A Survey on Semantic Communications System Based on Multi-modal Data

Thwe Thwe Win, Dongwook Won, Quang Tuan Do, Junsuk Oh,
Pham Thi Thu Hien, Jeongyeup Paek and Sungrae Cho
School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea
Email: {ttwin,dwwon, dqtuan, jsoh, ptthien}@uclab.re.kr, {jpaek, srcho}@cau.ac.kr

***Abstract*—The recent integration of deep learning with Semantic Communications (SC) has led to innovative advancements in the field of communication. In this survey paper, we extensively illuminate the research trends in SC systems that utilize deep learning, covering various single-modal data sources such as texts, images and videos, as well as multi-modal data sources. Through this, we provide a comprehensive understanding of how deep learning-based SC systems are rapidly advancing and how the current states and prospects of this technology are shaping up. This paper offers essential knowledge and information for researchers to grasp the technological developments and significance in this field and explore new research directions.**

***Index Terms*—Semantic communications, deep learning, multimodal data, multi-task semantic communication**

I. Introduction

In recent years, advanced technologies have been rapidly evolving to enhance the performance and reliability of communications systems. At the core of these advancements lies research into deep learning, particularly Semantic Communications (SC) systems. SC systems are centered around efficient communication based on the meaning or content of data, offering a different approach from conventional communication methods. The introduction of deep learning has made a significant contribution to improving the performance of such SC systems.

The backdrop of this research involves the interplay between deep learning and SC systems and their importance. Deep learning-based models possess the capability for complex pattern recognition and data processing, which can greatly enhance the efficiency and accuracy of SC systems. In particular, the approach using deep learning for various data types and sources is gaining prominence.

In this survey paper, we will delve deeply into the research trends of SC systems utilizing deep learning, spanning from single-modal data sources such as texts, images and videos to the processing of various data sources simultaneously in multimodal contexts. Through this exploration, a comprehensive understanding can be gained regarding the status and prospects of Semantic Communications and the contributions and possibilities presented by deep learning.

The primary goal of this paper is to assist researchers in gaining insight into the technological advancements and significance of deep learning-based SC systems, enabling them to explore new research directions and application areas.

II. Literature review

A. Text Transmission

In [1], the authors proposed DeepSC, a deep learning-based Semantic Communication (SC) system. This system consists of a transmitter with a semantic encoder for extracting semantic features from the text to be transmitted and a channel encoder for generating symbols for transmission as shown in Fig.1. The receiver comprises a channel decoder for symbol detection and a semantic decoder for text estimation. The system is trained using a loss function that considers sentence similarity and mutual information. The performance of the proposed system is evaluated using BLEU scores, which measure the similarity between the transmitted and received text. The authors specially found that under low SNR (Signal-to-Noise Ratio) conditions, the proposed system outperforms traditional methos and other deep learning-based networks in terms of BLEU scores. In [2], the research addresses various challenges associated with implementing a deep learning modal-based Semantic Communication (SC) system in IoT networks. These challenges include high latency when deploying deep learning models over limited bandwidth, the need for accurate data transmission in IoT networks, and the high latency and power consumption on IoT devices due to the excessive parameterization of deep learning models.

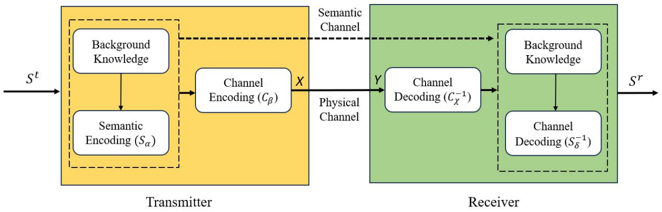
Semantic Channel
Background Knowledge
S^t
Semantic Encoding (S_α)
Channel Encoding (C_β)
X
Physical Channel
Y
Channel Decoding (C_χ^{-1})
Background Knowledge
Channel Decoding (S_δ^{-1})
S^r
Transmitter
Receiver

Fig. 1. The framework of proposed semantic communication system [1]

The authors propose L-DeepSC, a lightweight distributed SC system that jointly optimizes the receiver and feature extraction network. This system is text-based and aims to address these challenges. To mitigate the effects of fading channels, a refined minimum mean square estimator is employed. Additionally, to enable cost-effective data transmission and reception on devices with capacity constraints, finite-bit constellations are designed. To reduce the size of deep learning models, the authors propose model compression algorithms, including network sparsity and quantization.

This research seeks to make deep learning-based SC systems more practical and efficient in the context of IoT networks, by addressing these critical challenges.

B. Image Transmission

In [3], the author discusses an SC system related to image transmission, with two primary objectives: image reconstruction and task performance improvement. To evaluate these objectives, they use the PSNR (Peak Signal-to-Noise Ratio) metric for assessing image reconstruction and accuracy metrics for evaluating task performance.

A key aspect of the research in [3] is the comparison of performance based on compression ratio (CR). This is essential for understanding the impact of the compression ratio on the balance between task performance and image reconstruction, and for determining the most effective compression ratios for various tasks and datasets. However, this system faces the problem of distribution shift, where the distribution of observed data during inference differs from the empirical data used to train the system. To address this, the author Zhang propose a domain adaptation approach by adding a data adaptation network to the existing SC system. This allows the proposed system to adapt to observable datasets while maintaining high performance in data recovery and task execution aspects.

Therefore, the research in [3] contributes to the industry and academia by examining the impact of compression ratios on the balance between task performance and image reconstruction in SC systems. By comparing performance across various compression ratios, the research provides insights into how compression ratios affect this balance. Additionally, it helps determine the most effective compression ratios for various image-related tasks and datasets. This understanding is crucial for optimizing the performance of SC systems.

C. Speech Transmission

In [4], DeepSC-ST is introduced, a novel semantic communication system for speech transmission that utilizes deep learning. The system executes speech recognition and synthesis as the transmission tasks, respectively. Before recovering text, a joint semantic-channel encoder extracts speech recognition-related semantic features for transmission. Consequently, text recovery is enabled at the receiving end based on the received semantic data. This approach significantly decreases data transmission requirements without compromising performance. Within a neural network module, speech synthesis occurs at the end of the receiver by utilizing the recognized text and the speaker information. A robust model is identified to enable the DeepSC-ST in different channel conditions or environments. The simulation results of [4] show that DeepSC-ST surpasses conventional and existing deep learning-based communication systems, especially in scenarios with low signal-to-noise ratio (SNR). Finally, software, a proof-of-concept for DeepSC-ST, is developed to display its practical application.

D. Video Transmission

An innovative approach designed to revolutionize end-to-end video transmission on wireless communications channels is introduced in [5]. Nonlinear transformations and conditional strategies are merged by the proposed unique procedures to adaptively extract semantic characteristics over video frames. Moreover, the semantic features domain representations are conveyed across wireless channels by deep joint source-channel coding. The authors entitled to this comprehensive framework as deep video semantics transmission (DVST).

To leverage temporal context through the features domain is the robustness tip of this paper. The effectively progress modifying nonlinear transformation function takes advantage of this temporal context, generating with strong and exact entropy model. An inventive rate versatile transmission mechanism by guiding the transmission of each current frame. And it can tailor deep joint source-channel coding for video sources. By learning this procedure, the optimal channel within video frames can be allocated, thus the maximum result can be over from all transmission performance.

In this paper, the optimization problem is to reduce the rate-distortion performance of end-to-end transmission rate while reflecting on considering perceptual quality metrics or machine vision task performance metrics. The proposed extensive experiments proved DVST's superiority over conventional wireless video coding transmission schemes by crossing a broad spectrum of standard video source test sequences and diverse communication scenarios. The upcoming semantic communication can be reinforced by the suggested DVST framework with its video content-aware and machine vision task integration abilities. In [6], a novel approach to semantic video conferencing (SVC) is presented, leveraging key point transmission to achieve a remarkable reduction in the demand for transmission resources. The central innovation lies in the handling of transmission errors within SVC. Unlike conventional methods where errors directly impact individual pixels and degrade image quality, SVC primarily experiences changes in expressions when errors occur, preserving the overall image quality.

One notable challenge addressed in this research is the accurate detection of expression changes resulting from transmission errors. Conventional error detection methods, such as cyclic redundancy checks, often fall short in precisely reflecting the degree of expression changes withing the transmitted video data.To overcome this limitation, the authors introduce a groundbreaking solution called the "incremental redundancy hybrid automatic repeat-request framework for varying channels" (SVC-HARQ). SVC-HARQ is designed with flexibility in mind, allowing for optimized bit consumption and demonstrated strong performance in error correction and data recovery.

Furthermore, the paper introduces the concept of SVC-channel state information (CSI). This CSI feedback mechanism is carefully designed to enhance the allocation of key point transmission within the SVC system, ultimately leading to a significant improvement in the system's overall performance.

The efficiency of this wireless semantic communication system is supported by simulation results, which unequivocally demonstrate that the proposed approach significantly enhances the efficiency of video transmission, making it a promising advancement in the field of video conferencing and wireless communication.

E. Multi Modal Data Transmission

In [7], the research focuses on the efficiency of Semantic Communications (SC) for multitask scenarios. The proposed model demonstrates excellent performance in low SNR (Signal-to-Noise Ratio) conditions and exhibits significant reductions in transmission overhead and model size when compared to task-specific models designed for specific tasks. The paper explores the importance of SC for various data types, including images, text, and videos. It emphasizes the ability to identify rich features in crucial task-specific patches using a dynamic channel encoder. It was observed that patches in the central part of images are more likely to be preserved. For text, it confirmed that important words, such as emotional words, are more likely to be preserved in text classification tasks.

Various datasets, such as CIFAR-10, SST-2, VQAv2, MOSEI, were used in the research. Based on these datasets, the proposed Feature Selection Mechanism (FSM) is demonstrated to be effective compared to random selection strategies. This research provides a deep understanding of Semantic Communication and explores effective SC methods for various data types and tasks. And also, a novel solution called Unified Deep Learning-enabled Semantic Communication System (U-DeepSC) is introduced in this paper. To coherently handle multiple tasks and diverse data modalities within a single end-to-end system, this unified framework is designed. The implementation of vector-wise dynamic scheme is the main creation in U-DeepSC which allows to support the requirement of different tasks for adjusting the transmitted symbols number. Moreover, this novel can adapt to varying channel conditions, ensuring optimal transmission efficiency. To evaluate the significance of feature vectors, a lightweight feature selection model (FSM) is developed and enables the hierarchical removal of redundant vectors and significantly expediting the inference process.

This system uses a unified code-book for feature representation, serving multiple tasks efficiently for minimizing transmission overhead. For decreasing the overall data load, only task-specific features within the codebook are transmitted. While offering substantial reductions in both transmission overhead and model size, simulation results confirm that U-DeepSC achieves performance comparable to task-specific semantic communication systems tailored for specific tasks. This novel approach promising in enhancing the efficiency and flexibility of task-oriented semantic communication systems.

In [8], the novel scheme channel-level information fusion is proposed to challenge multi-user semantic communication. Especially, the unique features of signal transmission in wireless communications channels are considered by the approach of this paper. In the fusion of multi-modal data, wireless channels are regarded as mediums by leveraging the authors propose at the end of receiver in comparing with conventional signal processing techniques. To reduce the inherent randomness and variability due to wireless channels during the fusion process, semantic pre-coding is also provided in this paper. To confirm the efficacy of this fusion scheme, a comprehensive case learning is proposed that displays the practical benefits. In the field of multi-user communication, a thoughtful discussion remains and challenges. Towards more efficient and intelligent communication systems, the provided challenges represent exciting opportunities for the upcoming research and development in this domain.

III. Conclusion

The innovation in digital communication has been greatly accelerated through the fusion of Semantic Communication (SC) systems and deep learning. In this survey paper, we have meticulously illuminated the latest research trends in SC systems utilizing deep learning, ranging from single-modal data sources such as text and images to multi-modal data sources.

Through our investigation, we have confirmed that deep learning-based SC systems play a crucial role in significantly enhancing the efficiency and accuracy of communication. The flexible processing capabilities, particularly for diverse data sources, have proven to be a key advantage derived from the integration of deep learning into SC systems.

However, it is essential to emphasize that the fusion of deep learning and SC systems is still in its early stages, and ongoing research and technological developments are necessary to further improve efficiency and performance. Such research and technological advancements are expected to greatly contribute to the establishment of more advanced communication systems.

Finally, we hope that this paper serves as a valuable reference for researchers and developers in this field, allowing them to grasp the current technological trends and set directions for future research and development. The fusion of deep learning and SC systems holds the potential to significantly reshape the future of communication technology, and we eagerly anticipate the continued research and development in this exciting field.

Acknowledgment

This work was supported in part by the National Research Foundation of Korea (NRF), South Korea grant funded by the Korea government (MSIT) (No.2022R1A4A5034130), in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2022-00156353) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation), South Korea.

REFERENCES

[1] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, ”Deep learning based semantic communications: An initial investigation,” in GLOBECOM 2020-2020 IEEE Global Communications Conference, 2020, pp. 1-6.

[2] H. Xie and Z. Qin, ”A lite distributed semantic communication system for Internet of Things,” IEEE Journal on Selected Areas in Communications, vol. 39, no. 1, pp. 142-153, 2020.

[3] H. Zhang, S. Shao, M. Tao, X. Bi, and K. B. Letaief, ”Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data,” IEEE Journal on Selected Areas in Communications, vol. 41, no. 1, pp. 170-185, 2022.

[4] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu and G. Y. Li, ”Deep Learning Enabled Semantic Communications With Speech Recognition and Synthesis,” in IEEE Transactions on Wireless Communications, vol. 22, no. 9, pp. 6227-6240, Sept. 2023, doi: 10.1109/TWC.2023.3240969.

[5] S. Wang et al., ”Wireless Deep Video Semantic Transmission,” in IEEE Journal on Selected Areas in Communications, vol. 41, no. 1, pp. 214-229, Jan. 2023, doi: 10.1109/JSAC.2022.3221977.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] P. Jiang, C. -K. Wen, S. Jin and G. Y. Li, ”Wireless Semantic Communications for Video Conferencing,” in IEEE Journal on Selected Areas in Communications, vol. 41, no. 1, pp. 230-244, Jan. 2023, doi: 10.1109/JSAC.2022.3221968.

[7] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, ”A unified multi-task semantic communication system for multimodal data,” arXiv preprint arXiv:2209.07689, 2022.

[8] X. Luo, R. Gao, H. -H. Chen, S. Chen, Q. Guo and P. N. Suganthan, ”Multi-Modal and Multi-User Semantic Communications for Channel-Level Information Fusion,” in IEEE Wireless Communications, doi: 10.1109/MWC.011.2200288.