Semantic Communication for Cooperative Multi-Task Processing over Wireless Networks

Ahmad Halimi Razlighi, Carsten Bockelmann, and Armin Dekorsy
Department of Communications Engineering, University of Bremen, Germany
E-mails:{halimi, bockelmann, dekorsy}@ant.uni-bremen.de

***Abstract*—In this paper, we have expanded the current status of semantic communication limited to processing one task to a more general system that can handle multiple tasks concurrently. In pursuit of this, we first introduced our definition of the "semantic source", enabling the interpretation of multiple semantics based on a single observation. A semantic encoder design is then introduced, featuring the division of the encoder into a common unit and multiple specific units enabling cooperative multi-task processing. Simulation results demonstrate the effectiveness of the proposed semantic source and the system design. Our approach employs information maximization (infomax) and end-to-end design principles.**

***Index Terms*—Cooperative semantic communication, multi-task processing, infomax.**

I. Introduction

Applications involving machine-to-machine or human-to-machine communications often have to prioritize task execution over the exact reconstruction of transmitted information at the receiver. Unlike the traditional information theory established by Shannon, which emphasizes the accurate transmission and reception of bits, the design of communication systems for these applications takes a distinct approach, drawing attention to task performance rather than fidelity in information transmission. Thus, regarding the three levels of communication [1], technical layer (accurate transmission of symbols), semantic level (transmitting the desired meaning), and effectiveness level (effectiveness of the received meaning), one should investigate the second level of communication to meet the demands of emerging applications. Leveraging advancements in artificial intelligence, deep learning, and end-to-end (E2E) communication technologies, the concept of *semantic communication* has recently emerged [2]. Semantic communication prioritizes understanding the meaning and goals behind transmitted information, surpassing the traditional focus on the precise transmission of bits.

Four approaches to semantic communication are outlined in [3]. Firstly, the classical approach utilizes *logical* probability to quantify semantic information, primarily for text sources. Secondly, the knowledge graph (KG) approach represents semantics by KG structure. Thirdly, the machine learning (ML) approach leverages learned model parameters to represent semantics. Lastly, the significance approach emphasizes *timing* as semantics. Inspired by Weaver, an alternative approach extends Shannon's *statistical* probability (information theory) beyond the technical layer to the next two levels.

Works in semantic communication have been categorized into two types of research directions: data reconstruction and task execution. Data reconstruction is generally done by extracting semantic information on the transmitter side and recovering data with the received semantic information on the receiver side. This has been initially investigated by [4] and [5] for diverse sources following ML approach. Motivated by the mentioned works, papers like [6] and [7] have tried to explore communication efficiency and resource allocation issues in this research direction.

On the other hand, for task execution, also referred to as task-oriented communication or goal-oriented communication, [8] explored a communication scheme based on the information bottleneck (IB) framework, enabling information encoding for a single task while adapting to dynamic wireless channel conditions. The same authors in [9] studied distributed relevant information encoding for collaborative feature extraction to fulfill a task, leveraging distributed IB. Moreover, [10] offered a framework for collaborative retrieval of the message using multiple received semantic information and also expanded it using reinforcement learning in [11]. To consider some physical layer communication aspects, [12] contributed to resource allocation in a multi-user system according to the single task accuracy, channel conditions, and computing requests.

We aim to advance task-oriented research direction following the information theory perspective. Unlike existing works focusing on single-task processing, even in collaborative scenarios, this paper contributes to a more general approach capable of handling various tasks cooperatively by utilizing semantic wireless networks. Key contributions include:

- Introducing a semantic source utilizing probabilistic modeling and enabling different semantics extraction of a single observation.
- Proposing a semantic encoding structure employing neural networks (NNs), wherein the encoder is divided into a common unit (CU) and multiple specific units (SUs). This proposed design aims to process multiple tasks simultaneously.
- Demonstrating the proposed structure shows cooperation amongst SUs, improving their performance. This cooperation occurs through the shared CU, while SUs perform joint semantic and channel coding (JSCC) concurrently executing their individual tasks.

II. System Model

In this section, we initially introduce our probabilistic modeling of a semantic source. Subsequently, leveraging the

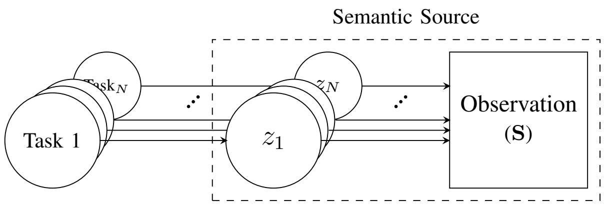

Fig. 1: Probabilistic graphical modeling of the proposed semantic source.

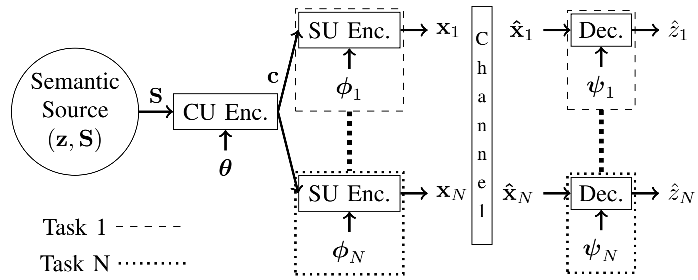

Fig. 2: Illustration of the proposed cooperative multi-task approach system model.

provided source model, we introduce our system model and formulate an optimization problem that addresses the execution of multiple tasks concurrently and cooperatively.

A. Semantic Source Modeling

We introduce our definition of *semantic source*, as shown in Fig. 1. We propose that there can be several semantics lying behind an observation shown by $\mathbf{S}$, and it is our tasks that specify one/multiple semantics to be our interest. In this paper, we assume the existence of N independent tasks, specifying N *semantic variables* indicated by $\mathbf{z} = [\, z_1 \, z_2 \, \ldots \, z_N \,]$. Having had the semantic variables defined and exploiting the probabilistic modeling, we define the tuple of $(\mathbf{z}, \mathbf{S})$ as our semantic source, fully described by the probability distribution of $p(\mathbf{z}, \mathbf{S})$. More precisely, we describe our semantic source by $p(\mathbf{z})p(\mathbf{S}|\mathbf{z})$, where $p(\mathbf{S}|\mathbf{z})$ is our *semantic channel* that reflects the semantic variables in our observation. Such a definition enables the simultaneous extraction of multiple semantic variables based on a single observation and addresses multiple tasks. For instance, consider an image featuring both a tree and a number. One task may entail determining the presence of a tree, resulting in a binary semantic variable. Meanwhile, another task could focus on identifying the number within the image, yielding a multinomial semantic variable, independent of the first one.

B. System Probabilistic Modeling

We argue that having the same observation, there can be some common relevant information useful for multiple semantic variables, and this can be shared among the SUs. Moreover, We argue this sharing might lead to better task performance, resulting in cooperative task-relevant information extraction. Thus, our idea is to split up the semantic encoder into a CU and SUs, introducing cooperative semantic communication to process multiple tasks. Our system model consists of a single observation and N independent semantic variables, each associated with a unique task. As illustrated in Fig. 2 initially, the CU encoder extracts the common relevant information from the semantic source. Then, N SU encoders extract and transmit task-specific information to their respective decoders. Output of SU encoders are shown as $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, and their noise-corrupted version received at the corresponding decoders are indicated by $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_N$. In our approach, we incorporate wireless transmission between encoders and decoders, employing the additive white Gaussian noise (AWGN) channel. Upon reception, semantic decoders deliver the semantic variables to their respective recipients. Specifically, the Markov representation of our system model for the i-th semantic variable is outlined as follows.

$$\begin{aligned} &p(\hat{z}_i, \hat{\mathbf{x}}_i, \mathbf{x}_i, \mathbf{c}|\mathbf{S}) = \\ &p^{\text{Dec}_i}(\hat{z}_i|\hat{\mathbf{x}}_i)\, p^{\text{Channel}}(\hat{\mathbf{x}}_i|\mathbf{x}_i)\, p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{c}) p^{\text{CU}}(\mathbf{c}|\mathbf{S}). \end{aligned} \tag{1}$$

In (1), $p^{\text{CU}}(\mathbf{c}|\mathbf{S})$ defines the CU that extracts the common relevant information amongst all tasks, from the observation. The i-th SU is described by $p^{\text{SU}_i}(\mathbf{x}_i|\mathbf{c})$ extracting task-specific information and provide $\mathbf{x}$ as the channel input. The corresponding decoder is then specified by $p^{\text{Dec}_i}(\hat{z}_i|\hat{\mathbf{x}}_i)$, where $\hat{\mathbf{x}}$ is the received information passed through the AWGN channel and modeled like $\hat{\mathbf{x}}_i = \mathbf{x}_i + \mathbf{n}$, where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}_m, \sigma_n^2 \mathbf{I}_m)$, and m is the size of the encoded task-specific information.

C. Optimization Problem

To design our split semantic encoder architecture, we formulate an optimization problem adopting the information maximization principle together with the E2E learning manner, which has been proven effective for task-oriented communication [13], as follows.

$$[\, p^{\text{CU}}(\mathbf{c}|\mathbf{s})^{\star},\, p^{\text{SU}}(\mathbf{x}|\mathbf{c})^{\star}] \; = \arg \max_{\substack{p^{\text{CU}}(\mathbf{c}|\mathbf{s}), \\ p^{\text{SU}}(\mathbf{x}|\mathbf{c}).}} \sum_{i=1}^{N} b_i \, I(\hat{\mathbf{x}}_i; z_i). \tag{2}$$

Thus, the objective is to maximize the mutual information between the channel output $\hat{\mathbf{x}}_i$, and the semantic variables z_i. In equation (2), b_i is a constant coefficient, representing a factor that will be fixed at one. This choice is made as we do not explore the relationship between semantic variables or prioritize them within the scope of this paper. As distributions, $p^{\text{CU}}(\mathbf{c}|\mathbf{s})$ and $p^{\text{SU}}(\mathbf{x}|\mathbf{c}) = [\, p^{\text{SU}_1}(\mathbf{x}_1|\mathbf{c}) \ldots p^{\text{SU}_N}(\mathbf{x}_N|\mathbf{c})]$, are completely unknown, we approximate them using NNs, resulting in $p^{\text{CU}}_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{s})$ and $p^{\text{SU}_i}_{\boldsymbol{\phi}_i}(\mathbf{x}_i|\mathbf{c})$, where $\boldsymbol{\theta}$ represents the NN's parameters approximating the CU and $\boldsymbol{\phi}_i$ is the parameters of the i-th NNs approximating the i-th SU. Consequently, we formulate the approximated objective function as follows.

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\theta}, \underline{\boldsymbol{\phi}}) = \sum_{i=1}^{N} I(\hat{\mathbf{x}}_i; z_i) \\ &\approx \mathbb{E}_{p^{\text{CU}}_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{s})} \left[\sum_{i=1}^{N} \left\{ \mathbb{E}_{p(\mathbf{S}, z_i)} \left[\mathbb{E}_{p^{\text{SU}_i}_{\boldsymbol{\phi}_i}(\hat{\mathbf{x}}_i|\mathbf{c})} [\, \log p(z_i|\hat{\mathbf{x}}_i)] \, \right] \right\} \right]. \end{aligned} \tag{3}$$

Details on the objective function's lower bound derivation are deferred to Appendix A. As shown in (3), by considering the channel outputs we aim to emphasize the role of joint semantic and channel coding performed by our SUs. That is why employing the fact that $p^{\text{SU}_i}_{\boldsymbol{\phi}_i}(\hat{\mathbf{x}}_i|\mathbf{c}) = \int p^{\text{SU}_i}_{\boldsymbol{\phi}_i}(\mathbf{x}_i|\mathbf{c})\, p^{\text{Channel}}(\hat{\mathbf{x}}_i|\mathbf{x}_i)\, d\mathbf{x}_i$, we try to optimize $p^{\text{SU}_i}_{\boldsymbol{\phi}_i}(\hat{\mathbf{x}}_i|\mathbf{c})$ and explicitly show how JSCC is performed. Moreover, (3) shows our adaptation of E2E design, where we jointly optimize the encoders and decoders. Including the AWGN channel directly in our E2E design works well as its transfer function is differentiable. Moreover, the outer expectation, appearing due to our structure, highlights the difference between our approach and others used for single-task processing and includes cooperation amongst the SU blocks.

Regarding the i-th decoder in (3), the $p^{\text{Dec}_i}(\hat{z}_i|\hat{\mathbf{x}}_i)$ can be fully determined using the known distributions and underlying probabilistic relationship in (1) as:

$$p^{\text{Dec}_i}(\hat{z}_i|\hat{\mathbf{x}}_i) = \frac{\int p^{\text{SU}_i}_{\boldsymbol{\phi}_i}(\hat{\mathbf{x}}_i|\mathbf{c})\, p^{\text{CU}}_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{S})\, p(\mathbf{S}, z_i)\, d\mathbf{s}\, d\mathbf{c}}{p(\hat{\mathbf{x}}_i)}. \quad (4)$$

However, due to the high-dimensional integrals, (4) becomes intractable and we need to follow the variational approximation technique [14], resulting in:

$$\mathcal{L}(\boldsymbol{\theta}, \underline{\boldsymbol{\phi}}, \underline{\boldsymbol{\psi}}) \approx \mathbb{E}_{p^{\text{CU}}_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{S})}\left[\sum_{i=1}^{N}\left\{\mathbb{E}_{p(\mathbf{S},z_i)}\left[\mathbb{E}_{p^{\text{SU}_i}_{\boldsymbol{\phi}_i}(\hat{\mathbf{x}}_i|\mathbf{c})}\left[\log q^{\text{Dec}_i}_{\boldsymbol{\psi}_i}(z_i|\hat{\mathbf{x}}_i)\right]\right]\right\}\right]. \quad (5)$$

Where in (5), $\boldsymbol{\psi}_i$ represents the i-th NN approximating the true distribution of the i-th decoder. To obtain the empirical estimate of the above objective function, we approximate the expectations using Monte Carlo sampling assuming the existence of a dataset $\{\mathbf{S}^{(j)}, z_1^{(j)}, \ldots, z_N^{(j)}\}_{j=1}^{J}$ where J represents the batch size of the dataset.

$$\mathcal{L}(\boldsymbol{\theta}, \underline{\boldsymbol{\phi}}, \underline{\boldsymbol{\psi}}) \approx \frac{1}{L}\sum_{l=1}^{L}\left[\sum_{i=1}^{N}\left\{\frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{K}\sum_{k=1}^{K}\left[\log q^{\text{Dec}_i}_{\boldsymbol{\psi}_i}(\hat{z}_i|\hat{\mathbf{x}}_{j,k})\right]\right]\right\}\right]. \quad (6)$$

To overcome the differentiability issues in (6), we have adopted the reparameterization trick [14], introducing $\mathbf{c}_{j,l} = \boldsymbol{\mu}_{\mathbf{c}_j} + \boldsymbol{\sigma}_{\mathbf{c}_j} \odot \boldsymbol{\epsilon}_{j,l}$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The issue of differentiability regarding $\boldsymbol{\phi}$ does not arise, as we presume a deterministic process occurring for $\mathbf{x}_j$, with noise sampling taking place for $\hat{\mathbf{x}}_{j,k} = \mathbf{x}_j + \mathbf{n}_k$. Thus, we fix the sample size of the reparameterization trick, L, and the channel sampling size, K, to one for each batch. Details on how the lower bound objective function is differentiable with respect to all parameters are deferred to Appendix B.

III. Simulation Results

To demonstrate the effectiveness of the proposed architecture, we use the MNIST dataset of handwritten digits [15], containing 60,000 images for the training set and 10,000 samples for the test set. For a specific number of tasks denoted by N, we shape our semantic source as stated before like $\{\mathbf{S}^{(j)}, z_1^{(j)}, \ldots, z_N^{(j)}\}_{j=1}^{J}$. This involves pre-processing the MNIST dataset by assigning multiple labels to each data sample $\mathbf{S}^{(j)}$. Thus, in this setup labels stand for our semantic variables. For our evaluations, we consider the execution of two tasks, binary classification (Task1) and categorical classification (Task2). Therefore, two semantic variables, $z_1 \sim \mathit{Bernoulli}$ and $z_2 \sim \mathit{Multinomial}$ will represent Task1 and Task2 respectively. Our experiments consider the classification of digit "2" as Task1 and digit identification for Task2. The implemented NN structure is described in Table I and found heuristically. Regarding the objective function in (6), the $\sum_{i=1}^{N}$ forces us to feed the dataset in training phase like $\{\mathbf{s}^{(j)}, z_n^{(j)}\}_{j=1}^{J}$. Then, the outer summation will try to capture the learned features across different tasks in the CU, performing a joint training procedure.

TABLE I: The NN structure for MNIST dataset.

	Layer	**Output size**
CU	Fully-connected (FC) + Tanh, FC	64
SU	(SU_1) FC + Tanh, FC	16
	(SU_2) FC + Tanh, FC	16
Dec	(Dec_1) FC + Tanh, FC + Sigmoid	1
	(Dec_2) FC + Tanh, FC + Softmax	10

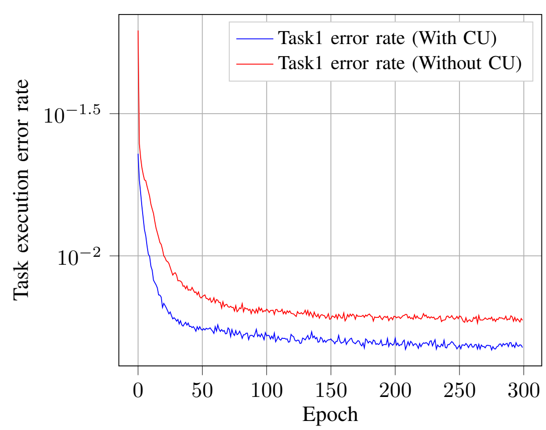

Fig. 3: Impact of the CU on task execution for Task1.

Fig. 3, compares the proposed approach (with CU) with the conventional single-task semantic communication (without CU) for Task1. It shows that our approach outperforms the conventional one resulting in a lower task execution error rate and faster improvement. The same holds for Task2, as shown in Fig. 4. We kept the structure of the without-CU case, the same as our SUs and respective decoders for fair comparison.

To illustrate that cooperation through the CU may not always be beneficial, Fig. 5, contrasts two scenarios: case1 and case2. The proposed approach demonstrates constructive outcomes in case1, involving two categorical classifications. However, in case 2, where both tasks are binary classifications (Task1: "2" and Task2: "4"), destructive cooperation is evident.

We interpret the findings in Fig 5 through the Kullback-Leibler (KL) divergence of our semantics distributions, $D_{KL}\left(p(z_1|\mathbf{S}) \parallel p(z_2|\mathbf{S})\right)$. In case1, semantic variables are maximally informative about each other, and constructive cooperation is observed. The converse is observed in case2 where

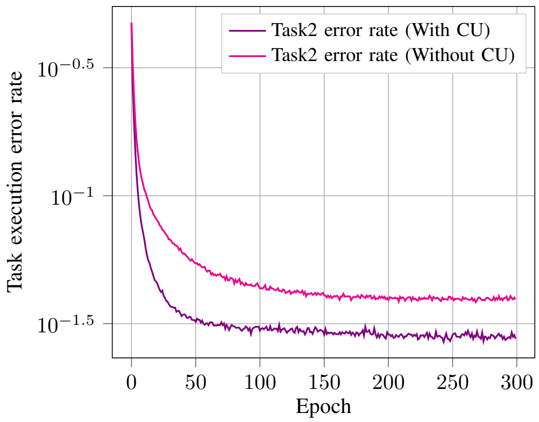

Fig. 4: Impact of the CU on task execution for Task2.

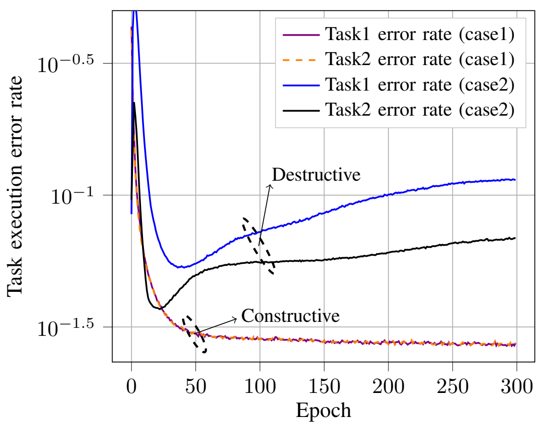

Fig. 5: Constructive and destructive behaviors.

they are less informative. Generally, increased KL divergence corresponds to constructive cooperation, while decreased KL divergence signifies a shift towards destructive behavior.

IV. Conclusion

In conclusion, we introduced a semantic encoding scheme, dividing the encoder into CU and SUs, enabling cooperative semantic communication and the simultaneous process of multiple tasks. We also introduced our semantic source definition, allowing the extraction of diverse semantic variables from a single observation. Future research directions involve optimizing the clustering of SUs to improve cooperative performance and exploring the integration of new SUs into the architecture.

Appendix A
Derivation of the Lower Bound

Using the approximated distributions for (2):

$$\mathcal{L}(\boldsymbol{\theta}, \underline{\boldsymbol{\phi}}) \approx \sum_{i=1}^{N} \int p(z_i, \mathbf{S})\, p_{\boldsymbol{\theta}}^{\text{CU}}(\mathbf{c}|\mathbf{S})$$

$$p_{\boldsymbol{\phi}_i}^{\text{SU}_i}(\mathbf{x}_i|\mathbf{c}) p^{\text{Channel}}(\hat{\mathbf{x}}_i|\mathbf{x}_i) \log p(z_i|\hat{\mathbf{x}}_i)\, dz_i\, d\mathbf{s}\, d\mathbf{c}\, d\mathbf{x}_i\, d\hat{\mathbf{x}}_i$$

$$\approx \sum_{i=1}^{N} \int p(z_i, \mathbf{S})\, p_{\boldsymbol{\theta}}^{\text{CU}}(\mathbf{c}|\mathbf{S}) p_{\boldsymbol{\phi}_i}^{\text{SU}_i}(\hat{\mathbf{x}}_i|\mathbf{c}) \log p(z_i|\hat{\mathbf{x}}_i)\, dz_i\, d\mathbf{s}\, d\mathbf{c}\, d\hat{\mathbf{x}}_i$$

Appendix B
Differentiability of the Lower Bound

Considering the lower bound (5):

$$\mathbb{E}_{p_{\boldsymbol{\theta}}^{\text{CU}}(\mathbf{c}|\mathbf{s})}\left[\sum_{i=1}^{N}\left\{\mathbb{E}_{p(\mathbf{s},z_i)}\left[\mathbb{E}_{p_{\boldsymbol{\phi}_i}^{\text{SU}_i}(\hat{\mathbf{x}}_i|\mathbf{c})}\left[f(z_i)\right]\right]\right\}\right]$$

We know that $z_i = g(\hat{\mathbf{x}}_i, \boldsymbol{\psi}_i)$, $\hat{\mathbf{x}}_i = h(\mathbf{c}, \boldsymbol{\phi}_i, \mathbf{n})$, and $\mathbf{c} = u(\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\epsilon})$ where $\mathbf{n}$ in $h(\cdot)$ and $\boldsymbol{\epsilon}$ in $u(\cdot)$ solve the differentiability issues with $\nabla_{\boldsymbol{\theta}} \mathbb{E}_{p_{\boldsymbol{\theta}}^{\text{CU}}(\mathbf{c}|\mathbf{s})}[\,\cdot\,]$ and $\nabla_{\boldsymbol{\phi}} \mathbb{E}_{p_{\boldsymbol{\phi}}^{\text{SU}}(\hat{\mathbf{x}}|\mathbf{c})}[\,\cdot\,]$. Employing the chain rule derivative for the lower bound ensures its differentiability with respect to all parameters, as exemplified below for $\underline{\boldsymbol{\phi}}$.

$$\frac{\mathcal{L}(\boldsymbol{\theta}, \underline{\boldsymbol{\phi}}, \underline{\boldsymbol{\psi}})}{\underline{\boldsymbol{\phi}}} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial h} \cdot \frac{\partial h}{\partial \underline{\boldsymbol{\phi}}}$$

References

[1] M. Sana and E. C. Strinati, "Learning semantics: An opportunity for effective 6G communications." Institute of Electrical and Electronics Engineers Inc., 2022, pp. 631–636.

[2] "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, pp. 5–41, 11 2022.

[3] D. Wheeler and B. Natarajan, "Engineering semantic communication: A survey," *IEEE Access*, vol. 11, pp. 13 965–13 995, 2023.

[4] H. Xie, Z. Qin, G. Y. Li, and B. H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[5] H. Xie and Z. Qin, "A lite distributed semantic communication system for internet of things," *IEEE Journal on Selected Areas in Communications*, vol. 39, pp. 142–153, 1 2021.

[6] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," *IEEE Wireless Communications Letters*, vol. 11, pp. 1394–1398, 7 2022.

[7] H. Tong, Z. Yang, S. Wang, Y. Hu, W. Saad, and C. Yin, "Federated learning based audio semantic communication over wireless networks." Institute of Electrical and Electronics Engineers Inc., 2021.

[8] 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, 2022.

[9] J. Shao, Y. Mao, and J. Zhang, "Task-oriented communication for multidevice cooperative edge inference," *IEEE Transactions on Wireless Communications*, vol. 22, no. 1, pp. 73–87, 2023.

[10] E. Beck, C. Bockelmann, and A. Dekorsy, "Semantic information recovery in wireless networks," *Sensors*, vol. 23, p. 6347, 7 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/14/6347

[11] E. Beck, C. Bockelmann, and A. Dekorsy, "Model-free reinforcement learning of semantic communication by stochastic policy gradient," in *IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN 2024)*, vol. 1. IEEE, May 2024, p. 7. [Online]. Available: https://icmlcn2024.ieee-icmlcn.org/

[12] F. Binucci, P. Banelli, P. Di Lorenzo, and S. Barbarossa, "Dynamic resource allocation for multi-user goal-oriented communications at the wireless edge," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 697–701.

[13] J. Shao and J. Zhang, "Bottlenet++: An end-to-end approach for feature compression in device-edge co-inference systems," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2020, pp. 1–6.

[14] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *CoRR*, vol. abs/1906.02691, 2019. [Online]. Available: http://arxiv.org/abs/1906.02691

[15] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.