Semantic Communication Meets Edge Intelligence

Wanting Yang, Zi Qin Liew, Wei Yang Bryan Lim, Zehui Xiong, Dusit Niyato, Xuefen Chi, Xianbin Cao, and Khaled B. Letaief

Abstract

The development of emerging applications, such as autonomous transportation systems, is expected to result in an explosive growth in mobile data traffic. As the available spectrum resource becomes more and more scarce, there is a growing need for a paradigm shift from Shannon's Classical Information Theory (CIT) to semantic communication (SemCom). Specifically, the former adopts a "transmit-before-understanding" approach while the latter leverages artificial intelligence (AI) techniques to "understand-before-transmit," thereby alleviating bandwidth pressure by reducing the amount of data to be exchanged without negating the semantic effectiveness of the transmitted symbols. However, the semantic extraction (SE) procedure incurs costly computation and storage overheads. In this article, we introduce an edge-driven training, maintenance, and execution of SE. We further investigate how edge intelligence can be enhanced with SemCom through improving the generalization capabilities of intelligent agents at lower computation overheads and reducing the communication overhead of information exchange. Finally, we present a case study involving semantic-aware resource optimization for the wireless powered Internet of Things (IoT).

Introduction

With the ongoing convergence of information and communication technologies (ICTs) and artificial intelligence (AI), the "Internet of Everything" has been considered as one of the key 6G visions, wherein *semantic communication* (SemCom) and *edge intelligence* are expected to be two key enablers [1].

SemCom is widely regarded as a promising communication paradigm to breakout the "Shannon's trap." In fact, SemCom is not an entirely new concept. Just after the introduction of Shannon's theorem, Weaver and Shannon went on to identify three levels of problems within the broad subject of communication [2]:

- *Technical level:* How accurately can the symbols of communication be transmitted?
- *Semantic level:* How precisely do the transmitted symbols convey the desired meaning?
- *Effectiveness level:* How effectively does the received meaning affect conduct in the desired way?

Shannon's Classical Information Theory (CIT) [3] focuses only on the technical level and achieves success in deriving a rigorous mathematical theory of communication based on probabilistic models, wherein the information is defined as what can be used to remove uncertainty and quantified based on the probability of its occurrence from the given source.

However, the achieved transmission rates in the CIT-driven conventional communication systems are approaching the Shannon limit[1] and the available spectrum resources are becoming increasingly scarce. Moreover, the rapid development of emerging applications, for example, autonomous transportation systems, leads to a never-ending growth in mobile data traffic. In this regard, SemCom has returned to relevance. Empowered by AI technologies such as computer vision (CV) and natural language processing (NLP), end devices such as sensor nodes or smartphones may eventually be equipped with human-like reasoning capabilities. Accordingly, semantic extraction (SE) can be integrated into the communication model to achieve SemCom. SemCom, thus, allows only the information of interest to the receiver for transmission, rather than raw data. As a result, bandwidth consumption can be reduced substantially and privacy preservation can be enhanced through avoiding entire data to be exchanged. However, there are still some factors hindering the implementation of SemCom. For instance, the training process of SE models requires significant computing and storage resources, thereby impeding the scalable implementation of SemCom on resource-constrained end devices. Furthermore, in building a common knowledge base toward improving the generalization capabilities of SE models, other issues such as privacy loss may arise.

Fortunately, edge intelligence is promising to facilitate the scalable implementation of SemCom systems. The precursor to edge intelligence is edge computing, which moves part of the service-specific processing and data storage from the central cloud to the edge of the network closer to the source of data. In 5G networks, edge computing has already made significant achievements in terms of improving performance, and supporting new services and functions. Empowered by AI technol-

[1] "Shannon limit" here refers to the well-known formula $C = B\log_2(1 + \gamma)$, where B represents the bandwidth, γ represents signal-to-noise ratio, and C is the theoretical tightest upper bound on the information rate that can be communicated at an arbitrarily low error rate for Gaussian noise channels.

Digital Object Identifier: 10.1109/MWC.004.2200050

Wanting Yang is with Jilin University, China, and also with Singapore University of Technology and Design, Singapore; Zi Qin Liew, Wei Yang Bryan Lim, and Dusit Niyato are with with Nanyang Technological University, Singapore; Zehui Xiong (corresponding author) is with Singapore University of Technology and Design, Singapore; Xuefen Chi is with Jilin University, China; Xianbin Cao is with Beihang University, China; Khaled B. Letaief is with The Hong Kong University of Science and Technology (HKUST), Hong Kong.

ogies in 6G, edge intelligence aims to offer more powerful computational processing and massive data acquisition at the edge networks to achieve the dynamic and adaptive edge maintenance and management [4]. Therefore, edge intelligence can provide a good basis for offloading SE model training and knowledge storage.

On the other hand, to realize the 6G vision of ubiquitous AI, *distributed learning and inference* have become instrumental to contribute toward the intelligentization of edge networks [5]. However, the data-driven approach implies that AI-enabled agents have to incur costly communication and computation overheads, which poses challenges for the communication network, especially amid the uncertain wireless environment and limited wireless resources. In this regard, SemCom can be seen as a key enabler of edge intelligence.

Our contributions in this article are as follows:

- We introduce a general system model for SemCom involving the three communication levels foreseen by Shannon and Weaver. We then discuss typical semantic metrics and key SE techniques.
- To address the costly implementation overheads of training, maintaining, and executing SE models for SemCom, we introduce edge enabled SemCom by studying a Federated Learning (FL) enabled SE system model and edge-sharing knowledge graph for the semantic management. We also introduce the role of SemCom in training intelligent agents and empowering communication-efficient distributed machine learning (ML).
- We insightfully discuss the open issues and future research directions that are at the intersection of AI and communications toward SemCom, a key component of 6G networks.
- We provide a case study of semantic-aware resource allocation in wireless powered IoT. Different from the CIT-driven IoT, our study utilizes an AI-driven allocation mechanism to derive the resource allocation policy that maximizes semantic performance across the network.

Preliminaries

SemCom differs from traditional Shannon communication in that it incorporates human-like "understanding" and "inference" into the encoding and decoding of communication data, no longer pursuing exact data replication. In this section, we provide a brief introduction of the SemCom framework and typical semantic metrics. We then discuss the key SE techniques in the existing works.

SemCom Framework

Different from the content-blind classical communication systems, what matters in SemCom design is the accuracy of semantic content, instead of the average information associated with the possibilities of source data that can be emitted by a source [6]. As such, the main changes in the SemCom system lie in the data processing before sending and after receiving. (Fig. 1). Before encoding, the source data goes through the semantic representation module, which can be seen as the "understanding-before-transmission" process, during which the redundant information

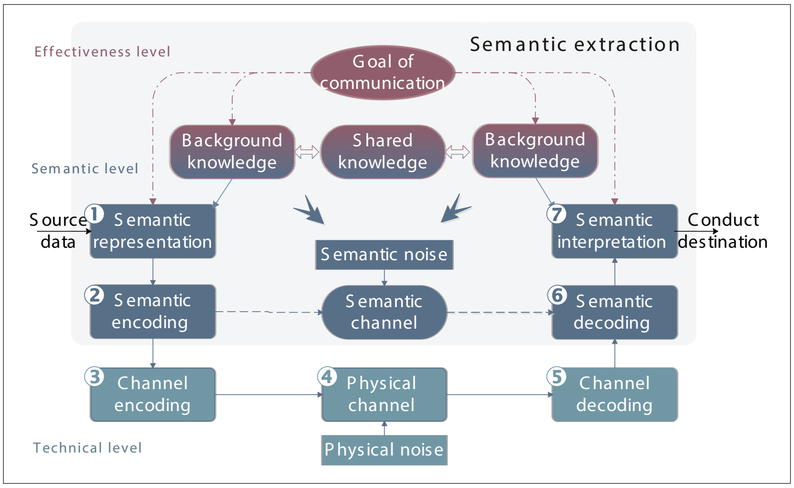

FIGURE 1. SemCom model [2, 7].

is removed. Then, the extracted relevant information goes to the semantic encoding module. In general SemCom scenarios, semantic decoding is the inverse process of encoding, which are jointly determined based on the AI technologies and their prior knowledge. For brevity, we refer to both semantic encoding and decoding as SE, and semantic encoding (decoding) in the subsequent text is considered to be integrated with the semantic representation (interpretation) module.

As with human conversation, effective conversation requires common knowledge of each other's language and communication context. In SemCom, the background knowledge (BK) of the communication parties has to be shared in real time to ensure that the processes of understanding and inference can be well-matched for all the source data. If the BK fails to match, *semantic noise* is generated, thereby resulting in performance degradation even in the absence of syntactic errors during the physical transmission.

Moreover, in some cases wherein the goal of communication may change, all possibilities for communication goals should be included in the BK and the communication goal should instruct SE to filter out irrelevant semantic information (SI) in each transmission.

Semantic Metrics

The design of network performance metrics has long been a nucleus concern in network design and optimization. As the study on SemCom is still in its early stage, most semantic metrics are derived from NLP (Table 1). Different from bit-error rate (BER) or symbol-error rate (SER) in classical communication systems, SemCom avoids treating packets equally, and measure the differences in the meaning conveyed by the recovered sentence and transmitted sentence. Besides such error-based metrics, some other metrics focus on timeliness. Age of information (AoI)-based metrics highlight the importance of data packet freshness, which allows scheduling schemes based on AoI minimization to filter out the irrelevant packets given the bandwidth constraints. By jointly considering the accuracy and timeliness of information, the authors in [9] introduce the metric of age of incorrect information (AoII) to SemCom, which measures the network performance by looking at

Semantic metrics		Advantages	Drawbacks
Bilingual evaluation understudy (BLEU)	BLEU is a method for automatic evaluation for machine translation. It is used to compare word groups with different size of the candidate with that of the reference translation and count the number of matches.	It considers the linguistic laws, such as that semantically consistent words usually come together in a given corpus.	It only calculates the differences of words between two sentences and has no insight into the meaning of the whole sentence.
Consensus based Image Description Evaluation (CIDEr)	CIDEr was proposed as an automatic consensus metric of image description quality in, which was originally used to measure the similarity of a generated sentence against a set of ground truth sentences written by humans. It can also be used to evaluate the communication performance for text transmission.	Compared to BLEU, it does not evaluate semantic similarity on the basis of a reference sentence, but a set of sentences with the same meaning.	Similar to BLUE, it is also based on the comparisons between word groups, and the semantic similarity can only be captured at the word level.
Sentence similarity	Sentence similarity is calculated as the cosine similarity of the semantic features extracted bidirectional encoder representations from transformers (BERT). BERT is a fine-tuned word representation model, which employs a huge pre-trained model including billions of parameters used for extracting the SI.	The SI in this metric is viewed from a sentence level owing to the sensitivity of BERT to polysemy, (e.g., word"mouse" in biology and machine).	The pre-trained BERT model introduces much resource consumption in the training process and makes it hard to generalize in other tasks.

TABLE 1. Some semantic metrics derived from NLP [8].

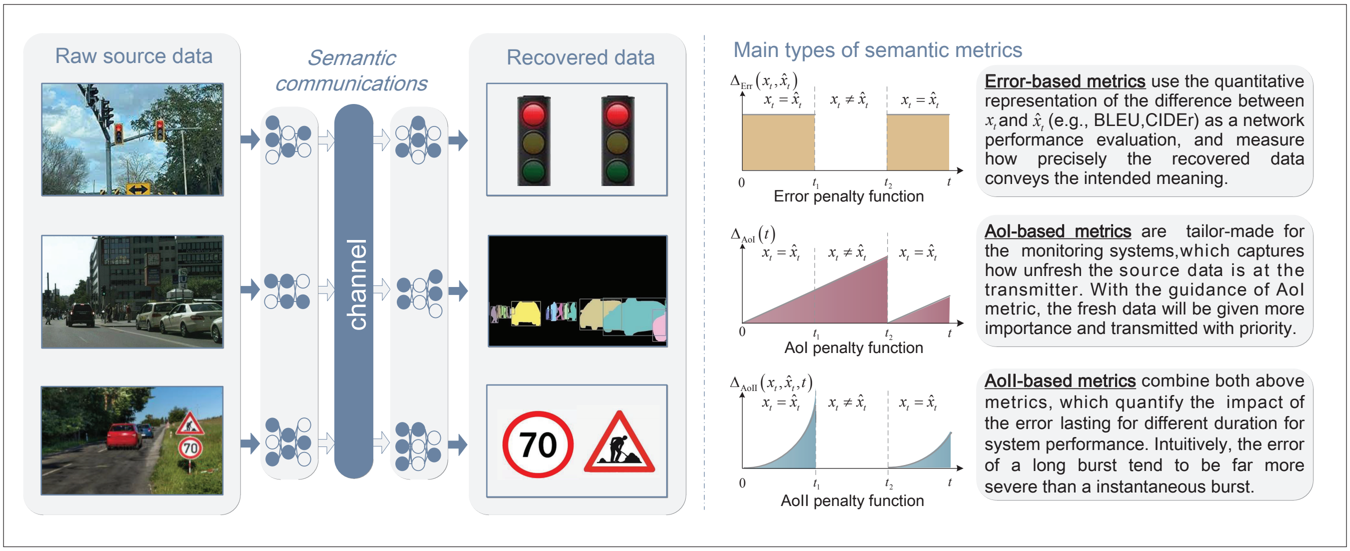

FIGURE 2. Some SemCom examples and metrics, where x_t denotes the transmitted information, and $\hat{x}_t$ denotes the estimated information inferred from the transmission [9].

a bigger picture of the packet's role in achieving the overall communication goal.

Moreover, for the cases where the benefits of some packets to be transmitted are evaluated to be important for the system objective, the value of information (VoI) is of more concern than accuracy. Hence, the VoI-related metrics can be used toward goal-oriented communications that capture the importance, relevance, and priorities of packets. Some SemCom examples and the three typical types of metrics are presented in Fig. 2.

Semantic Extraction Techniques

We now discuss some key SE techniques, the general models of which are shown in Fig. 3.

Deep Learning Based SE: With the advancement of Transformer, squeeze-and-excitation network, and deep residual network, deep learning (DL) has been widely employed in SE for text, speech, and image transmission. DL-based SE aims to enhance the system robustness at a low signal-to-noise ratio (SNR) with a shorter bit-flow. The encoder and decoder are usually modeled as two separate learnable sections, and linked through a random channel, which are trained jointly [10]. During the training process, the Generative Adversarial Networks (GANs) are commonly used to model the channel dynamics and noise. However, as the loss function is generally required to be differentiable, the common loss functions such as cross entropy are adopted during the model training process. This treats the semantic contribution of data with equal importance, which is inconsistent with human perception.

Deep Reinforcement Learning Based SE: Deep Reinforcement Learning (DRL) can integrate non-differentiable semantic metrics like BLEU into SE training. In the DRL-based SE for text transmission in [8], long short-term memory networks are employed in the encoder and decoder. The state is defined as the recurrent state of the decoder and the previously generated words. The transition between the two adjacent states is determined by the next generated word, and the action of the DRL agent is to generate a new word, with the action space defined by the dictionary dimension. As the semantic metrics can only be used as the long-term return in DRL, self-critic training is employed to address the challenging issue for

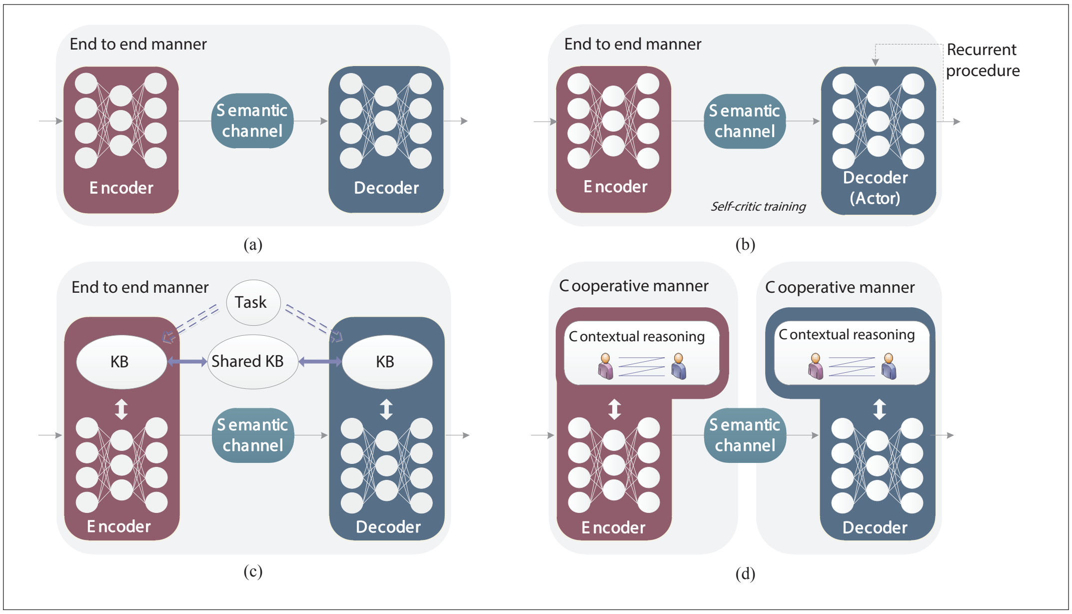

FIGURE 3. General models of main semantic extraction methods [7, 8, 10, 11]: a) System model of DL-based SE; b) System model of RL-based SE; c) System model of KB-assisted SE; d) System model of semantic-native SE.

Knowledge Base (KB) is a special database for knowledge management, which is widely used in automated AI systems to store the data with formal representation allowing for inference.

identifying the intermediate rewards, that is, the impact of each step on the semantics of the whole sentence. Moreover, for other non-sequential task, the decoding process needs to be transformed into a recurrent procedure beforehand.

Knowledge Base Assisted SE: Knowledge Base (KB) is a special database for knowledge management, which is widely used in automated AI systems to store the data with formal representation allowing for inference. The KB-assisted SE integrates the KB into the encoder and decoder, aiming to extract more SI related to the communication task for a given transmission bit limitation [7]. Specifically, the KB in Semcom is composed of source information, communication tasks, and the possible ways of reasoning that can be understood, recognized, and learned by communication participants. During the SE process, the KB is employed to quantify the level of relevance of SI for different communication tasks and instruct SE to capture the SI that is closely related to the task in each transmission. Meanwhile, as KB-assist SE is in an end-to-end manner, the KBs at both sides need to be kept in synchronization.

Semantic-Native SE: In the aforementioned methods, the SI is fixed. In [11], the authors propose a semantic-native SE, wherein the SI can be learned from iterative communications between intelligent agents, which makes it feasible in the cases where SI varies over time. Moreover, the communication parties can have the capability of contextual reasoning about the semantics in the local context of social interactions, which makes communication more efficient. Hence, it can promote intelligentization of communication systems with high degree of flexibility and efficiency.

In summary, DL-based SE is the most used. RL-based SE, while achieving better performance than DL-based ones, comes at the cost of huge computational resource consumption. Moreover, KB-based SE is only validated for image classification and semantic-native SE remains a theoretical proposition currently.

Edge-Enabled SemCom

In contrast to the classical *transmission-before-understanding* communications, the *understanding-before-transmission* paradigm of SemCom requires a shared knowledge background and computationally costly operations for SE model training and inference. This undoubtedly poses new challenges summarized as follows:

- Limited computing power and energy constraint of the end devices results in long latency in training and updating of the SE model, thereby degrading communication reliability.
- Comprehensive knowledge sharing among end devices to improve an SE model is at the cost of bandwidth and privacy. On the other hand, incomplete knowledge reduces the generalization capabilities of AI-based SE.
- Most SE methods are task-specific and trained separately, which is far from brain-like cognition and is computationally inefficient due to the duplication of work.

To address the above challenges, we propose an *edge-enabled SemCom architecture* in this section.

Federated Learning Enabled SE

We address the first two challenges by integrating edge intelligence with SemCom. Given the powerful computation and caching capabilities of the edge servers, the BK storage, and SE model training can be performed at the edge. In this way, computation and communication latency for training (mentioned in the first challenge) can be reduced [10]. Meanwhile, the edge server can serve as an authoritative intermediary for knowledge sharing, thereby eliminating the need for all communication parties to fully share each other's BK [7]. This can reduce communication parties' burden and also enhance their privacy.

We consider a common urban scenario shown in Fig. 4a. The end devices are typically clustered into different groups according to their associated access points or transmission requirements. Then,

Although the KG construction is also a computation-intensive task, the structure of KG is much more fixed than that of having to retrain separate SE models for various tasks.

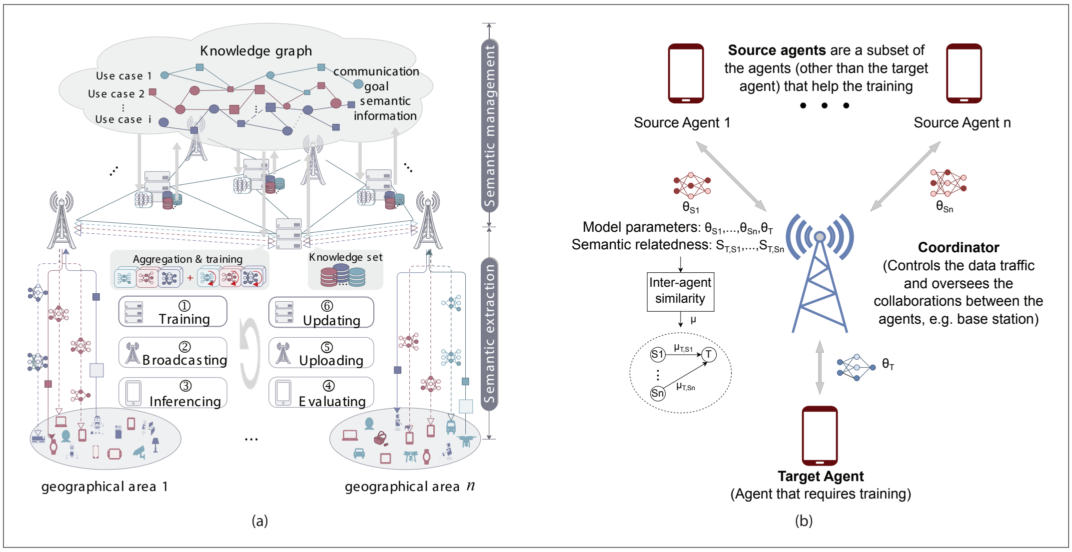

FIGURE 4. System models for the edge-enabled SemCom and Semantic-aware edge intelligence: a) architecture for edge-enabled SemCom; b) example structure of collaborative deep reinforcement learning [12].

the edge server conducts SE model training, and end devices employ well-trained models to perform SE. With FL, the trained SE model parameters in edge servers can be exchanged directly with other edge servers with the identical tasks to accelerate the training process. Therefore, the generalization performance of the model can be improved in a privacy-preserving manner. The key procedures are outlined as follows:

Step 1 Edge servers perform the pre-training or fine-tuning for specific SE tasks based on each communication group's shared BK. Model parameter exchange and federated aggregation are performed over separate communication groups with the same communication goals but not a shared knowledge background. (*Edge server*)

Step 2 The derived global models are broadcast separately to each communication group. (*Access point*)

Step 3 The source devices generate the raw data. The destination devices receive SI. Then, the SE model is utilized to encode and decode SI. (*End device*)

Step 4 The destination devices evaluate the accuracy of SI during the communications for data labeling. (*End device*)

Step 5 The newly labeled SI and/or corresponding raw data are uploaded to the edge servers. (*Access point*)

Step 6 The edge servers perform the regular updates for the knowledge sets according to uploaded information and raw data for fine-tuning of the SE model. (*Edge server*)

Efficient Semantic Extraction Based on Edge-Sharing Knowledge-Graph

In this subsection, we focus on the third challenge. Inspired by the KB-assisted SE [7], we propose to construct an edge-sharing knowledge graph (KG), which stores the underlying relations between communication goals and SI for semantic management to enable computationally-efficient SE.

In general, a sophisticated KG heavily relies on a large deep learning model and a complete knowledge set. Fortunately, this is feasible in the framework of edge intelligence, where the KG can be cached at the edge servers and the available related knowledge sets can be accessed at reduced link distances. Although the KG construction is also a computation-intensive task, the structure of KG is much more fixed than that of having to retrain separate SE models for various tasks. Once the KG is constructed, it can be cached at the edge servers to facilitate the computation-efficient SE.

As an illustration, we consider the use case of KG toward SemCom in intelligent transportation networks. Since a well-trained convolutional neural network (CNN) for multiple object identification embeds all the feature maps related to different objects, the gradients of the output of the CNN with respect to feature maps can be treated as the importance weights of the feature map to different objects [7]. Accordingly, the KG can be established by storing the importance weights of all feature maps for the tasks with different identification targets [7]. In this sense, the SE for single object identification can be executed according to the important feature map, thereby avoiding the need for specialized training and also removing redundant details from the image for more efficient transmission. Meanwhile, although autonomous vehicles and unmanned aerial vehicles (UAVs) work in distinct environments and have unique task specifications, they also share several similar characteristics and communication goals such as collision avoidance and path planning. Therefore, KG and transfer learning techniques (for the initialization of SE model parameters) can be applied to the training of SE to save much computation resources of vehicles and UAVs.

Research Directions

While the above-mentioned network architecture can facilitate the development of efficient SemCom, there are still open issues to be solved before it can be implemented in practice, some of which are highlighted below.

Interpretability and Explainability of SE: As unexpected information often appears in communications, the black-box nature of SE methods hinders its implementation. Hence, interpretability in

SE needs to be studied to associate possible causes and results and to guide improvements to the SE model. Meanwhile, explainability in SE can identify the SI hidden in deep nets, which paves the way to the KG-based efficient SE across multiple modalities and tasks described above. However, most existing SE methods are not explainable.

Semantic-Noise Based Privacy Preserving: For the communication groups with similar BK and communication goals, eavesdropping becomes easy. Considering the success of covert communication in which artificial noise is introduced for secure wireless transmissions, artificially increasing the mismatch to generate semantic noise may also serve as a potential method to enable secure SemCom.

Variable Length Semantic Encoding: Existing works merely consider the dynamic channel gains in SE without the concern of resource constraints. However, in a multi-user scenario, the fluctuation in resources, such as available spectrum and transmit power, can have a non-negligible impact on the SemCom performance. How to achieve variable-length semantic encoding to cope with dynamic network resources remains an open issue.

Semantic-Aware Edge Intelligence

As depicted earlier, more intelligent agents have to been deployed for edge orchestration in the SemCom context. Along with this, the bandwidth and energy consumption caused by the information interaction among intelligent agents becomes exacerbated. To this end, in this section, we discuss the potential of semantic awareness for the performance enhancement of edge intelligence with limited energy and bandwidth resources.

Semantic-Aware Intelligent Agent

The ever-increasing complexity of tasks in smart systems, such as autonomous driving, calls for the intelligent agents with an ability to learn adaptively based on their own experience. DRL with the capability of trial-and-error search has conceived seen as a promising method. However, due to the limited training data, which fails to represent the complex real-world environment, the monotonous experience induces overfitting issues, long convergence time, and sub-optimal performance of the DRL model. To this end, collaborative DRL (CDRL) is proposed to generalize the model experience by exchanging their model parameters or policies (Fig. 4b).

In real-world applications of CDRL, not all source agents can be selected for cooperative training due to the limited bandwidth. Since different agents have specialized environments, tasks, and action spaces, a general metric is required to evaluate the effectiveness of source agents in enhancing the learning performance of the target agent. The likelihood of a source agent being selected is only positively correlated with its structural similarity to the target agent, such as the cosine similarity between the agents' model parameters.

However, the lower structural similarity does not necessarily mean a negative collaboration [12], and agents with similar model structures may not perform a similar task. To this end, the authors in [12] further integrate semantic relatedness into the metric design for source agent selection in CDRL, where semantic relatedness is defined as the average return value received by the source agent from a target environment in limited training episodes. From the results in [12], using the same bandwidth, the average return of the DRL agents is improved when the semantic relatedness is considered, up to 83 percent higher than the baseline methods.

Semantic-Aware Distributed Deep Learning at Wireless Edge Networks

A drawback of CDRL and distributed deep learning is the high communication overheads incurred for exchanging model parameters. Therefore, finding an efficient way to compress the model parameters is essential for the implementation of edge intelligence.

Gradient sparsification and model parameter pruning are the two common methods for model compression, where a subset of the original model parameters is extracted considering the semantics or importance of the parameters for model accuracy and convergence speed. For example, in [13], gradient sparsification is adopted to compress the model at the transmitter by setting all but k elements with the highest magnitudes of entries to zero. Since only the positions of the non-zero elements are to be sent, the receiver can recover the received data in a more reliable manner with advanced noisy measurements. In [14], an adaptive model pruning method is employed for dropping a fraction of the model parameters to reduce the communication overheads in FL, where the importance of the model parameters are measured by their contribution to the future training.

Research Directions

Below, we highlight the open challenges for the SemCom enabled edge intelligence.

DL-Based SE for Task Similarity: Although the proposed semantic relatedness metric in [12] can enhance the learning performance, the extra training steps to obtain the return value greatly reduces the system efficiency. Furthermore, it remains unclear how the number of the extra training steps is determined, thereby limiting the scalability of this hand-crafted method. For future works, the semantic relatedness between the agents can be extracted based on DL. The deep learning network can take the model parameters as input and output embeddings as the semantic representations. The network can be trained by minimizing the similarity between the output embeddings of different tasks, and maximizing the similarity of the output embeddings of the similar tasks. In this way, the semantic representation can be extracted to calculate the task similarity between the agents.

Semantic Compression of Model Parameters: The existing model compression methods often require the setting of hyperparameters, for example, degree of gradients sparsity. However, identifying the optimal values for the hyperparameters could be computationally inefficient. Inspired by the semantic encoder/decoder structure in [10], which shows superiority in transmission efficiency and noise tolerance compared to conventional source coding, future works can adopt semantic-aware model parameter exchange between the end devices. Specifically, it consists of five modules similar to the structure in [10]: semantic encoder, channel encoder, channel, channel decoder, and semantic decod-

In real-world applications of CDRL, not all source agents can be selected for cooperative training due to the limited bandwidth. Since different agents have specialized environments, tasks, and action spaces, a general metric is required to evaluate the effectiveness of source agents in enhancing the learning performance of the target agent.

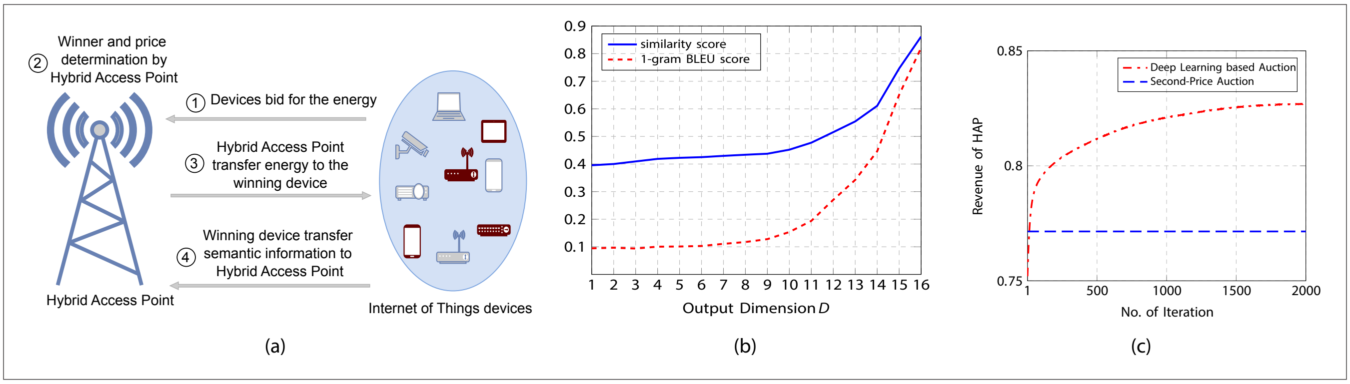

FIGURE 5. System model and experiment results of the case study: a) IoT devices bid for the energy from Hybrid Access Point; b) BLEU score and sentence similarity; c) Revenue of Hybrid Access Point.

er. Differently, the semantic encoder and decoder are for extracting the essential information at the transmitter and reconstructing the full model parameters at the receiver.

Case Study: Resource Allocation for the Convergence of SemCom and Edge Intelligence

To facilitate the convergence of SemCom and edge intelligence, it is necessary to redesign the resource allocation policies. The reason is that while classical communication systems aim to improve communication efficiency in terms of reducing the BER or SER, SemCom aims to transmit the data relevant to the transmission goal. In other words, most existing resource allocation frameworks are designed to maximize throughput without considering the semantic importance of the bit flow, especially from different users. Moreover, an edge-enabled SemCom system will have to involve entities with conflicting goals. For example, the transmitting end user aims to maximize the efficiency of the SE model while minimizing computation cost, whereas the edge server that charges the end user for its services aims to maximize revenue while reducing operational costs.

As a case study, we propose a system model in which energy-constrained IoT devices harvest the energy wirelessly for text transmission [15]. We implement all the experiments with PyTorch libraries to ensure the reproducibility on other machines. We train and test the semantic model on NVIDIA Tesla K80 Graphics Processing Unit. Different from existing studies that maximize the bit transmission rate, our proposed framework aims to maximize the *semantic performance* of the system. We consider a wireless-powered communication network where there are a hybrid access point (HAP) and multiple wireless-powered IoT devices. The IoT devices are equipped with a semantic encoder/decoder to encode/decode SI from text data. For example, SI of a sentence with 32 words is encoded as a 2-dimensional matrix with size 32 × 16, where 16 is the number of output dimensions of the semantic features. As suggested previously, this is achieved through utilizing the trained SE models cached on edge servers.

In the system (Fig. 5a), the HAP is considered to transmit energy to only one IoT device at a specific time. To decide the energy allocation, an auction mechanism is proposed where the IoT devices will bid for the energy, and the HAP will determine the winner and price. As the received energy is limited, some IoT devices need to reduce the output feature dimension to fit the energy budget. The sentence similarity and BLEU score under different output dimensions are shown in Fig. 5b. The number of bits that the devices can send upon receiving the energy is first obtained. Then, the IoT devices will adjust the output dimension to fit the data budget. As the objective of the transmission is to transfer SI, the IoT devices have more incentive to bid higher if they can achieve better semantic performance. The bids are derived from the sentence similarity and BLEU score (discussed in Table 1). In general, the higher the sentence similarity and BLEU score, the higher the bid submitted by the devices.

The winner and price are determined by a DL-based auction mechanism to maximize the revenue of the HAP. As shown in Fig. 5c, the DL-based auction mechanism achieves higher revenue as compared to the traditional Second-Price Auction in which the highest bidder wins the energy allocation and pays the price of the second-highest bidder. By maximizing the revenue of the access point, the price paid by the winning IoT device is also maximized. Hence, the energy is delivered to the device that values it the most (pay the maximized price) to ensure effective SemCom, all while fulfilling the desired properties of individual rationality and incentive compatibility for the auction. In the future, we can explore the semantic aware resource allocations for more data types, for example, image, video, and speech signal.

Conclusion

In this article, we first provided a tutorial on SemCom. We discussed the SE techniques and performance indicators that vary from the CIT. Then, we motivated the edge-driven SemCom and the SemCom-driven edge, highlighting how the component of the two technologies can play an instrumental role toward the efficient intelligentization of future networks. We also discussed open research issues, as well as provided a case study of semantic-aware resource allocation.

Acknowledgments

This research is supported in part by the National Research Foundation (NRF) and Infocomm Media Development Authority under the Future Communications Research & Development Programme (FCP), under the AI Singapore Programme (AISG) (AISG2-RP-2020-019), under Energy Research Test-Bed and Industry Partnership Funding Initiative, part of the Energy Grid (EG) 2.0 programme, under DesCartes and the Campus for Research Excellence and Technological Enterprise (CREATE) programme, Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), and in part by the supported by Wallenberg-NTU Presidential Postdoctoral Fellowship. This work is also supported by the SUTD SRG-ISTD-2021-165, the SUTD-ZJU IDEA Grant (SUTD-ZJU (VP) 202102), the SUTD-ZJU IDEA Seed Grant (SUTD-ZJU (SD) 202101), and the Ministry of Education, Singapore, under its SUTD Kickstarter Initiative (SKI 20210204). This work is also supported in part by National Natural Science Foundation of China under Grant 62271228, the Jilin Scientific and Technological Development Program under Grant 20220101103JC, Natural Science Foundation of Sichuan Province under Grant 2022NSFSC0897, and the China Scholarship Council CSC NO. 202106170088.

References

[1] K. B. Letaief *et al.*, "Edge Artificial Intelligence for 6G: Vision, Enabling Technologies, and Applications," *IEEE JSAC*, vol. 40, no. 1, 2021, pp. 5–36.

[2] W. Weaver, "Recent Contributions to the Mathematical Theory of Communication," *ETC: A Review of General Semantics*, 1953, pp. 261–81.

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

[4] X. Wang *et al.*, "Convergence of eDge Computing and Deep Learning: A Comprehensive Survey," *IEEE Commun. Surveys & Tutorials*, vol. 22, no. 2, 2020, pp. 869–904.

[5] J. Zhang and K. B. Letaief, "Mobile Edge Intelligence and Computing for the Internet of Vehicles," *Proc. IEEE*, vol. 108, no. 2, 2019, pp. 246–61.

[6] Y. Yang *et al.*, "Semantic Communications with AI tasks," arXiv preprint arXiv:2109.14170, 2021.

[7] E. C. Strinati and S. Barbarossa, "6G Networks: Beyond Shannon Towards Semantic and Goal-Oriented Communications," *Computer Networks*, vol. 190, 2021, p. 107,930.

[8] K. Lu et al., "Reinforcement Learning-Powered Semantic Communication via Semantic Similarity," arXiv preprint arXiv:2108.12121, 2021.

[9] A. Maatouk, M. Assaad, and A. Ephremides, "The Age of Incorrect Information: An Enabler of Semantics-Empowered Communication," arXiv preprint arXiv:2012.13214, 2020.

[10] H. Xie and Z. Qin, "A Lite Distributed Semantic Communication System for Internet of Things," *IEEE JSAC*, vol. 39, no. 1, 2020, pp. 142–53.

[11] H. Seo *et al.*, "Semantics-Native Communication With Contextual Reasoning," arXiv preprint arXiv:2108.05681, 2021.

[12] F. Lotfi, O. Semiari, and W. Saad, "Semantic-Aware Collaborative Deep Reinforcement Learning Over Wireless Cellular Networks," arXiv preprint arXiv:2111.12064, 2021.

[13] M. M. Amiri and D. Gündüz, "Federated Learning Over Wireless Fading Channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, 2020, pp. 3546–57.

[14] Y. Jiang *et al.*, "Model Pruning Enables Efficient Federated Learning on Edge Devices," arXiv preprint arXiv:1909.12326, 2019.

[15] Z. Q. Liew *et al.*, "Economics of Semantic Communication System in Wireless Powered Internet of Things," *Proc. ICASSP 2022-2022 IEEE Int'l. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, to be published.

Biographies

Wanting Yang received the B.S. degree in 2018 from the Department of Communications Engineering, Jilin University, Changchun, China, where she is currently working toward the Ph.D. degree with the College of Communication Engineering. Her research interests include wireless video transmission, learning, ultra-reliable, and low-latency communications.

Zi Qin Liew received the degree (Hons.) in electronic and electrical engineering from the Nanyang Technological University in 2018. He is currently an Alibaba Ph.D. Candidate with Alibaba Group and the Alibaba-NTU Joint Research Institute, Nanyang Technological University, Singapore. His research interests include wireless communications and resource allocation.

Wei Yang Bryan Lim is currently pursuing the Ph.D. degree (Alibaba Talent Programme) with the Alibaba-NTU Joint Research Institute (JRI), Nanyang Technological University, Singapore. His research interests include edge intelligence and resource allocation.

Zehui Xiong is an Assistant Professor at Singapore University of Technology and Design. Prior to that, he was a researcher with Alibaba-NTU Joint Research Institute, Singapore. He received the Ph.D. degree in Computer Science and Engineering at Nanyang Technological University, Singapore. His research interests include wireless communications, network games and economics, blockchain, and edge intelligence.

Dusit Niyato [F] is currently a Professor with the School of Computer Science and Engineering and, by courtesy, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore. He is an inventor of four U.S. and German patents. He was named the 2017–2021 Highly Cited Researcher in Computer Science. He is currently the Editor-in-Chief for IEEE Communications Surveys and Tutorials.

Xuefen Chi is currently a Professor with the Department of Communications Engineering, Jilin University, Changchun, China. She was a Visiting Scholar with the School of Electronics and Computer Science, University of Southampton, Southampton, U.K., in 2015. Her current research interests include machine type communications, indoor visible light communications, delay-QoS guarantees, and queuing theory and its applications.

Xianbin Cao received the Ph.D. degree in signal and information processing from the University of Science and Technology of China, Hefei, China, in 1996. He is the Dean and a Professor with the School of Electronic and Information Engineering, Beihang University, Beijing, China. His research interests include intelligent transportation systems, airspace transportation management, and intelligent computation

Khaled B. Letaief [F] has been with HKUST since 1993 where he was the Dean of Engineering, and is now a Chair Professor and the New Bright Professor of Engineering. From 2015 to 2018, he was with HBKU in Qatar as Provost. He is an ISI Highly Cited Researcher. He has served in many IEEE leadership positions including ComSoc President, Vice-President for Technical Activities, and Vice-President for Conferences.