Disentangling Learnable and Memorizable Data via Contrastive Learning for Semantic Communications

Christina Chaccour, *Graduate Student Member, IEEE,* and Walid Saad, *Fellow, IEEE,*
Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA USA.

***Abstract*—Achieving artificially intelligent-native wireless networks is necessary for the operation of future 6G applications such as the metaverse. Nonetheless, current communication schemes are, at heart, a mere reconstruction process that lacks reasoning. One key solution that enables evolving wireless communication to a human-like conversation is semantic communications. In this paper, a novel machine reasoning framework is proposed to pre-process and disentangle source data so as to make it *semantic-ready*. In particular, a novel contrastive learning framework is proposed, whereby instance and cluster discrimination are performed on the data. These two tasks enable increasing the cohesiveness between data points mapping to semantically similar content elements and disentangling data points of semantically different content elements. Subsequently, the semantic deep clusters formed are ranked according to their level of confidence. Deep semantic clusters of highest confidence are considered *learnable, semantic-rich* data, i.e., data that can be used to build a language in a semantic communications system. The least confident ones are considered, *random, semantic-poor, and memorizable data* that must be transmitted classically. Our simulation results showcase the superiority of our contrastive learning approach in terms of semantic impact and minimalism. In fact, the length of the semantic representation achieved is minimized by 57.22% compared to vanilla semantic communication systems, thus achieving minimalist semantic representations.**

***Index Terms*— Semantic communications, contrastive learning, semantic language, AI-Native, 6G, beyond 6G.**

I. Introduction

To enable emerging services like the metaverse, digital twins, Web 3.0, and Industry 5.0 [1] over wireless systems, such as 6G and beyond, there is a need for a fundamentally novel approach to performing communication in such systems. In particular, instead of relying on classical data-centric artificial intelligence (AI) techniques that limit wireless network generalizability and autonomy we must move towards *generalizable, knowledge-based, and reasoning-driven* wireless networks [2], through the concept of semantic communications. Semantic communications is a communication paradigm that involves transforming radio nodes into intelligent, reasoning agents that can unravel semantic content elements (meaning) and can communicate a *minimal, efficient, and generalizable* representation that expresses the aforementioned meaning. In semantic communications the transmitter-receiver pair become a *teacher-apprentice* that exchange a *semantic language* rather than acting as a mere bit-pipeline.

Hence, semantic communication is the path to AI-native wireless networks [2]. Nonetheless, building a mature semantic language that can express structure, minimally and efficiently is a very challenging task. Low-layer source data tends to be very entangled and its structure is not directly interpretable as random information overlay its meaningful aspects. Thus, identifying the semantic content elements in a datastream and describing them via a *minimal, efficient, and generalizable* representation becomes a very complex learning task and would lead to an overly complex semantic language. This complexity is not a result of the content complexity, but rather a consequence of attempting to learn the semantic content elements of possibly spurious random data points. It is thus necessary to devise a fully-fledged mechanism that enables disentangling *semantic rich data* from *spurious random information* in raw source data. This is a fundamental step in communicating a language that *captures meaningful structure* and leads to a minimalist and efficient semantic communication system. Essentially, semantic communication systems cannot be robust, nor tend to generalizability, if their languages capture spurious information in the raw data.

A. Prior Works

Recently, a number of works [3]–[7] studied the concept of semantic communication systems. In [3], a deep learning based semantic communication system for text transmission was devised. The work in [4] proposes deep learning approach to perform semantic communication for speech signals. The authors in [5] introduce a comprehensive semantic communications framework for enabling goal-oriented task execution. In [6], an emergent semantic communication system framework via a signaling game and a neuro-symbolic AI approach is proposed. The authors in [7] explored task-oriented multi-user semantic communications to transmit data with single-modality and multiple modalities. While the works in [3]–[7] are interesting, they fail to acknowledge the entangled and intertwined nature of raw data. Furthermore, although the works in [3]–[7] study a language, yet they fail to distinguish the characteristics of a semantic language as well as the AI and computational approaches needed to build a language from raw entangled data. Also, the works [3]–[7] do not take into account the fact that *semantic-poor* data points are a complex learnable task, but a simple *memorization* task, and thus are more efficiently transmitted classically. Finally, none of the works in [3]–[7] propose a fully fledged framework to disentangle raw data into meaningful semantic content elements that can be used to build a semantic language. Clearly, there is a gap in the research community to build semantic communication systems that can leverage raw data so as to build a language and transmit knowledge.

This research was supported by the U.S. National Science Foundation under Grants CNS-2007635.

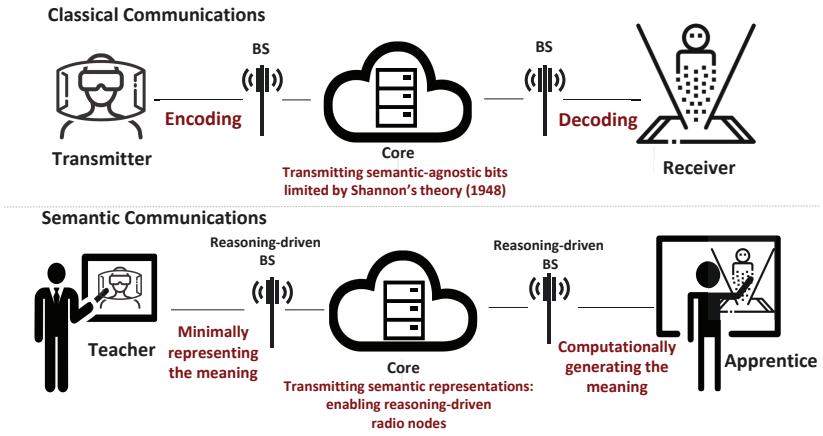

Fig. 1: Illustrative figure showcasing end-to-end (E2E) classical communication systems versus E2E semantic communication systems.

B. Contributions

The main contribution of this article is a novel contrastive learning approach to pre-process and disentangle raw data making it semantic-ready. In particular, a contrastive learning scheme is adopted at the transmitter side so as to disentangle raw source data into deep semantic clusters. This deep semantic clustering enables the separation of *learnable data* from *memorizable data* as well as data points belonging to different semantic content elements. In essence, *learnable data* are data points with *rich semantic information*, i.e., such data points can be expressed via a representation that has a meaningful semantic content element. Meanwhile, *memorizable data points* are random data points that do not necessarily map to a particular significance, and are thus *semantically poor*. Essentially, first, an instance discrimination contrastive learning task is performed whereby data points are compared to their neighbours at the instance-level. That is, instances undertake a perturbation which mimics a *distribution shift*, then the teacher learns whether such data points are semantically close or not. Second, building on this instance discrimination task, a deep cluster discrimination task is performed so as to increase the cohesiveness between data points that belong to the same semantic content element, and disentangle semantic content elements of different meaning. The deep contrastive-semantic clusters built are further ranked per confidence level so as to label the least confident, as semantic poor data points, i.e., memorizable data that must be classically transmitted. *To the best of our knowledge this is the first work that pre-processes raw data via contrastive learning so as to disentangle semantic rich data points from semantic poor ones. This is also the first work that enables increasing the cohesiveness between semantically similar data points, and increases the disentanglement between semantically different raw data points.* Simulation results demonstrate that our contrastive learning disentanglement method enables realizing a large semantic impact ($71.9\,\%$ higher than conventional semantic communication systems) despite increases in the content complexity. Our simulation results also show that the proposed contrastive learning approach yields representations of smaller length and thus enables a minimalist semantic communication system.

II. System Model and Problem Formulation

A. Semantic Communication Model

Consider a teacher b, who observes source information, which can stem from any use case (e.g. hologram, tactile feedback, image data) and is willing to transmit knowledge contained in the raw, low-layer, source datastream $\boldsymbol{X}$ to an apprentice d, as shown in Fig. 1.

1) *At the reasoning-driven transmitter side:* The goal of the teacher is to first *disentangle* the semantic content elements $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$, ($n$ is the number of semantic content elements in the source information) contained in the datastream $\boldsymbol{X}$, i.e. separate the different meaning elements contained in the source data. Then, for every content element the teacher must learn a corresponding *semantic representation* $\boldsymbol{Z} = \{Z_1, \ldots, Z_n\}$ with desirable properties. Essentially, $\boldsymbol{Z}$ is the minimal, efficient, and generalizable description of $\boldsymbol{Y}$.
2) *At the reasoning-driven receiver side:* The goal of the apprentice is to *understand* the meaning contained in the semantic representation Z_i and use it to computationally generate semantic content elements Y_i.

To efficiently and effectively transmit semantic information between a teacher and an apprentice, it is necessary to devise a semantic language, that transforms the transmitter-receiver bit-pipeline into a structured information transfer. As we discussed in [2], we can define a semantic language as follows:

Definition 1. *A* semantic language $\mathcal{L} = (X_i, Z_i)$, *(or* $\mathcal{L} = (X_{l,i}, Z_i)$*) is a dictionary (from a data structure perspective) that maps the source raw datastream (or an interventional learnable* $X_{l,i}$*) to their corresponding semantic representation* Z_i*, based on the identified semantic content elements* Y_i*.*

B. Semantic Language Model

In classical communications, the goal is to minimize the system entropy and maximize the channel capacity. In essence,

statistical entropy characterizes the number of "yes/no" questions we would have to ask in order to get complete information on the datastream we are dealing with. Entropy cannot be used in semantic communication systems as it solely depends on the freedom of choice of the transmitter rather than the meaning of the message. Thus, in a semantic communication system, the equivalent of "entropy" is the "language complexity" as discussed in [2]. Henceforth, in a semantic communication system, the goal of the complexity of the language is to characterize the difficulty of identifying and learning the semantic content elements in the raw datasteram $\boldsymbol{X}$. It is given by [2]:

$$\Gamma(\mathcal{L}) = \min_{p(\boldsymbol{Z}|\boldsymbol{X})} L_{\mathcal{L}}(p) + K(p), \tag{1}$$

where, $L_{\mathcal{L}}(p) = \sum_{i=1}^{N} -\log p(Z_i|X_i)$ is the cross-entropy loss, and $K(p)$ is the Kolmogorov complexity of the distribution $p(\boldsymbol{Z}|\boldsymbol{X})$. From (1), one can observe:

- The language complexity, unlike Shannon's information-theoretic perspective, is a metric that depends on the meaning of representations. (1) is a function of: a) the fidelity of the representation model in capturing the semantic content elements in the datastream, and b) the Kolmogorov complexity that characterizes the individuality of the semantic content elements as well as capturing the shortest effective binary description of X_i [2].
- If the majority of the data lacks *structure*, the *learning task becomes increasingly difficult*. Nonetheless, this is not necessarily a result of complex content, but rather of a high number of random data points that lack semantic significance. In this setting, such data points must be *memorized rather than learned*. That is, the semantic communication system must retain them as is and classically transmit them.

Furthermore, (1) is an ideal theoretical equation that measures complexity in terms of the idealized, shortest, and best obtainable model that can represent the semantic content elements in the data. The Kolmogorov complexity term in (1) is not computable, i.e., there is no single function that will return the complexity of a particular representation binary string. Thus, for tractability and to capture a realistic representation model, based on fundamentals from transfer learning [8], we can rewrite the language complexity as follows:

$$\Gamma(\mathcal{L}, \zeta, \Lambda) = \mathbb{E}_{\theta\ \Lambda(\theta|\mathcal{L})}\{L_{\mathcal{L}}(p_\theta(Z_i|X_i)\} + \beta KL(\Lambda(\theta|\mathcal{L})||\zeta(\theta),$$

where the second term $KL(\Lambda(\theta|\mathcal{L})||\zeta(\theta)$ is the KL divergence that characterizes the information in the parameters of the representation model. Moreover, ζ and Λ are, respectively, the pre-distribution and post-distribution of the representation model. Such distributions do not correspond to a prior or posterior of a Bayesian setting [8]. A pre-distribution is used to initialize the semantic representation model so as to initiate the gradual learning of a semantic language. Meanwhile the post-distribution is used to fine-tune the language acquired between the teacher and the apprentice. Both of these distributions depend on the level of symmetry between teacher and apprentice and their knowledge bases[1].

C. Problem Formulation

As previously discussed, a language that lacks *structure* will have a very large complexity due to the large amount of spurious random data points captured via representations. Remarkably, raw data tends to be very entangled, and it contains a large amount of random information. Hence, relying on the raw datastream $\boldsymbol{X}$ will lead to: a) an overly complex language with representations that cannot express meaningful semantic content elements, and b) such a problem mimicks the overfitting machine learning (ML) scenario whereby the radio nodes have not learned representations but merely *memorized* spurious patterns in the source information. To alleviate this problem, it is necessary to devise a comprehensive method that enables disentangling the learnable $\boldsymbol{X}_l$ data points from the memorizable ones $\boldsymbol{X}_m$. Performing such disentanglement technique will allow the radio node to:

- Use $\boldsymbol{X}_l$ as their structure-rich datastream to build a semantic language, and learn efficient and minimal semantic representations. Thus, $\boldsymbol{X}_l$ will be transmitted semantically.
- Transmit $\boldsymbol{X}_m$ classically due to the lack of structure and rich semantic content elements in these data points. Attempting to perform reasoning or learning on $\boldsymbol{X}_m$ is a computationally inefficient scheme. Meanwhile, the classical communication scheme (which essentially targets characterizing the uncertainty in information) remains very efficient as it is not concerned with meaning.

The goal of this work is to propose a novel contrastive learning technique that enables pre-processing and disentangling raw source data, to ultimately separate learnable and memorizable data via instance and cluster discrimination. This approach also increases the cohesiveness between semantically similar data points and disentangling semantic different ones. Next, we delve into the analytical foundation of our proposed contrastive learning framework.

III. Contrastive Learning for Disentangling Semantic Content Elements

Given a set of unlabeled[2] datastreams, we let κ be the feature embedding network that extracts key structural information encoded in the low-dimensional vector subspace $\psi_\kappa : \boldsymbol{X} \rightarrow \boldsymbol{Z} \in \mathbb{R}^N$. Furthermore, we let $\phi_\varkappa$ be a classifying function that associates every semantic representation with its ground-truth semantic content element pseudo-label $\phi_\varkappa : \boldsymbol{Z} \rightarrow \boldsymbol{Y}$, where $\boldsymbol{Y} \in \mathcal{Y}$, $\mathcal{Y}$ is the universal set of possible semantic content elements that can be communicated. Essentially, the goal is that representations of the same cluster must share similar pseudo-labels vis-à-vis the semantic content elements they

[1]Learning and optimizing such distributions is an important research avenue that is outside the scope of this paper.

[2]Unlabeled here refers to datastreams with unknown semantic content elements.

are representing. Next, we discuss how we perform instance discrimination to improve the generalizability of the representations that should be learned from the raw source data. Then, cluster discrimination is performed among the data points in $\boldsymbol{X}$ to optimize the boundaries between different semantic content elements and ultimately separate memorizable data. In essence, the higher confidence clusters are *semantic rich* structures and are *learnable* clusters, meanwhile lower confidence clusters are *semantic poor* clusters with memorizable properties that must be classically transmitted. For instance, semantic-rich data can be thought as "meaningful structural elements" of a hologram, meanwhile memorizable data may be random details in a hologram that do not have a meaning or structure (e.g. information that human beings would not notice about that hologram if trying to describe it). Furthermore, the instance discrimination technique performs this clustering while making sure that the mapping κ leads to minimalist representations.

A. Improving the Robustness of Semantic Representations via Instance Discrimination

To perform instance discrimination [9], we first consider every mapping function ψ_κ, and we construct another momentum encoder function[3] $\psi_{\tilde{\kappa}}$ that shares an identical structure but independent parameters $\tilde{\kappa}$. Furthermore, for every raw datastream in $\boldsymbol{X}$, we randomly apply a set of transformations $\mathcal{X}$ so as to mimic future distribution changes, and improve the robustness of reasoning with respect to vertical generalizability (distribution shifts). For each instance of the datastream X_i, we represent two perturbed copies, namely, $\chi_1(X_i)$ and $\chi_2(X_i)$. As such, after passing such perturbed copies with the momentum encoder, we obtain $\boldsymbol{a}_i = \psi_\kappa(\chi_1(X_i))$ and $\boldsymbol{b}_i = \phi_\varkappa(\chi_2(X_i))$. In the instance discrimination, the contrastive learning task must match $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$, against the contrastive set, composed by K stale representations of the pseudo-negative samples of the semantic content element pseudo-labels [10]:

$$\tilde{\mathcal{B}}_i = \{\tilde{\boldsymbol{b}} | \tilde{\boldsymbol{b}} \in C_l \ \ \forall l \in [1, M] \text{ and } \ l \neq i\} = \{\tilde{\boldsymbol{b}}_1, \ldots, \tilde{\boldsymbol{b}_K}\},$$

where C_l is the memory bank used to manage the semantic content element and its respective cluster, and M is the possible number of semantic content elements that might occur for the observed data stream. The loss resulting from the instance discrimination step is given by [10]–[12]:

$$L_I(\boldsymbol{X}_i) = -\log \frac{\exp(\cos(\boldsymbol{a}_i, \boldsymbol{b}_i)/\tau)}{\sum_{\tilde{\boldsymbol{b}} \in \tilde{\boldsymbol{B}} \cup \{\boldsymbol{b}_i\}} \exp(\cos(\boldsymbol{a}_i, \tilde{\boldsymbol{b}})/\tau)}. \quad (2)$$

Next, we contrast random information with respect to clusters containing semantically similar content elements.

B. Disentangling Semantic Content Elements via Semantic Cluster Discrimination

Essentially, data points that map to a single semantic content element (and its representation) must belong to a single semantic cluster free from any random information. Thus, in order to discover the random information existing within semantically similar clusters, we investigate the decision boundaries between such clusters. Then, we adopt an approach that enables closing the gap between semantically similar content elements, and eliminating random information from semantic rich clusters. In essence, cluster discrimination increases the compactness between intra-content data points, and increases the separation between inter-content data points, learnable, and memorizable data. Thus, given a sample $\boldsymbol{a_i}$, its probability of being in a semantically similar cluster is given by [10]–[12]:

$$P_{i,l} = \frac{\sum_{\boldsymbol{b} \in C_l} \exp(\cos(\boldsymbol{a}_i, \tilde{\boldsymbol{b}}))/\tau}{\sum_{l'=1}^{N} \sum_{\tilde{\boldsymbol{b}} \in C_{l'}} \exp(\cos(\boldsymbol{a}_i, \tilde{\boldsymbol{b}})/\tau)}, \quad (3)$$

Then, we formulate the cluster discrimination loss as:

$$L_D = \frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{M} -P_{i,l} \log P_{i,l}. \quad (4)$$

Thus, the total loss will be formulated as [10]–[12]: $L_T = \eta L_I + \epsilon L_D$, where we can set $\eta = \epsilon = 1$ in case we want to minimize the exhaustive per-dataset parameter tuning. To minimize this loss, the weights κ and the decision boundaries of the semantic clusters are updated by back-propagation. Meanwhile, the momentum encoder $\tilde{\kappa}$ is updated via $\tilde{\kappa} \leftarrow \omega\tilde{\kappa} + (1-\omega)\kappa$, where ω is the momentum encoder coefficient. To distinguish $\boldsymbol{X}_l$ from $\boldsymbol{X}_m$, the deep semantic clusters are ranked from highest confidence to the lowest confidence. Thus, the distinction is not binary: the higher the level of confidence, the more *learnable and semantic rich* those data points are. Meanwhile, semantic deep clusters with the lowest level of confidence contain the data points that are the most *memorizable and semantic poor*. Categorizing particular semantic deep clusters and their corresponding data points as $\boldsymbol{X}_m$ is a function of the lowest tolerable level of confidence. This is based on the tolerable language complexity in (1), the computation capability of the system, and the content complexity of the source data.

IV. Simulation Results and Analysis

A. Performance Evaluation via Semantic Key Performance Indicators (KPIs)

Given that semantic communication systems are knowledge-centric, and thus leverage communication and computing resources; it is necessary to evaluate the performance of such systems using novel metrics. As such, in our simulation results, to evaluate the superiority of disentangling learnable and memorizable data via contrastive learning, we rely on the notion of *semantic impact*. Essentially, as we have defined it in [2], the semantic impact ι_{Y_i} measures the *equivalent computational* packets generated via the transmission of a semantic representation. In other words, it measures the number of packets that a semantic representation can computationally generate, and thus highlights the amount of communication resources and time that could be saved when deploying semantic systems.

[3]In this work, we consider the momentum contrast as our scheme for instantiation and instance discrimination.

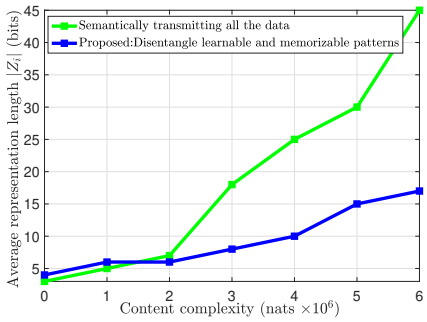

Fig. 2: Average representation length $|Z_i|$ (bits) versus content complexity (nats).

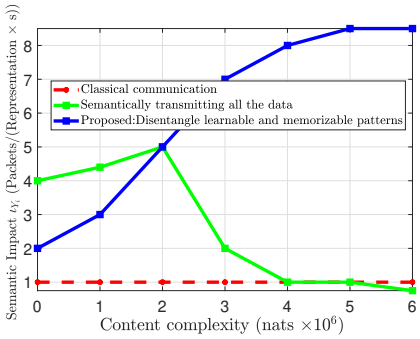

Fig. 3: Semantic impact ι_{Y_i} (Packets/(Representation $\times$s)) versus content complexity (nats).

B. Simulation Results

For our simulations, we consider $\boldsymbol{X}$ to be the low-level binary equivalent of a mixture model of CIFAR-10(/100) [13] and ImageNet-10/Dogs [14]. The simulations adopted the stochastic gradient descent (SGD) optimizer, $\omega = 0.9$, and $\tau = 0.1$. The learning rate was set to 0.02 across $200\,\text{epochs}$ and a batch size of 32 [15].

In Fig. 2, we can see that the average representation length increases with the content complexity. In fact, we can see that, for a low content complexity, semantically transmitting all the data might result in a smaller representation length. This is because the amount of random information, and thus $\boldsymbol{X}_m$ is considerably small. Meanwhile, as we increase the content complexity, we can see that semantically transmitting all the data is not a feasible approach anymore as the representation length steeply increases as we increase the content complexity. In contrast, when adopting our contrastive learning approach to pre-process the data, we can see that the average representation length is contained and minimalism can be achieved in the representation used, even for higher complexity orders. In fact, our representation is minimized by 57.22% compared to the vanilla semantic approach.

In Fig. 3, we compare the semantic impact ι_{Y_i} for classical communications, fully transmitting all data via semantic communications, and adopting our proposed contrastive learning approach to disentangle $\boldsymbol{X}_l$ and $\boldsymbol{X}_m$. We can see that while for a low content complexity, semantically transmitting all the data might be a feasible solution, however, as we increase the content complexity we can see how the semantic impact drastically decreases. Meanwhile, when contrastive learning is adopted to pre-process the language, we can see that the semantic impact grows despite increases in the content complexity. In fact, a 71.9% improvement is achieved compared to a semantic communication system that lacks language pre-processing.

V. Conclusion

In this paper, we have proposed a novel contrastive learning approach to pre-process and disentangle the raw data used in semantic communication systems. In particular, our contrastive learning approach performs instance and cluster discrimination on raw data points. In essence, we increase the cohesiveness between data points that belong semantically similar content elements and disentangle data points belonging to semantically different content elements. Subsequently, we rank the deep semantic clusters and consider the least confident ones as memorizable, semantic-poor data. Simulation results demonstrate the superiority of our contrastive learning approach compared to pure semantic and classical communications vis-à-vis semantic impact and minmalism of the representation.

References

[1] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, P. Popovski, and M. Debbah, "Seven defining features of terahertz (THz) wireless systems: A fellowship of communication and sensing," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 967–993, Jan. 2022.

[2] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. V. Poor, "Less data, more knowledge: Building next generation semantic communication networks," *arXiv preprint arXiv:2211.14343*, 2022.

[3] 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, April 2021.

[4] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, Jun. 2021.

[5] M. K. Farshbafan, W. Saad, and M. Debbah, "Curriculum learning for goal-oriented semantic communications with a common language," *arXiv preprint arXiv:2204.10429*, 2022.

[6] C. K. Thomas and W. Saad, "Neuro-symbolic causal reasoning meets signaling game for emergent semantic communications," *arXiv preprint arXiv:2210.12040*, 2022.

[7] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2584–2597, Jul. 2022.

[8] A. Achille, G. Paolini, G. Mbeng, and S. Soatto, "The information complexity of learning tasks, their structure and their distance," *Information and Inference: A Journal of the IMA*, vol. 10, no. 1, pp. 51–72, Mar. 2021.

[9] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, USA, Jun. 2018, pp. 3733–3742.

[10] D. Huang, X. Tao, F. Gao, and J. Lu, "Deep learning-based image semantic coding for semantic communications," in *Proc. of IEEE Global Communications Conference (GLOBECOM)*. Madrid, Spain: IEEE, Dec. 2021, pp. 1–6.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, Jul. 2020, pp. 1597–1607.

[12] H. Hu, J. Cui, and L. Wang, "Region-aware contrastive learning for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Virtual, Mar. 2021, pp. 16 291–16 301.

[13] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, Apr. 2015.

[15] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," *ICT express*, vol. 6, no. 4, pp. 312–315, May 2020.