifications" [7].

In [10], the authors follow another approach to capture semantics: Semantic similarity is used as a semantic error measure quantifying the distance between 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.

Only recently, enabled by the rise of ML techniques in communications research, semantic communication has been reinvented in the ML context in [11], [12], [13] and applied to

the task of text and speech transmission. As a result, semantic communication is still a nascent field: It remains still unclear what this term exactly means and especially its distinction from joint source channel coding [14], [15]. As a result, many survey paper aim to provide an interpretation, see, e.g., [16], [17], [18]. We will revisit this issue in Sec. II.

B. Main Contribution

The main contributions of this article are manifold:

- We revisit Weaver's notion of semantic communication and show that he argues for the generality of information theory to include the semantic context into communication.
- Inspired by Weaver, we propose an information-theoretic framework for inclusion of the semantic context. There, we explicitly model the latter as a semantic hidden random variable.
- In this framework, we define semantic communication system design as an Information Bottleneck (IB) optimization problem and consider important implementation aspects, e.g., application of Deep Neural Networks (DNNs).
- We uncover the restrictions of the classic 5G communication system design w.r.t. semantic context. We also provide an example where we design a semantic receiver given a classic transmitter.
- Notably, we reveal the huge potential of a semantic communication system design based on the example of distributed image classification. Numerical results show a tremendous saving in bandwidth of 20 dB with our proposed approach ISCNet compared to a classic PHY layer design.

In the following, we reinterpret Weaver's philosophical considerations in Sec. II paving the way for our proposed theoretical framework in Sec. III. In IV, we reveal the design flaws of a classic design w.r.t. semantic context. Finally, in Sec. V and VI, we provide our examples of semantic communication and summarize the main results, respectively.

II. Philosophical Considerations

Despite much renewed interest, research on semantic communication is still in its infancy and recent work reveals a differing understanding of the word *semantic*. Therefore, we think that the meaning of semantics is still elusive and requires resharpening. In this work, we contribute our interpretation which aims to make the view on semantic communication more consistent. To motivate our interpretation, we 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 [1]: *"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.* ***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.*** *The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design."*

In fact, this viewpoint abstracts all kinds of information one may transmit, e.g., oral and written speech, music, pictorial arts, ..., all human behavior, and lays also the foundation for the research area of information theory. Thus, it found its way into many other research areas where data or information is 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 comprehensible review of [1], he first states that *"there* ***seem*** *to be [communication] problems at three levels"* [4]:

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.)

These levels are quoted in recent works where Level C is oftentimes referred to as goal- or task-oriented communication instead. But in the concluding section, he then 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"***. He further points out the bizarre definition of information as a measure of uncertainty and not meaning or semantics. In the end, Weaver indicates based on examples from physics that meaning may be the opposite of information: The less Shannon information a signal carries, the more meaning it has and vice versa. He also mentions *"one of the most significant but difficult aspects of meaning, namely the influence of context"*.

In fact, we argue that it is this context which generates meaning in semantic communication. By introducing context, we are able to introduce meaning, i.e., to reduce uncertainty and hence Shannon information, and thus to save communication bandwidth. This semantic context can be included by different levels such as B and C but the separation is rather arbitrary from our point of view as indicated by Weaver. In fact, we are able to add arbitrarily many levels of details to the communication problem and optimize communications for a specific application.

If we add context/applications that belongs to the domain of humans, it becomes difficult to describe the actual meaning in terms of mathematical modeling. How can we measure if two sentences have the same meaning, i.e., how does the semantic space looks like? In case of image classification, we then need to make use of labeled datasets, i.e., samples, as usually done in the domain of machine learning. According to the argumentation, we can thus distinguish between model and data-driven semantics. Note that both can be handled within information theory. We conclude that as long we stay in the domain of mathematical modeling and machine learning, Shannon's information theory is sufficient for the design of semantic communication systems.

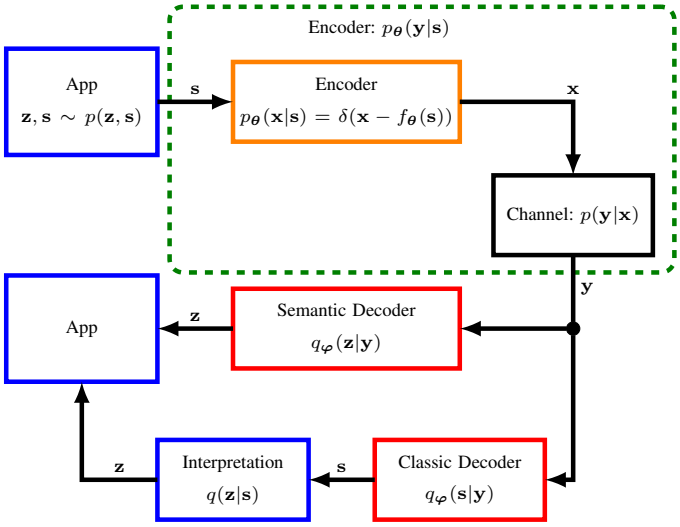

Fig. 1. Block diagram of the considered semantic system model.

Finally, we note that our information theoretic view for inclusion of semantic aspects is in accordance with the works [7], [8] and extends upon them. There, the authors admit that the model-theoretical approach leads to simplifications and does not include level C. In addition, we note that logic-based expert systems that were dominant in earlier work on AI showed limitations, could not satisfy the high expectations in the 1970s and led to the first "AI winter" [2]. In contrast, we propose to use probabilistic models from system theory to capture semantic aspects. Note that the former approaches can be included as special cases of functions. This viewpoint is motivated by recent success of pattern recognition tools which advanced the field of AI in the 2010s and may be used to extract semantics.

III. Theoretical Framework

A. Information-theoretic System Model

With the information theoretic view in mind, we are now ready to define our proposed mathematical probabilistic model shown in Fig. 1 accordingly. Let us assume that, e.g., an application (App), generates a source signal $\mathbf{s} \in \mathcal{M}_s^{N_s \times 1}$, a multivariate Random Variable (RV) of dimension N_s, we observe. 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 [19]. 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.

In classic design, the source $\mathbf{s}$ enters the communication system and the semantic context does not matter. To introduce semantics or the application into the problem, we assume the existence of a hidden target RV $\mathbf{z} \in \mathcal{M}_z^{N_z \times 1}$ jointly distributed with $\mathbf{s}$ according to the joint probability density or mass function (pdf / pmf) $p(\mathbf{s}, \mathbf{z}) = p(\mathbf{s}|\mathbf{z}) \cdot p(\mathbf{z})$ where $p(\mathbf{s}|\mathbf{z})$ is the pdf or pmf of $\mathbf{s}$ conditioned on $\mathbf{z}$. The RV $\mathbf{z}$ is the semantic space or application context of $\mathbf{s}$ which we, e.g., want to infer from $\mathbf{s}$.

Note that semantic context can be included on increasing lagers of complexity. First, a RV $\mathbf{z}_1$ might capture the floating-point representation of continuous RVs. Moving beyond the first application context, from App 1 to App 2, then a RV $\mathbf{z}_2$ might expand this towards the interpretation of said RV like the classification of images or sensor data. Further, image classification could be a sub task of a more general goal. In fact, we can add or remove context arbitrarily often according to the application and we can optimize the overall (communication) system w.r.t. $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_i$, respectively.

Our challenge is to encode the source signal $\mathbf{s}$ onto the transmit signal vector $\mathbf{x} \in \mathcal{M}_x^{N_{\mathrm{Tx}} \times 1}$ (see Fig. 1) for reliable semantic communication through the physical communication channel $p(\mathbf{y}|\mathbf{x})$ where $\mathbf{y} \in \mathcal{M}_y^{N_{\mathrm{R}} \times 1}$ is the received signal vector. We assume the encoder $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})$ to be parametrized by a parameter vector $\boldsymbol{\theta}$. Note that $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})$ is probabilistic here but usually assumed to be deterministic with $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s}) = \delta(\mathbf{x} - f_{\boldsymbol{\theta}}(\mathbf{s}))$ since we aim for uncertainty reduction at the receiver and that $p(\mathbf{y}|\mathbf{x})$ is independent of $\boldsymbol{\theta}$. There, $\delta(\cdot)$ is the Dirac delta function. In summary, the Markov chain reads

$$\mathbf{z} \leftrightarrow \mathbf{s} \leftrightarrow \mathbf{x} \leftrightarrow \mathbf{y}\,. \tag{1}$$

In the following, we summarize transmitter and channel into the probabilistic encoder $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$ for better readability. The Markov chain thus reduces to $\mathbf{z} \leftrightarrow \mathbf{s} \leftrightarrow \mathbf{y}$.

At the receiver side, the optimal decoder of a classic PHY layer design is given by the posterior $p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{y})$ that can be deduced from prior $p(\mathbf{s})$ and likelihood $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$ by application of Bayes law. If calculation of the posterior is intractable, oftentimes the optimal decoder is replaced by an approximation $q_{\boldsymbol{\varphi}}(\mathbf{s}|\mathbf{y})$ with parameters $\boldsymbol{\varphi}$. Based on the estimate of $\mathbf{s}$, then the application interprets the actual semantic content $\mathbf{z}$. In fact, the non-semantic PHY layer design is equal to joint source channel coding of $\mathbf{s}$ without taking $\mathbf{z}$ into account.

We propose to include the semantic hidden target RV $\mathbf{z}$ into the design using a semantic decoder $q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})$. The benefit of doing so lies in the following reason: As outlined in the last section, the actual semantic uncertainty or information content, i.e., the entropy $\mathcal{H}(\mathbf{z})$, can be assumed smaller than the entropy $\mathcal{H}(\mathbf{s})$ of the source. This should finally allow for bandwidth savings: 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 $\mathbf{z}$. Only considering the task of source compression, this means

$$\mathcal{H}(\mathbf{z}) \leq \mathcal{H}(\mathbf{y}) \leq \mathcal{H}(\mathbf{s})\,. \tag{2}$$

Note that in our system model $\mathbf{x}$ and $\mathbf{y}$ should include redundancy w.r.t. $\mathbf{z}$ for robust channel encoding. Therefore, it may also be that $\mathcal{H}(\mathbf{y}) \geq \mathcal{H}(\mathbf{s})$. But note that also $\mathbf{s}$ needs channel encoding and the entropy of its encoding grows.

Finally, we have outlined a complete system model including semantic and communication RVs. By this means, we extend the model from the works [12], [20] that consider joint

source channel coding of text by the semantic space or RV $\mathbf{z}$ as in [8]. Further, we generalize the model-theoretical approach on the semantic and source RV from [7], [8] to arbitrary RVs and probability distributions. We also extend [7] by explicitly distinguishing between source signal $\mathbf{s}$ and transmitted signal $\mathbf{x}$.

B. Learning of a Semantic Communication System via Infomax Principle

After explaining the system model, we are finally able to approach a semantic communication system design: Since we aim to find a hidden representation $\mathbf{y}$ of our source $\mathbf{s}$, we in fact need to solve an unsupervised learning problem. To define an optimization criterion for our discriminative model or encoder $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$, it is useful to follow the infomax principle from an information theoretic perspective [19]. As suggested, this means our aim is to find a representation $\mathbf{y} \sim p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$ that retains a significant amount of information about the semantic RV $\mathbf{z}$, i.e., maximization of the Mutual Information (MI) $I(\mathbf{z};\mathbf{y})$ w.r.t. the encoder $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$ [21]:

$$\begin{aligned}
&\underset{p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})}{\arg\max}\ I(\mathbf{z};\mathbf{y}) && (3)\\
=&\underset{\boldsymbol{\theta}}{\arg\max}\ \mathrm{E}_{\mathbf{z},\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{y})}\left[\ln \frac{p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{y})}{p(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{y})}\right] && (4)\\
=&\underset{\boldsymbol{\theta}}{\arg\max}\ \mathcal{H}(\mathbf{z}) - \mathcal{H}\left(p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{y}), p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y})\right) && (5)\\
=&\underset{\boldsymbol{\theta}}{\arg\max}\ \mathrm{E}_{\mathbf{z},\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{y})}[\ln p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y})]\ . && (6)
\end{aligned}$$

There, $\mathrm{E}_{\mathbf{x}\sim p(\mathbf{x})}[f(\mathbf{x})]$ denotes the expected value of $f(\mathbf{x})$ w.r.t. both discrete or continuous RVs $\mathbf{x}$ and $\mathcal{H}(p(\mathbf{x}), q(\mathbf{x})) = \mathrm{E}_{\mathbf{x}\sim p(\mathbf{x})}[-\ln q(\mathbf{x})]$ is the cross entropy between two pdfs / pmfs $p(\mathbf{x})$ and $q(\mathbf{x})$. Note independence from $\boldsymbol{\theta}$ in $\mathcal{H}(\mathbf{z})$ and dependence in $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y})$ and $p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{y})$ through the Markov chain $\mathbf{z} \to \mathbf{s} \to \mathbf{y}$. Further, note that the form of $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$ has to be constrained to avoid learning a trivial identity mapping $\mathbf{y} = \mathbf{s}$. In our example, we indeed do this by assuming a physical channel $p(\mathbf{y}|\mathbf{x})$.

If calculation of the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y})$ in (6) is intractable, we are able to replace it by a variational distribution $q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})$ with parameters $\boldsymbol{\varphi}$. Similar to the transmitter, Deep Neural Networks (DNNs) are usually used in semantic communication literature [12], [20] for design of the approximate posterior $q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})$ at the receiver. To enhance the performance complexity trade-off, the application of Deep Unfolding can be considered, a model-driven learning approach that introduces model knowledge of $p_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{x},\mathbf{y},\mathbf{z})$ to create $q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})$ [22], [3]. With $q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})$, we are able to define a Mutual Information Lower BOund (MILBO) [21] similar to the well-known Evidence Lower BOund (ELBO) [2]:

$$\begin{aligned}
I_{\boldsymbol{\theta}}(\mathbf{z};\mathbf{y}) &\geq \mathrm{E}_{\mathbf{z},\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{y})}[\ln q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})] && (7)\\
&= -\mathrm{E}_{\mathbf{y}\sim p(\mathbf{y})}[\mathcal{H}(p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y}), q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y}))] && (8)\\
&= -\mathcal{L}^{\mathrm{CE}}_{\boldsymbol{\theta},\boldsymbol{\varphi}}\,. && (9)
\end{aligned}$$

The lower bound holds since $-\mathcal{H}(p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{y}), p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y}))$ itself is a lower bound and $\mathrm{E}_{\mathbf{z},\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{y})}[\ln p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y}) / \ln q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})] \geq 0$. Optimization of $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ can now be done w.r.t. this lower bound:

$$\underset{\boldsymbol{\theta},\boldsymbol{\varphi}}{\arg\max}\ -\mathcal{L}^{\mathrm{CE}}_{\boldsymbol{\theta},\boldsymbol{\varphi}}\,. \tag{10}$$

We note that the MILBO in (7) is equivalent to the negative cross entropy amortized across observations $\mathbf{y}$ [3], i.e., $\mathcal{L}^{\mathrm{CE}}_{\boldsymbol{\theta},\boldsymbol{\varphi}}$, in (8). This means that approximate maximization of the mutual information justifies the minimization of the cross entropy in the Auto Encoder (AE) approach [23] oftentimes seen in recent semantic communication literature [12], [20]. 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 system model, we do not auto encode the hidden $\mathbf{z}$ itself, i.e., $\mathbf{z} = \mathbf{s}$, but encode $\mathbf{s}$ for decoding of $\mathbf{z}$. This can be seen by rewriting the amortized cross entropy:

$$\begin{aligned}
\mathcal{L}^{\mathrm{CE}}_{\boldsymbol{\theta},\boldsymbol{\varphi}} &= \mathrm{E}_{\mathbf{y}\sim p(\mathbf{y})}[\mathcal{H}(p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y}), q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y}))] && (11)\\
&= \mathrm{E}_{\mathbf{s}\sim p(\mathbf{s})}\left[\mathrm{E}_{\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})}[\mathcal{H}(p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y}), q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y}))]\right] && (12)\\
&= \mathrm{E}_{\mathbf{s}\sim p(\mathbf{s})}\left[\mathrm{E}_{\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})}\left[\mathrm{E}_{\mathbf{z}\sim p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y})}[-\ln q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})]\right]\right] && (13)\\
&= \mathrm{E}_{\mathbf{z},\mathbf{y},\mathbf{s}\sim p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{y},\mathbf{s})}[-\ln q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})] && (14)\\
&= \mathrm{E}_{\mathbf{s},\mathbf{z}\sim p(\mathbf{s},\mathbf{z})}\left[\mathrm{E}_{\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})}[-\ln q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})]\right]\ . && (15)
\end{aligned}$$

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

$$\begin{aligned}
&\mathcal{L}^{\mathrm{CE}}_{\boldsymbol{\theta},\boldsymbol{\varphi}}\\
&= \mathrm{E}_{\mathbf{y}\sim p(\mathbf{y})}\left[\mathrm{E}_{\mathbf{z}\sim p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y})}[-\ln q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y}) + \ln p(\mathbf{z}|\mathbf{y}) - \ln p(\mathbf{z}|\mathbf{y})]\right] && (16)\\
&= \mathrm{E}_{\mathbf{y}\sim p(\mathbf{y})}[D_{\mathrm{KL}}(p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y}) \parallel q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y}))]\\
&\quad + \underbrace{\mathrm{E}_{\mathbf{y}\sim p(\mathbf{y})}\left[\mathrm{E}_{\mathbf{z}\sim p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y})}[-\ln p(\mathbf{z}|\mathbf{y})]\right]}_{\mathcal{H}(\mathbf{z}|\mathbf{y}) = -I_{\boldsymbol{\theta}}(\mathbf{z};\mathbf{y}) + \mathcal{H}(\mathbf{z})} && (17)\\
&= \mathcal{H}(\mathbf{z}) - \underbrace{I_{\boldsymbol{\theta}}(\mathbf{z};\mathbf{y})}_{\text{encoder objective}} + \underbrace{\mathrm{E}_{\mathbf{y}\sim p(\mathbf{y})}[D_{\mathrm{KL}}(p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y}) \parallel q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y}))]}_{\text{decoder objective}}\,. && (18)
\end{aligned}$$

This means, optimization of the MILBO balances maximization of the mutual information $I_{\boldsymbol{\theta}}(\mathbf{z};\mathbf{y})$ and minimization of the KL divergence $D_{\mathrm{KL}}(p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{y}) \parallel q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y}))$. The former criterion 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.

To summarize, the cross entropy loss of our approach and of the AE usually used in literature is well-motivated as shown in (18) since its amortized version is equal to the negative MILBO. This is in contrast to the Variational AE (VAE) being used in [24] that maximizes the ELBO on the evidence $p_{\boldsymbol{\theta}}(\mathbf{y})$. Further, we notice that no explicit variational regularization term is present in the MILBO compared to the ELBO. For further discussion, the reader is referred to [25].

C. Information Bottleneck View

So far, we have shown that the InfoMax principle (3) is a valid approach for unsupervised learning of our discriminative model and that of an AE. The InfoMax objective aims at an

probabilistic encoding $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$ of $\mathbf{s}$ which retains most of the information of $\mathbf{z}$ in the received signal $\mathbf{y}$. At this point, we become aware of one limitation: Note that the best but trivial encoder without noise or rate restrictions is $\mathbf{y} = \mathbf{s}$ since then definitely all available information of $\mathbf{z}$ is contained in $\mathbf{y}$.

Hence, we have to restrict the information flow to achieve bandwidth savings. This means we now want to maximize $I_{\boldsymbol{\theta}}(\mathbf{z};\mathbf{y})$ while restricting the information flow to I_{C} with a compressed representation $\mathbf{y}$. Hence, this problem

$$\arg\max_{\boldsymbol{\theta}} I_{\boldsymbol{\theta}}(\mathbf{z};\mathbf{y}) \quad \text{s.t.} \quad I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y}) \leq I_{\mathrm{C}} \tag{19}$$

is called Information Bottleneck (IB) problem [26], [27]. It is proposed in [17] for goal-oriented communications. The IB problem (19) is closely related to rate distortion theory. In particular, the rate-distortion function under logarithmic loss can be shown to coincide with (19) for remote source coding [27].

Now, we turn our attention to how IBM problem (19) can be solved. In fact, algorithms that solve (19) are only known for discrete or continuous Gaussian RVs [27]. Note that in the case of discrete RVs, we aim at an encoder that compresses $\mathbf{s}$ into a compact representation $\mathbf{y}$ by clustering and in the case of continuous RVs by dimensionality reduction. In our first numerical example of the following Sec. V, we will assume all RVs to be continuous. Hence, we will restrict to solve (19) directly by constraining the output dimension of the encoder $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})$ or $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$ to N_{Tx} or N_{R}, respectively, and hence the information flow. In other words, we explicitly introduce an IB. Then, we are able to use the cross entropy in (14) as the optimization criterion and to avoid tuning of I_{C}.

It remains the question whether solving (19) directly as in other works [12], [20], [28] may hold some benefits. To show the additional required effort, we shortly review existing approaches. First, the constrained optimization problem (19) can be turned into an unconstrained one by introduction of the Lagrange function with multiplier β

$$\arg\max_{\boldsymbol{\theta}} I_{\boldsymbol{\theta}}(\mathbf{z};\mathbf{y}) - \beta I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y}) \tag{20}$$

for some fixed $\beta \geq 0$. The Lagrange multiplier β allows to define a trade-off between rate $I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y})$ and distortion $I_{\boldsymbol{\theta}}(\mathbf{z};\mathbf{y})$ which indicates the relation to both classic measures from rate distortion theory. With $\beta = 0$, the objective (20) aims at minimal distortion whereas for $\beta \to \infty$ rate is minimal. Calculation of the mutual information terms may be computational intractable as in the InfoMax problem (3). Notable exceptions include if the RVs are all discrete or Gaussian distributed. Hence, it is necessary to devise to variational approximations. Like for the InfoMax problem (3), we introduce a lower bound, i.e., the MILBO (7), to the first term. The second term includes a marginal w.r.t. $p_{\boldsymbol{\theta}}(\mathbf{y})$ which can become computational intractable. Since the second term has a negative sign, we thus need to find an upper bound:

$$\begin{aligned} I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y}) &= \mathrm{E}_{\mathbf{s},\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{y})}\left[\ln \frac{p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})}{p_{\boldsymbol{\theta}}(\mathbf{y})}\right] && (21)\\ &= \mathrm{E}_{\mathbf{s},\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{y})}[\ln p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})] - \mathrm{E}_{\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{y})}[\ln p_{\boldsymbol{\theta}}(\mathbf{y})] && (22)\\ &\leq \mathrm{E}_{\mathbf{s},\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{y})}\left[\ln \frac{p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})}{q_{\boldsymbol{\vartheta}}(\mathbf{y})}\right] . && (23) \end{aligned}$$

The last inequality follows from $D_{\mathrm{KL}}(p_{\boldsymbol{\theta}}(\mathbf{y}) \parallel q_{\boldsymbol{\vartheta}}(\mathbf{y})) \geq 0$ with $\mathrm{E}_{\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{y})}[\ln p_{\boldsymbol{\theta}}(\mathbf{y})] \geq \mathrm{E}_{\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{y})}[\ln q_{\boldsymbol{\vartheta}}(\mathbf{y})]$ for some variational distribution $q_{\boldsymbol{\vartheta}}(\mathbf{y})$ with parameters $\boldsymbol{\vartheta}$. In total, the lower bound on the IB problem, i.e., the variational IB problem, now reads:

$$\begin{aligned} &I_{\boldsymbol{\theta}}(\mathbf{z};\mathbf{y}) - \beta I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y})\\ \geq\ &\mathrm{E}_{\mathbf{z},\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{y})}[\ln q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})]\\ &- \beta\, \mathrm{E}_{\mathbf{s}\sim p(\mathbf{s})}[D_{\mathrm{KL}}(p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s}) \parallel q_{\boldsymbol{\vartheta}}(\mathbf{y}))] . \end{aligned} \tag{24}$$

This lower bound allows for optimization of $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\vartheta}$ by means of the reparametrization trick with conditions that hold for VAEs [19] and will be partly covered in Sec. III-D for our approach. Additionally, the variational regularization term $D_{\mathrm{KL}}(p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s}) \parallel q_{\boldsymbol{\vartheta}}(\mathbf{y}))$ needs to be analytically computable and differentiable w.r.t. $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ both being possible for members of the exponential family. This "Deep Variational Information Bottleneck" was introduced in [29] and applied in [28] to communications design to find a compressed representation of an image to be transmitted for classification at the receiver side. Note that in some works, e.g., in the foundational work [26], the IB objective (20) is turned into a minimization problem by reversing the sign.

We note that the IB problem (19) is motivated from a semantic point of view w.r.t. to some target or task $\mathbf{z}$. But we can also change this view and apply the IB to joint source channel coding, i.e., the classic communication setup, where $\mathbf{s}$ is the source, $\mathbf{x}$ the transmitted signal/compressed representation and $\mathbf{y}$ the received signal/target variable. Then, the Markov chain reads $\mathbf{s} \leftrightarrow \mathbf{x} \leftrightarrow \mathbf{y}$ and the sign reversed objective from (20) becomes:

$$\begin{aligned} &I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{x}) - \beta I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y}) && (25)\\ \leq\ &\mathcal{H}(\mathbf{s}) - I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y}) - (\beta-1) I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y}) && (26)\\ \leq\ &\mathcal{H}(\mathbf{s}) - I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y}) - (\beta-1) I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y})\\ &+ \mathrm{E}_{\mathbf{y}\sim p(\mathbf{y})}[D_{\mathrm{KL}}(p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{y}) \parallel q_{\boldsymbol{\varphi}}(\mathbf{s}|\mathbf{y}))] && (27)\\ =\ &\mathcal{L}^{\mathrm{CE}}_{\boldsymbol{\theta},\boldsymbol{\varphi}} - (\beta-1) I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y}) . && (28) \end{aligned}$$

The latter upper bound can be derived with $I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{x}) \leq \mathcal{H}(\mathbf{s})$ and by introduction of the variational posterior $q_{\boldsymbol{\varphi}}(\mathbf{s}|\mathbf{y})$ as was the case in Section III-B. It is very similar to the loss function designed in [12], [20] except that we replace $I_{\boldsymbol{\theta}}(\mathbf{x};\mathbf{y})$ by $I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y})$ and have a factor $(\beta-1)$ instead of $(\alpha+1)$. This means we relax $\alpha \geq \beta - 2 \geq -2$ with $\beta \geq 0$ compared to [20] and are able to motivate the loss function in [20] and [12] from an information theoretic perspective. However, note that our choice (25) forces $\mathbf{y}$ to contain much information about $\mathbf{s}$ which is not necessarily true when maximizing $\beta I_{\boldsymbol{\theta}}(\mathbf{x};\mathbf{y})$ while minimizing $I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{x})$. Using the lower bound in (24), we can further avoid estimating the mutual information $I_{\boldsymbol{\theta}}(\mathbf{s};\mathbf{y})$

or $I_{\boldsymbol{\theta}}(\mathbf{x};\mathbf{y})$ at the cost of variational approximation. Finally, we note that finding a compressed representation $\mathbf{x}$ of the source $\mathbf{s}$, as for example obtained by the application of compressed sensing, does not necessarily mean that we obtain the relevant information w.r.t. the application. Hence, we think it is crucial to include the semantic context $\mathbf{z}$.

D. Implementation Considerations

Now, we will provide important implementation considerations for optimization of (10)/(14) and (19). We note that computation of the MILBO leads to similar problems like for the ELBO [19]. If calculating the expected value in (14) cannot be solved analytically or is computational intractable, we can approximate it using Monte Carlo sampling techniques. For Stochastic Gradient Descent (SGD) - based optimization like, e.g., in the AE approach, the gradient w.r.t. $\boldsymbol{\varphi}$ can then be calculated by

$$\frac{\partial}{\partial\boldsymbol{\varphi}}\mathcal{L}^{\mathrm{CE}}_{\boldsymbol{\theta},\boldsymbol{\varphi}} = \frac{\partial}{\partial\boldsymbol{\varphi}}\,\mathrm{E}_{\mathbf{z},\mathbf{s},\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})p(\mathbf{s}|\mathbf{z})p(\mathbf{z})}[-\ln q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})] \tag{29}$$

$$= -\,\mathrm{E}_{\mathbf{z},\mathbf{s},\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})p(\mathbf{s}|\mathbf{z})p(\mathbf{z})}\left[\frac{\partial \ln q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})}{\partial\boldsymbol{\varphi}}\right] \tag{30}$$

$$\approx -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ln q_{\boldsymbol{\varphi}}(\mathbf{z}_i|\mathbf{y}_i)}{\partial\boldsymbol{\varphi}} \tag{31}$$

and by application of the backpropagation algorithm to $\frac{\partial}{\partial\boldsymbol{\varphi}}\ln q_{\boldsymbol{\varphi}}(\mathbf{z}_i|\mathbf{y}_i) = \frac{\partial}{\partial\boldsymbol{\varphi}}q_{\boldsymbol{\varphi}}(\mathbf{z}_i|\mathbf{y}_i)/q_{\boldsymbol{\varphi}}(\mathbf{z}_i|\mathbf{y}_i)$ in automatic differentiation frameworks like TensorFlow. Computation of the so called Reinforce gradient w.r.t. $\boldsymbol{\theta}$ leads to high variance of the gradient estimate since we sample w.r.t. the pdf $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$ dependent on $\boldsymbol{\theta}$ [19]. Typically, the reparametrization trick is used to overcome this problem as in the VAE approach [19]. Here it is applicable if the latent variable $\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$ can be decomposed into a differentiable function $f_{\boldsymbol{\theta}}(\mathbf{n},\mathbf{s})$ and a RV $\mathbf{n}\sim p(\mathbf{n})$ independent of $\boldsymbol{\varphi}$. Fortunately, the typical forward model of a communication system $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$ fulfills this criterion. Assuming a deterministic DNN encoder $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{s})$ and additive noise $\mathbf{n}$ with covariance $\boldsymbol{\Sigma}$, we can thus rewrite into $f_{\boldsymbol{\theta}}(\mathbf{n},\mathbf{s}) = \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{s}) + \boldsymbol{\Sigma}^{1/2}\cdot\mathbf{n}$ and accordingly the Monte Carlo approximation of the amortized cross entropy gradient into:

$$\frac{\partial}{\partial\boldsymbol{\theta}}\mathcal{L}^{\mathrm{CE}}_{\boldsymbol{\theta},\boldsymbol{\varphi}} \tag{32}$$

$$= -\,\mathrm{E}_{\mathbf{z},\mathbf{s},\mathbf{y}\sim p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})p(\mathbf{s}|\mathbf{z})p(\mathbf{z})}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ln q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})\right] \tag{33}$$

$$= -\,\mathrm{E}_{\mathbf{s},\mathbf{n}\sim p(\mathbf{n})p(\mathbf{s}|\mathbf{z})p(\mathbf{z})}\left[\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{n},\mathbf{s})}{\partial\boldsymbol{\theta}}\cdot\frac{\partial \ln q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})}{\partial\mathbf{y}}\right] \tag{34}$$

$$\approx -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{n}_i,\mathbf{s}_i)}{\partial\boldsymbol{\theta}}\cdot\frac{\partial \ln q_{\boldsymbol{\varphi}}(\mathbf{z}_i|\mathbf{y})}{\partial\mathbf{y}}\bigg|_{\mathbf{y}=f_{\boldsymbol{\theta}}(\mathbf{n}_i,\mathbf{s}_i)}. \tag{35}$$

Although the use of this trick becomes evident from a theoretical perspective, in most of recent AE works in communications (where $\mathbf{z}=\mathbf{s}$) it seems that the trick is used without being aware of the underlying problem [23]. This is because in these works optimization of AEs is treated as a supervised learning problem with a simple noise layer in between usually used for regularization in ML literature. We have seen that AEs, although being conceptual simple and easy to implement, and our semantic IB approach require sophisticated techniques from unsupervised learning/information theory to be interpreted and implemented correctly.

E. Supervised Learning and Infomax Principle

As a final remark, we arrive at a special case of the infomax principle if we fix the encoder with $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s}) = p(\mathbf{y}|\mathbf{s})$ and hence the transmitter. Then, only the receiver approximate posterior $q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{y})$ needs to be optimized in (11). Thus, in this case, maximization of the MILBO is equivalent to a supervised learning problem and minimization of KL divergence between true and approximate posterior [3]. This setup has several benefits: In practice, we avoid the Reinforce gradient and especially we do not need any (ideal) connection between transmitter and receiver. Further, even today in 5G, we can apply a semantic receiver design to standardized systems with fixed transmitter capabilities to possibly achieve semantic performance gains. We will investigate a ML-based semantic receiver design given a fixed transmitter in our second example of Sec. V. Finally, we note that the SotA transmitters target at establishing near-deterministic links which may not really be needed from a semantic perspective and a waste of resources. Hence, it is also worth considering adaptation of the transmitter to achieve more efficient use of bandwidth. We will elaborate on this point in the next section.

IV. Semantic Communication in a Classic Design

Including several details of an application, i.e., semantic context, into the communication problem, challenges the conventional communication system design of 5G. Based on Fig. 2, we will explain if it is possible to include the semantic context in 5G design and where the pitfalls lie.

In today's conventional systems, the application (App) still plays a minor role since source encoding completely decouples the application context from the communication system. Furthermore, services or Quality of Service like in 5G seem to be a crude interface to reflect its requirements. First, the source signal $\mathbf{s}$ is encoded by the source encoder for redundancy reduction, encoded with a channel code for error protection and finally modulated for transmission through a channel. All these steps are reversed with respective separated functional blocks at the receiver side. Separation into single blocks is usually preferred since optimization of all blocks together was too difficult/complex in the past. Assuming probabilistic models with factorization between these blocks at the receiver, we arrive at message passing schemes enabling the flow of soft information, e.g., between equalizer $q(\mathbf{x}|\mathbf{y})$ and channel decoder $q(\mathbf{b}|\mathbf{x})$. Message passing is indicated by the integral/summation operation to obtain $q(\mathbf{s}|\mathbf{y})$.

In particular, Shannon proved with the separation theorem that separate source and channel coding is optimal for large block-lengths and point-to-point transmission [30]. As a result, source coding (also known as data compression) and channel coding mainly have been investigated independently in the last decades. However, the theorem does not hold for multi-point communication and does not imply that coding must be used at all [30].

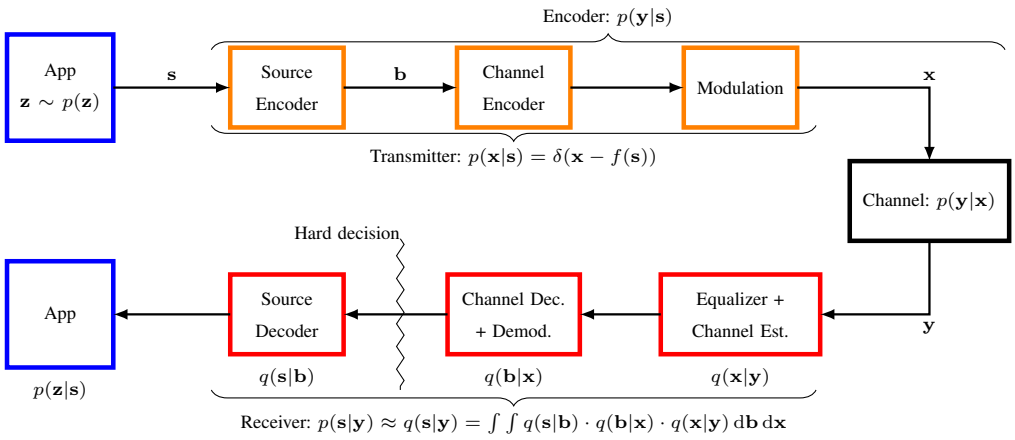

Fig. 2. Conventional communication system design and introduction of semantic context.

Now, we explain the major drawback of the conventional communications design when it comes to semantics: It only accounts for the entropy $\mathcal{H}(\mathbf{s})$ of the source $\mathbf{s}$ but not the entropy $\mathcal{H}(\mathbf{z})$ of the application $\mathbf{z}$ behind. For example for lossless semantic transmission according to the separation theorem, the product of code rate R and channel capacity C needs to be higher than the entropy:

$$\mathcal{H}(\mathbf{z}) \leq \mathcal{H}(\mathbf{s}) \leq RC\,. \tag{36}$$

In fact, this means a code with higher and thus more bandwidth efficient code rate would be sufficient w.r.t. $\mathbf{z}$ although reception of $\mathbf{s}$ becomes lossy. But moving RC below $\mathcal{H}(\mathbf{s})$ is critical since errors are not tolerated by design: Most of the source coding standards use Variable-Length Codes (VLC) such as Huffman coding which makes it very sensitive to errors at the decoding stage [31]. More specifically, decoding errors can lead to different bit sequence lengths after source decoding and hence make the communication system output meaningless (in terms of semantic output). Therefore, channel decoders are usually designed to achieve a low frame error rate, i.e., it only allows hard decisions to be propagated. The last point means that there is usually an "information barrier" between channel and source decoder as indicated in Fig. 2: Uncertainty being equivalent to Shannon information cannot be propagated to higher layers and used by the application. In particular, this makes designing a semantic receiver given a standard transmitter and with or without standard receiver blocks like in Sec. V-B very difficult in practice.

Further, powerful channel codes oftentimes have waterfall regions which amplifies the "cliff effect" [15]: Either channel capacity is above the code rate and transmission is nearly deterministic or the link fails. This means that several codes with rates adapted for certain SNR regions are required and the complexity of the communication system grows.

A second weak point of the conventional design is that the required large block lengths for source and channel coding as well as interleavers for statistical decoupling of the processed symbols or bits, e.g., between channel decoder and equalizer, add a huge amount of latency. Interleavers avoid non-i.i.d. input data with memory for which these systems are not designed.

To overcome these 2 major design flaws w.r.t. semantics, we conclude to remove the block-wise structure at transmitter and receiver. This can be achieved by means of

1) joint source channel coding. Recent work considers AEs for this task and has shown performance improvements at low SNR for language, speech and image transmission [14], [15], [12].
2) the DNN-based semantic IB approach we outlined in Sec. III.

Now, we will give numerical examples demonstrating superior performance of the semantic IB approach compared to the classic design.

V. Examples of Semantic Communication

In this section, we provide two examples to explain what we understand under a context-aware or semantic communication design and to show the benefits of such a design.

A. Distributed Image Classification

In our first example, the task of image classification, we introduce higher level context into the communication problem. Since the world model stems from labeling of human beings, our scenario is an example of data-driven semantics.

Note that if we have unlimited processing resources at an agent that captures the images in this scenario, classification of the image by the agent and subsequent transmission of the result would be optimal and most bandwidth efficient. In [28], task-oriented communication is motivated by limited processing capability of the agent that captures the image. Therefore, the authors aim for an informative and compact representation for inference at the powerful edge with limited bandwidth.

In contrast, we assume a distributed setting shown in Fig. 3 where each of 4 agents gathers an image $\mathbf{s}$ independently which is generated by a hidden process z. Based on these images, a central unit shall perform classification. For example, we aim to detect if a burglary happens in a security scenario or if we have found a person to be rescued in an exploration scenario. We note that transmitting each image through a noisy channel from agent to central unit would consume a

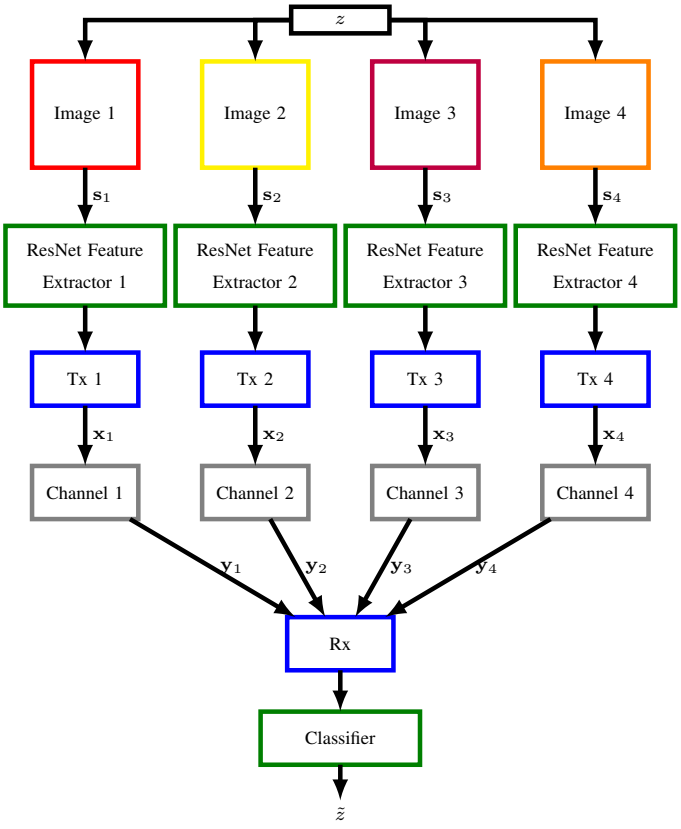

Fig. 3. Integrated semantic communication system (ISCNet) for distributed agents. Each agent extracts features for bandwidth efficient transmission. Based on the received signal, the central unit performs classification.

lot of bandwidth. Hence, we propose to optimize a bandwidth efficient feature extractor jointly with transmitter, receiver and concluding classifier to maximize the systems overall classification accuracy (see Fig. 3). Note that communication is optimized for the effectiveness level C.

As a first demonstration example, we use the classic datasets MNIST and CIFAR10 with 10 image classes to detect [32]. We assume that z generates an image that we divide into 4 equally sized quadrants and each agent observes one quadrant $\mathbf{s}_1, \ldots, \mathbf{s}_4$. Albeit this does not resemble a realistic scenario note that we can still show the basic working principle and ease implementation. In this as well as more realistic scenarios, transmission of features, i.e., soft information, is crucial to obtain higher classification accuracy since some (sub) images contribute more useful information than others.

1) ResNet: For design of the overall system, we rely on a famous DNN approach for feature extraction breaking records at the time of invention: ResNet [32], [33]. The key idea of ResNet is that it consists of multiple residual units: Each units input is fed directly to its output and if the dimensions do not match, a convolutional layer is used. This structure allows for fast training and convergence of DNNs since the training error can be backpropageted 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 layer, does not necessarily increase the expressive power. However, this holds for nested functions like ResNet

TABLE I
DISTRIBUTED IMAGE CLASSIFICATION WITH INTEGRATED SEMANTIC COMMUNICATION NETWORK (ISCNET).

Component	Layer	Dimension
Input	Image	$(14, 14, 1)$, $(16, 16, 3)$
$4\times$	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)
$4\times$ Tx	ReLU	N_{Tx}
	Linear	N_{Tx}
	Normalization (dim.)	N_{Tx}
Channel	AWGN	N_{Tx}
Rx	ReLU ($4\times$ same)	$(2, 2, N_{\text{Rx}})$
	GlobalAvgPool2D	N_{Rx}
Classifier	Softmax	10

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 to extract translation invariant and local features across two spatial dimensions. Color channels like in CIFAR10 add a third dimension 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).

In this work, we use the preactivation version of ResNet without bottlenecks from [32], [33] implemented for classification on the dataset CIFAR10. In Tab. I, we show its structure for the distributed scenario. It is also valid for the central case if we remove the components Tx, Channel, Rx and increase each spatial dimension by 2 to contain all quadrants of the original image. Each ResNetBlock consists of multiple residual units (res. un.) and we use 2 for MNIST and 3 for CIFAR10 which means we use ResNet14 and ResNet20, respectively. For further implementation details, we refer the reader to the original work [33].

2) Proposed Distributed Communications Design: Our key idea here is to modify ResNet w.r.t. the communication task by splitting it at a suitable point where semantic information with low-bandwidth is present (see Fig 3). ResNet and CNNs in general can be interpreted to extract features: With full images, we obtain a feature map of size $8\times8\times N_{\text{Feat}}$ after the last ReLU activation (see Tab. I). These local features are aggregated by the global average pooling layers across the 2 spatial dimensions. Based on these N_{Feat} global features, the softmax layer finally classifies the image. We note that the features contain the relevant information and are of low dimension compared to the original image or even its sub images, i.e., $16 \times 16 \times 3 = 768$ compared to 64 for CIFAR10. Therefore, we aim to transmit each agent's local features instead of all sub images and add DNN layers, i.e., the component Tx in Tab. I, to encode the features for transmission trough a channel. Normalization of Tx output across the batch or the

encode vector dimension (dim.) is required to constrain the output power to one. At the receiver side, we accordingly use a Rx module: For its design, we are able to employ prior knowledge that the features can be found everywhere on the sub images. Further, we assume the channels to be AWGN. Hence, we choose to enforce the results of all 4 feature extractors and channel encoders to be the same and to use the same receive layers for all 4 transmit signals. Then, we aggregate the decoded feature map of size $(2, 2, N_{\text{Rx}})$ across the spatial dimension. Based on the received features, finally classification is performed by a softmax layer with 10 units yielding the MAP estimate $\tilde{z}$. In the following, we name our proposed approach for semantic communication Integrated Semantic Communication Network (ISCNet).

3) Optimization Details: We evaluate ISCNet in TensorFlow 2 on MNIST and CIFAR10. We split the data set into 60k/50k training data and 10k validation data samples, respectively. We do not make use of data augmentation in contrast to [32], [33] which leads to slightly worse accuracy. The ReLU layers are initialized with uniform distribution according to He and all other layers according to Glorot [34]. In case of CIFAR10 classification on the central unit with original ResNet, we need to train $|\boldsymbol{\theta}| = 273,066$ parameters. This number grows more than 4 times to $|\boldsymbol{\theta}| + |\boldsymbol{\varphi}| = 1,127,754$ with $N_{\text{Tx}} = N_{\text{Feat}} = 64$ for the distributed feature extraction communication system due to having 4 agents with additional channel encoder (Tx). Only $|\boldsymbol{\varphi}| = 4810$ parameters amount to channel decoder (Rx) and classification, i.e., the central unit. We note that the number of added Tx and Rx parameters of 33560 and 3192 is relatively small. For l_2 regularization, we use a weight decay of 0.0001 as in [32], [33]. For optimization of the cross entropy (14), we use Stochastic Gradient Descent (SGD) with momentum of 0.9 and a batch size of 64. The learning rate of 0.1 is reduced to 0.01 and 0.001 after 100 and 150 epochs for CIFAR10 and after 3 and 6 for MNIST. In total, we train for 200 epochs with CIFAR10 and for 10 with MNIST. In order to optimize the transceiver for a wider SNR range of channels, we choose the SNR to be uniformly distributed within $[-4, 6]$ where $\text{SNR} = 1/\sigma_{\text{n}}^2$.

All layers of the transmitter have width N_{Tx} and those of the receiver N_{Rx}. Note that N_{Tx} determines the number of channel uses and hence the information bottleneck in (19). A higher width N_{Rx} is equivalent to more computing power at the receiver. Since the number of parameters only weakly grows with N_{Rx} in our design, we choose high $N_{\text{Tx}} = 4N_{\text{Rx}}$. One exception is default ISCNet with $N_{\text{Tx}} = N_{\text{Rx}} = N_{\text{Feat}}$.

4) Numerical Results: The numerical results of our proposed approach ISCNet in terms of classification error rate on MNIST are shown in Fig. 4. First, we observe that the classification error rate of 0.5% based on a central ResNet unit with full image information (central) is smaller than that of 0.9% in the distributed setting (distributed) without noise and Tx (only normalization in Tx). Note that we assume ideal communication links. However, the difference seems negligible considering that the local agents only see the quarter of the full images and learn features independently based on it. With noisy communication links (with noise), the performance degrades especially for SNR $<$ 10 dB and

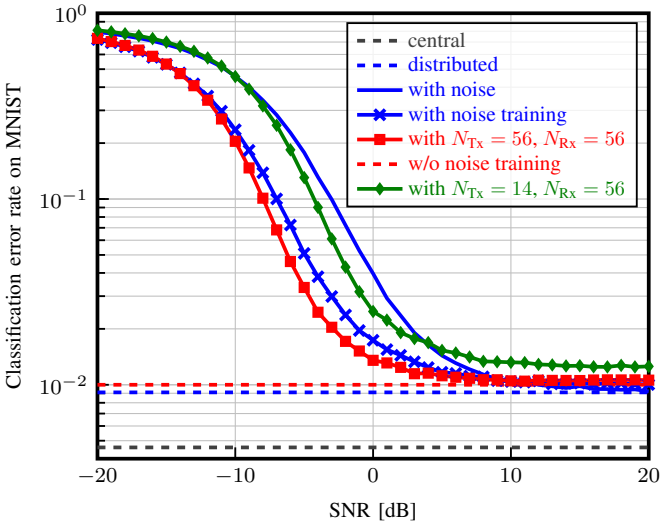

Fig. 4. Classification error rate of the central unit on MNIST with full image information and with distributed agents with ISCNet as a function of SNR.

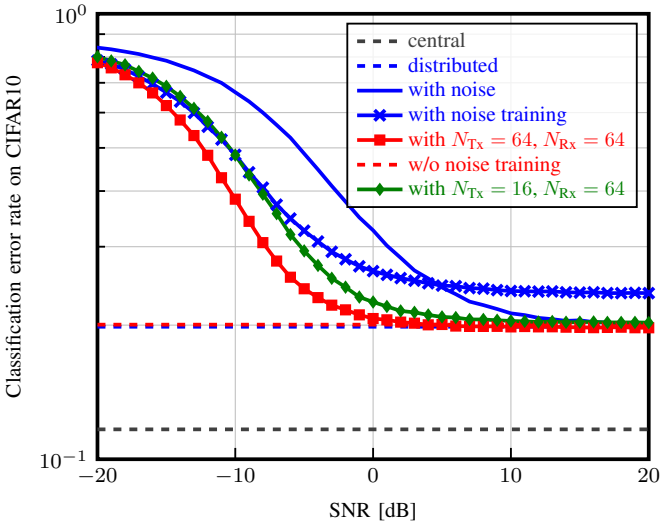

Fig. 5. Classification error rate of the central unit on CIFAR10 with full image information and with distributed agents with ISCNet as a function of SNR.

we can avoid it partly by training with noise (with noise training). Introducing channel encoding (Tx) and decoding (Rx) (with $N_{\text{Tx}} = N_{\text{Feat}} = 56$, $N_{\text{Rx}} = 56$), we further improve classification accuracy at low SNR. Assuming and training with ideal links (w/o noise training) reveals that accuracy drops slightly compared to the approach without communication layers since we loose the spatial separation by adding dense ReLU layers. If we encode the features from $N_{\text{Feat}} = 56$ to $N_{\text{Tx}} = 14$ at the Tx (with $N_{\text{Tx}} = 14$, $N_{\text{Rx}} = 56$) to use less bandwidth, accuracy is higher than in a distributed system optimized with ideal links for low SNR. At high SNR, we observe a small error offset which indicates lossy compression. In fact, the communication system ISCNet learns joint source channel coding for the task of image classification and improves performance of the overall system with non-ideal links.

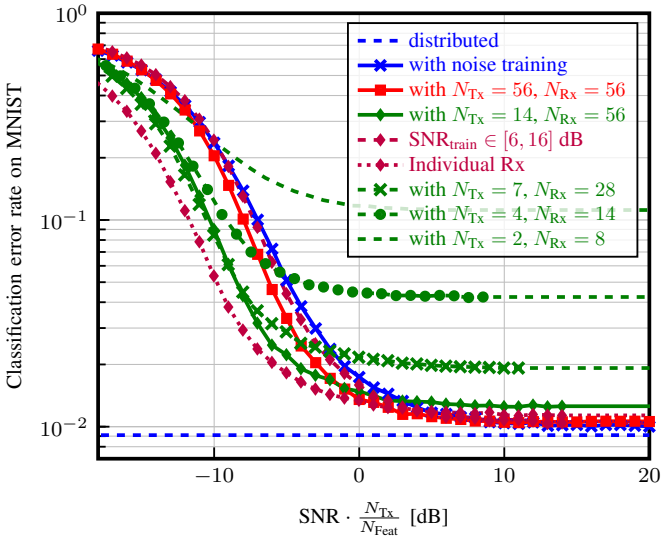

Fig. 6. Classification error rate of the central unit on MNIST with distributed agents for different ISCNet designs as a function of SNR.

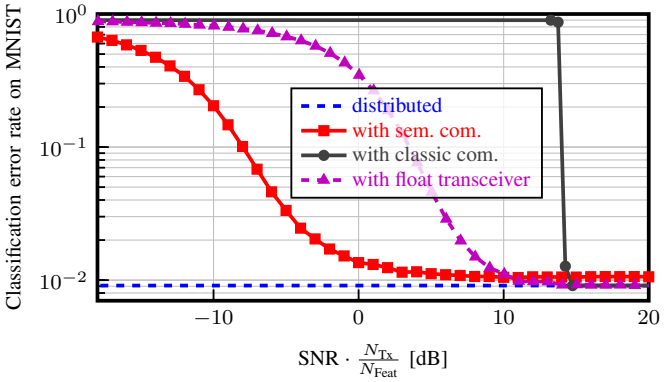

Fig. 7. Classification error rate of the central unit on MNIST with distributed agents for semantic and classic communication system design as a function of SNR.

Comparing these results to the classification accuracy on CIFAR10 shown in Fig. 5, we observe a similar behavior. But a few main differences become apparent: The central unit performs much better with 12% error rate than the distributed setting with ideal links with 20%. We expect the reason to lie in the more challenging dataset with more color channels. Further, the distributed system without Tx and Rx trained with noise (with noise training) and hence 64 channel uses runs into an error floor. Even the overall system with Tx/Rx, i.e., ISCNet, and only $N_{\text{Tx}} = 16$ channel uses (with $N_{\text{Tx}} = 16$, $N_{\text{Rx}} = 64$) achieves channel coding with negligible compression.

Since one of the main advantages of semantic communication lies in savings of bandwidth, we finally investigate the influence of the number of channel uses N_{Tx} on MNIST classification accuracy. For fair comparison between channel encodings of different length, we normalize the SNR in Fig. 6 by $N_{\text{Tx}}/N_{\text{Feat}}$ where $N_{\text{Feat}} = 56$. With less channel uses from $N_{\text{Tx}} = 14$ to 2, the channel coding gain seen at low effective SNR remains the same. In contrast, the error floor moves higher which hints at increased compression rate. We can avoid this error floor by training for higher $\text{SNR}_{\text{train}} \in [6, 16]$ dB but this reduces the coding gain. Also for $N_{\text{Tx}} = 64$, the channel coding gain may be smaller but no error floor occurs. We conclude that we are able to trade-off channel and source coding by choosing different training SNR as well as by varying Tx and Rx in dimensions N_{Tx} and N_{Rx}.

We also investigated whether alternative ISCNet designs improve performance: Using more expressive Tx and Rx DNN layers, i.e., 2 or 3 consecutive ReLU layers for $N_{\text{Tx}} = 14$ (see Tab. I), we achieve similar or worse accuracy, even with $N_{\text{epoch}} = 20$. If the receiver is not shared across the 4 Tx but individual for each Tx (Individual Rx, see Fig. 6), we observe a small decrease in error rate. The error rate does not decrease further using one large joint Rx layer that all received signals enter jointly. We conjecture that the 4 image sections contain slightly different relevant information about the classes and need to be encoded differently. However, the gain is minor and justifies the design choice of using only one receiver for all Tx.

Finally, we compare semantic and classic communication system design. For the classic design, we replace Tx and Rx in Tab. I: We first compress each element of the feature vector that is computed in 32-bit floating-point precision in the distributed setting (with noise) 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 [35] with rate 0.25 and long block length of 15360 and modulate the code bits with BPSK. At the receiver, we assume belief propagation decoding where the noise variance is perfectly known for LLR computation. The results in Fig. 7 reveal tremendous bandwidth savings for the semantic design with ISCNet: We observe an enormous SNR shift of 20 dB compared to the classic design. Note that the classic PHY layer 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. In conclusion, this surprisingly clear result justifies a semantic PHY layer design and shows its huge potential to provide bandwidth savings.

B. Floating-point transmission

In our second example, we demonstrate that also classic problems like unequal error protection that are very close to the technical level can be tackled in our theoretical framework. The example is floating-point transmission with subsequent computations on a digital system. Note that it is rather a numerical toy example and introduces context on a very abstract level compared to our first example. Since the world model is created for interpretation by machines, we deal with model-driven semantics.

For example in distributed systems like multi-agent exploration or our first example, the data in the form of floating-point numbers is exchanged between agents and needs to be communicated. Referring to Fig. 1, an application generates continuous data z processed as discrete floating-point numbers, i.e., as bits $\mathbf{s}$, on digital hardware. Usually, these source bits

TABLE II
DNN BASED TRANSMITTER AND RECEIVER FOR SEMANTIC PHY LAYER DESIGN WITH FLOATING-POINT NUMBERS.

Component	Layer	Dimension
Tx	ReLU	$2N_\mathrm{b}$
	ReLU	$2N_\mathrm{b}$
	Linear	$2N_\mathrm{b}$
	Normalization (dim.)	N_b
Channel	AWGN	N_b
Rx	ReLU	$2N_\mathrm{b}$
	ReLU	$2N_\mathrm{b}$
	Linear	1

enter the transmitter that encodes the bits for deterministic reception at the receiver. This means the transmitter is agnostic about the meaning or context z of $\mathbf{s}$ we want rather to reconstruct at the receiver. Note that usually every bit would be considered stochastic independent in a classic communication system, i.e., $p(\mathbf{s}) \approx \prod_{i=1}^{N_\mathrm{b}} p(s_i)$ where N_b is the number of floating-point bits, and detected separately. We call this approach single bit detector and assume that the prior probability $p(s_i)$ of each single bit s_i is known.

As a first step in semantic communications design, we now focus on design of a semantic receiver given a classic transmitter. To achieve this, we use a simple abstraction of the transmission system: We assume a BPSK transmission of the N_f bits $\mathbf{s}$ of each floating-point number over an AWGN channel with noise variance σ_n^2 to a receiver. Based on the statistics/prior $p(\mathbf{s})$ of the data, we are able to compute the ideal posterior $p(s|\mathbf{y}) = p(\mathbf{y}|s) \cdot p(s)/p(\mathbf{y})$ by marginalization of $p(\mathbf{y})$. From our simulations, we note that for computational tractability the resolution needs to be lower than 16 bit.

Further, note that we are not interested in reconstructing the bits $\mathbf{s}$ exactly, i.e., MAP detection, since these bits are mapped via a deterministic function $f(\cdot)$ into the real-valued domain $z = f(\mathbf{s})$, i.e., the semantic space. Hence in this scenario, we thus want our receiver estimate to be close to the target variable z in the Mean Square Error (MSE) sense. Therefore, the mean estimator based on $p(z|\mathbf{y})$ is optimal. More precisely, one floating-point value consists of a signed bit, significand and exponent bits that contribute through the form of function $z = f(\mathbf{s})$ defined in the standard [36]. This means each bit of the float has a different meaning and is of different importance for our task of reconstructing z.

For approximate estimation, we can approximate the model based inference at the receiver by a Gaussian $q_{\boldsymbol{\varphi}}(z|\mathbf{y})$. We parametrize its mean by a small DNN designed as the component Rx in Tab. II and optimize it by minimization of the cross entropy (11) which is equivalent to the MSE loss in this case. Moving beyond receiver design, we can also parametrize the encoder $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{s})$ by a DNN and optimize the resulting semantic transceiver, i.e., ISCNet, via (11). Our selected structure is shown in Tab. II. Note that normalization of the encoder output across the batch or the encode vector dimension is required to constrain the output power to one. For training of the DNN receiver and transceiver, we initialize ReLU layers with uniform distribution according to He and

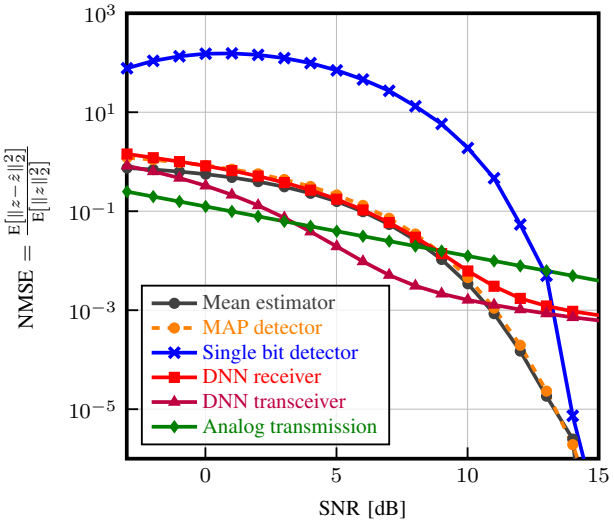

Fig. 8. NMSE as a function of SNR for different non- and semantic transceiver approaches and 8-bit floating-point resolution.

all other layers according to Glorot [34]. As optimizer we use Adam with batch size of 1000 for 10000 iterations and perform 10 steps per iteration. To optimize the DNNs for a wider SNR range, we choose the SNR in dB to be uniformly distributed within $[6, 16]$ where in linear range SNR $= 1/\sigma_\mathrm{n}^2$.

In Fig. 8, we show the Normalized MSE (NMSE) performance of the proposed (sub) optimal approaches as a function of SNR. We assume the target values z to be Gaussian distributed and that NaN as well as $\pm$inf values do not occur. For computational tractability, we consider 8-bit floating-point numbers (minifloats) with one signed bit, 4 exponent and 3 significand bits. The NMSE of analog transmission of z over the AWGN channel (with N_b channel uses for fair comparison) is shown as reference curve. The classic approach with subsequent $z = f(\mathbf{s})$ is clearly inferior in the considered SNR range. Note that we correct NaN and $\pm$inf to the most probable bits based on $p(s_i)$. We observe clearly superior performance of the DNN receiver compared to the classic single bit detector approaching that of the mean estimate at low SNR with much lower computational complexity. If we also optimize the encoder, we are able to surpass the NMSE of the classic receivers in the low SNR regime. We assume that the DNN transceiver neglects bits that have limited contribution to z and performs lossy compression so that more important bits can be transmitted more reliable. By this means, the transceiver uses the bandwidth more efficiently. For high SNR, both DNN receiver and transceiver are not able to increase the precision arbitrarily. We think that this drawback can be overcome by training at higher SNR and the introduction of the noise variance into the design.

Finally, we also investigate performance of the float transceiver in the scenario of distributed image classification. The classification error rate curve shown in Fig. 7 from Sec. V-A lies almost exactly in the middle between that of a full semantic and a classic design, separated by 10 dB. We conclude that even with partial semantic knowledge, a semantic design yields tremendous gains and can be realized

with manageable effort. As a final remark, we note that the example of this section can be treated also in the framework of rate distortion theory. In fact, analog transmission of a Gaussian through a AWGN channel is optimal from this point of view.

VI. Conclusion

In this article, inspired by Weaver, we proposed an information-theoretic framework where the semantic context is explicitly introduced as hidden random variable into the communication design. In particular, for bandwidth efficient transmission, we proposed to define semantic communication system design as an Information Bottleneck (IB) optimization problem and covered important implementations aspects like the infomax principle and the reparametrization trick. Further, we uncovered that variable length source codes and huge interleavers decouple the semantic context from a classic 5G communication system, a major design flaw. Finally, based on two examples with different levels of semantic context, we motivated our view on semantic communication. Notably, based on the example of distributed image classification, we revealed the huge potential of a semantic communication system design. Numerical results show a tremendous saving in bandwidth of 20 dB with our proposed approach ISCNet compared to a classic PHY layer design. In the second example of floating-point number transmission, we showed that also classic problems can be tackled within our framework and that semantic gains can be achieved by only changing the receiver given a classic transmitter.

With these theoretical findings, our future work will focus on technical aspects of semantic communication. For example, it remains unclear if solving the variational IB problem with explicit constraint holds benefits compared to our proposed approach. Also further research is required to clarify how a semantic design can be implemented in practice.

References

[1] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[2] O. Simeone, "A Very Brief Introduction to Machine Learning with Applications to Communication Systems," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 648–664, Dec. 2018.

[3] E. Beck, C. Bockelmann, and A. Dekorsy, "CMDNet: Learning a Probabilistic Relaxation of Discrete Variables for Soft Detection With Low Complexity," *IEEE Transactions on Communications*, vol. 69, no. 12, pp. 8214–8227, Dec. 2021.

[4] W. Weaver, "Recent Contributions to the Mathematical Theory of Communication," in *The Mathematical Theory of Communication*, 1949, vol. 10, pp. 261–281. [Online]. Available: https://www.jstor.org/stable/42581364

[5] L. Floridi, "Philosophical Conceptions of Information," in *Formal Theories of Information: From Shannon to Semantic Information Theory and General Concepts of Information*, ser. Lecture Notes in Computer Science, G. Sommaruga, Ed. Berlin, Heidelberg: Springer, 2009, pp. 13–53. [Online]. Available: https://doi.org/10.1007/978-3-642-00659-3_2

[6] W. Hofkirchner, *Emergent Information: A Unified Theory of Information Framework*. World Scientific, 2013.

[7] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler, "Towards a theory of semantic communication," in *2011 IEEE Network Science Workshop*, Jun. 2011, pp. 110–117.

[8] P. Basu, J. Bao, M. Dean, and J. Hendler, "Preserving quality of information by using semantic relationships," *Pervasive and Mobile Computing*, vol. 11, pp. 188–202, Apr. 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1574119213000953

[9] R. Carnap and Y. Bar-Hillel, "AN OUTLINE OF A THEORY OF SEMANTIC INFORMATION," *Research Laboratory of Electronics, Massachusetts Institute of Technology*, p. 54, 1952.

[10] B. Güler, A. Yener, and A. Swami, "The Semantic Communication Game," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 787–802, Dec. 2018.

[11] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep Learning based Semantic Communications: An Initial Investigation," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, Dec. 2020, pp. 1–6.

[12] ——, "Deep Learning Enabled Semantic Communication Systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[13] Z. Weng, Z. Qin, and G. Y. Li, "Semantic Communications for Speech Signals," in *ICC 2021 - IEEE International Conference on Communications*, Jun. 2021, pp. 1–6.

[14] N. Farsad, M. Rao, and A. Goldsmith, "Deep Learning for Joint Source-Channel Coding of Text," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 2326–2330.

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

[16] P. Popovski, O. Simeone, F. Boccardi, D. Gündüz, and O. Sahin, "Semantic-Effectiveness Filtering and Control for Post-5G Wireless Connectivity," *Journal of the Indian Institute of Science*, vol. 100, no. 2, pp. 435–443, Apr. 2020. [Online]. Available: https://doi.org/10.1007/s41745-020-00165-6

[17] E. Calvanese Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, May 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128621000773

[18] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, "What is Semantic Communication? A View on Conveying Meaning in the Era of Machine Intelligence," *Journal of Communications and Information Networks*, vol. 6, no. 4, pp. 336–371, Dec. 2021.

[19] O. Simeone, "A Brief Introduction to Machine Learning for Engineers," *Foundations and Trends® in Signal Processing*, vol. 12, no. 3-4, pp. 200–431, Aug. 2018.

[20] M. Sana and E. C. Strinati, "Learning Semantics: An Opportunity for Effective 6G Communications," in *2022 IEEE 19th Annual Consumer Communications Networking Conference (CCNC)*, Jan. 2022, pp. 631–636.

[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *Journal of Machine Learning Research*, vol. 11, no. 110, pp. 3371–3408, 2010. [Online]. Available: http://jmlr.org/papers/v11/vincent10a.html

[22] N. Farsad, N. Shlezinger, A. J. Goldsmith, and Y. C. Eldar, "Data-Driven Symbol Detection Via Model-Based Machine Learning," in *2021 IEEE Statistical Signal Processing Workshop (SSP)*, Jul. 2021, pp. 571–575.

[23] T. O'Shea and J. Hoydis, "An Introduction to Deep Learning for the Physical Layer," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.

[24] Y. M. Saidutta, A. Abdi, and F. Fekri, "Joint Source-Channel Coding Over Additive Noise Analog Channels Using Mixture of Variational Autoencoders," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 2000–2013, Jul. 2021.

[25] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy, "Fixing a Broken ELBO," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Jul. 2018, pp. 159–168. [Online]. Available: https://proceedings.mlr.press/v80/alemi18a.html

[26] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv:physics/0004057*, Apr. 2000. [Online]. Available: http://arxiv.org/abs/physics/0004057

[27] Z. Goldfeld and Y. Polyanskiy, "The Information Bottleneck Problem and its Applications in Machine Learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 19–38, May 2020.

[28] J. Shao, Y. Mao, and J. Zhang, "Learning Task-Oriented Communication for Edge Inference: An Information Bottleneck Approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 197–211, Jan. 2022.

[29] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep Variational Information Bottleneck," *arXiv:1612.00410 [cs, math]*, Oct. 2019. [Online]. Available: http://arxiv.org/abs/1612.00410

[30] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: lossy source-channel communication revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.
[31] A. Zribi, R. Pyndiah, S. Zaibi, F. Guilloud, and A. Bouallegue, "Low-Complexity Soft Decoding of Huffman Codes and Iterative Joint Source Channel Decoding," *IEEE Transactions on Communications*, vol. 60, no. 6, pp. 1669–1679, Jun. 2012.
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[33] ——, "Identity Mappings in Deep Residual Networks," in *Computer Vision – ECCV 2016*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 630–645.
[34] ——, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," 2015, pp. 1026–1034. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html
[35] J. Hoydis, S. Cammerer, F. A. Aoudia, A. Vem, N. Binder, G. Marcus, and A. Keller, "Sionna: An Open-Source Library for Next-Generation Physical Layer Research," *arXiv:2203.11854 [cs, math]*, Mar. 2022. [Online]. Available: http://arxiv.org/abs/2203.11854
[36] "IEEE Standard for Floating-Point Arithmetic," *IEEE Std 754-2019 (Revision of IEEE 754-2008)*, pp. 1–84, Jul. 2019.

Edgar Beck (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Bremen, Germany, in 2014 and 2017, respectively, where he is currently pursuing the Ph.D. degree in electrical engineering at the Department of Communications Engineering (ANT). His research interests include cognitive radio, compressive sensing, massive MIMO systems, semantic communication and machine learning in wireless communications.

Edgar Beck was a recipient of the OHB Award for the best M.Sc. degree in Electrical Engineering and Information Technology at the University of Bremen in 2017.

Dr. Carsten Bockelmann (Member, IEEE) received the Dipl.-Ing. and Ph.D. degrees in electrical engineering from the University of Bremen, Germany, in 2006 and 2012, respectively. Since 2012, he has been a Senior Research Group Leader with the University of Bremen coordinating research activities regarding the application of compressive sensing and machine learning to communication problems. His research interests include massive machine-type communication, ultra reliable low latency communications and industry 4.0, compressive sampling and channel coding.

Prof. Dr. Armin Dekorsy (Senior Member, IEEE) is currently the head of the Department of Communications Engineering, University of Bremen. He is also a Directory Board Member of the Gauss-Olbers Space Technology Transfer Center (GOC), University of Bremen.

He is distinguished by more than ten years of industrial experience in leading research positions, namely DMTS at Bell Labs Europe and the Head of Research Europe Qualcomm Nuremberg, and by conducting (inter)national research projects (more than 25 BMBF/BMWI/EU projects) in affiliation with his scientific expertise shown by more than 200 journal and conference publications and more than 19 patents. He investigates new lines of research in wireless communication and signal processing for the baseband of transceivers of future communication systems, the results of which are transferred to the pre-development of industry through political and strategic activities. His current research focuses on distributed signal processing, compressive sampling, information bottleneck method, semantic communication, and machine learning leading to the further development of communication technologies for 6G and beyond, industrial wireless communications and NewSpace satellite communications.

He is a Senior Member of the IEEE Communications and Signal Processing Society and the Head of VDE/ITG Expert Committee "Information and System Theory".