Deep learning (LLM) for Wireless Communications

Amogu J. Uduka
Electrical Engineering
George Mason University
Virginia, USA
auduka@gmu.edu

***Abstract*—This literature review examines the growing impact of deep learning, with a focus on large language models (LLMs) such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT), in advancing semantic communication systems and driving other technological innovations in wireless communications. The review consolidates insights from various studies that leverage these models to optimize data transmission across different modalities—text, audio, sensor data, and images—within next-generation wireless networks. It explores the potential of Large Language Models (LLM) to enhance bandwidth efficiency, reduce transmission overhead, and improve signal processing through shared knowledge bases and pre-trained model integration.**

***Index Terms*—6G, Deep Learning, Semantic Communication, Additive White Gaussian Noise (AWGN)**

I. Introduction

The emergence of 6G technology is set to transform wireless communication by incorporating artificial intelligence (AI) to develop intelligent, self-optimizing networks (SON). As global connectivity demands grow for faster, more reliable, and energy-efficient solutions, 6G networks powered by AI have the potential to meet these challenges and usher in a new era of seamless communication. Central to 6G's innovation is the integration of AI models within the network's core architecture. These intelligent systems will process vast data streams, detect patterns, and make data-driven decisions to optimize network performance. With machine learning (ML), 6G networks can dynamically manage resources, reduce interference, and strengthen security protocols, enabling more efficient and secure communications [1].

Deep learning is a branch of AI that enables computers to process and interpret data by mimicking the way the human brain operates. These models are capable of identifying intricate patterns in images, text, audio, and other types of data, providing highly accurate insights and predictions. Deep neural networks (DNNs) have driven significant advancements across a wide range of applications, including speech recognition, natural language processing, image classification, data analytics, and autonomous vehicles, among many others. These breakthroughs have transformed industries by enhancing the ability of AI systems to understand, process, and act on complex data.

In complex scenarios where accurately modeling channels is difficult, these approximations may fail to capture the full intricacies of the environment. A more effective alternative is a data-driven approach that utilizes various deep neural network (DNN) architectures, each suited to specific tasks. For instance, convolutional neural networks (CNNs) are adept at extracting spatial features from wireless signals, making them ideal for tasks like channel estimation and signal classification. Recurrent neural networks (RNNs) and long short-term memory (LSTM) networks are well-suited for handling sequential data, enabling them to excel in tasks such as predicting time-varying channels. Transformer-based models, known for their efficiency in capturing long-range dependencies, can be used in scenarios that require understanding complex relationships in signal transmission. By bypassing traditional channel estimation and instead adapting to real-time data using these specialized DNN architectures, wireless systems can achieve greater accuracy and efficiency in environments where traditional models fall short. LLMs are DNN architectures designed to process and generate human-like text by analyzing vast amounts of data and identifying complex patterns. By leveraging their ability to model intricate relationships and capture long-range dependencies, LLMs can transform various aspects of wireless networks such as encoding and decoding transmitted data based on its semantic meaning, rather than relying solely on the traditional bit-level transmission. This enables more efficient data compression and improved communication accuracy, especially in bandwidth-constrained environments. Moreover, LLMs can enhance predictive capabilities in wireless systems, such as optimizing channel prediction by learning from historical data to anticipate future conditions. The rest of this review is organized as follows. Section II presents a detailed description of the context, the problems addressed, and the main results from the reviewed papers. Section III focuses on the novelty of these papers, exploring their unique contributions and how they relate to one another. Finally, Section IV discusses future directions and potential extensions for the integration of AI and deep learning techniques into wireless communications.

II. Part I: Description of Context

A. LLM4CP: Adapting Large Language Models for Channel Prediction

1) Description of Context: [2] proposes the adaptation of LLMs for multi-input single-output (MISO) orthogonal frequency-division multiplexing (OFDM) channel prediction, aiming to enhance predictive capability and generalization. The approach utilizes a channel prediction neural network based on a pre-trained GPT-2 model, which is fine-tuned to predict future downlink channel state information (CSI) sequences based on historical uplink CSI data. The primary focus of this work is to develop an LLM-based technique that reduces the information overhead associated with CSI estimation, thereby improving spectrum efficiency, which is often negatively impacted by excessive overhead especially in Frequency Division Duplexing (FDD) systems where there is no reciprocity.

2) Approach Taken and Problem solved:

1) **Pre-processing Module**:
   This module's primary goal was to accurately predict the future downlink CSI of a channel given a specific dimensional resource block. The pre-processing was parallelized for each antenna (transmitter and receiver) to reduce the complexity and training time of the DNN. Input data was normalized to facilitate network training and convergence. To handle local temporal features and reduce computational complexity, a patching operation along the temporal dimension was adopted.
2) **Embedding Module**:
   This module was employed for preliminary feature extraction before passing the data into the LLM. It included the collection of CSI attention and position embedding. CNN layers were utilized to extract temporal and frequency features within each patch and integrate features across patches. This block helped extract different weights for each patch. Global average pooling was applied to generate channel-wise statistics. Fully connected (FC) layers modeled the correlation between different patches.
3) **Output Module**:
   The output module converted the output features of the LLM into the final prediction results by using the first two fully connected layers to transform the dimensions of the LLM output. Finally, the output was de-normalized to generate the final network output. After which, its output is fed into the LLM.

Key Components of the Experiment:

1) **Dataset**: QuadRIGA
2) **Antenna Spacing**: Half wavelength at the carrier frequency
3) **Channel Model**: 3GPP Urban Macro (Uma) and No Line of Sight (NLOS) scenarios
4) **Neural Networks and Models**:
   - **PAD**: Designed to overcome mobility issues in TDD.
   - **RNN)**: For channel prediction tasks.
   - **LSTM**: Designed with memory cells and multiplicative gates to manage long-term dependencies (solves vanishing gradient problem).
   - **Gated Recurrent Units (GRU)**: Tackles vanishing gradient issues.
   - **CNN**: Treats CSI as a 2D image processing task.
   - **Transformer**: Addresses error propagation and serves as a basis for comparison.
5) **Performance Metrics**: Normalized Mean Square Error (NMSE), Spectral Efficiency, Bit Error Rate (BER).
6) **Hyper-parameters for Network Training**: Batch size, Epochs, Optimizer, Starting learning rate, Learning rate decay rate.

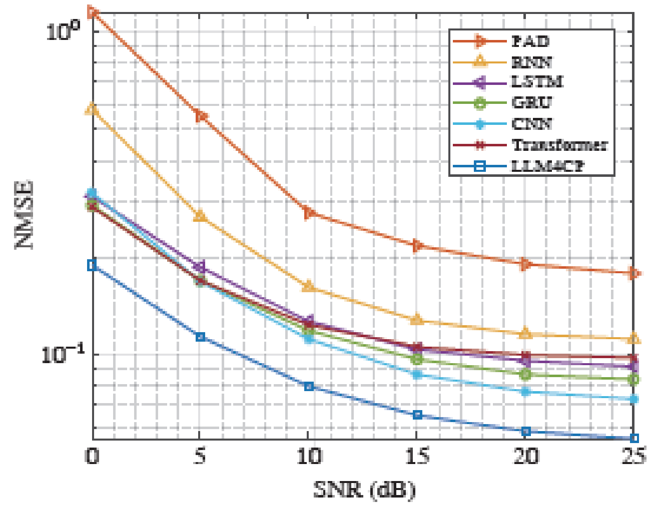


Fig. 1. NMSE performance of LLM4CP compared to baseline methods across various user velocities.

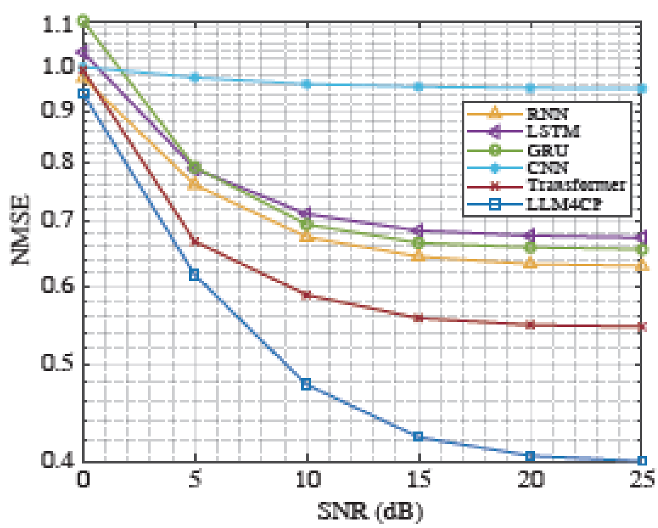


Fig. 2. LLM4CP demonstrates distinct advantages in FDD systems over various user velocities.

3) Main Results: In TDD systems, while LLM4CP performs well, its advantages are more evident in FDD systems under different user velocities. Figure 1 illustrates that the NMSE performance of LLM4CP outperforms baseline methods across different user velocities, showing higher prediction accuracy. Moreover, Figure 2 highlights the clear superiority of LLM4CP in FDD systems, attributed to its enhanced capability to model complex time-frequency relationships.

B. Deep Learning-Enabled Semantic Communication Systems

1) Context Overview: Building on recent advancements in Deep Learning for NLP, [3] tackles the issue of spectrum resource allocation through a semantic-based communication system. By interpreting the semantic meaning of digital bits, the system improves communication accuracy and efficiency. It operates within the semantic domain, extracting valuable information while discarding irrelevant data, leading to compressed communication that retains essential meaning. The paper seeks to explore the following questions:

- How can the meaning behind digital bits be defined? (Self-Attention Mechanism)
- How can the semantic error in sentences be measured? (Sentence Similarity)
- How can semantic and channel coding be designed jointly? (Bi-Directional LSTM)

2) Approach and Problem Solved: The system model consists of two layers:

- Semantic level: Handles the processing of semantic information for encoding and decoding to extract meaningful content.
- Transmission level: Ensures the accurate exchange of semantic information over the communication medium.

The model initializes parameters like weights and bias, using an embedding vector to represent the input words. The training process for the Deep Semantic Communication (Deep-SC) model involves two phases, driven by distinct loss functions. The Deep-SC model minimizes text estimation error using two loss functions: Cross-Entropy and Mutual Information. The first term minimizes the semantic difference between the original and decoded sentences (Cross-Entropy), while the second term optimizes the data rate during transmitter training (Mutual Information).

1) Phase 1 (Training of the Mutual Information Estimation Model): It assumes that both the transmitter and receiver possess background knowledge, referring to pre-existing information that supports efficient communication. This knowledge can be sourced from previous training data or experiences relevant to the communication scenario. Utilizing this background knowledge helps the system encode and decode semantic information effectively, maintaining meaning despite the stochastic nature of the physical channel. The knowledge set (K) generates a mini-batch of sentences, which are transformed into dense word vectors via the embedding layer. The semantic encoder layer extracts the semantic information (M) from the sentence, which is encoded into symbols (X) to manage physical channel effects. After passing through the channel, the receiver receives distorted signals (Y) influenced by noise, and the Mutual Information loss is calculated, optimizing weights and bias using stochastic gradient descent (SGD).
2) Phase 2 (Whole Network Training) : The mini-batch (S) from the knowledge set (K) is first encoded into semantic information (M) at the semantic level. Next, (M) is further encoded into symbols (X) for transmission across the physical channels. At the receiver's end, the distorted symbols (Y) are captured and decoded by the channel decoder layer. The semantic decoder layer then estimates the transmitted sentences. Finally, the entire network is optimized using stochastic gradient descent (SGD), with the loss being computed and minimized accordingly.

Key Components of the Experiment:

TABLE I
THE SETTING OF THE DEVELOPED SEMANTIC NETWORK

	Layer Name	Units	Activation
Transmitter (Encoder)	3×Transformer Encoder	128 (8 heads)	Linear
	Dense	256	ReLU
	Dense	16	ReLU
Channel	AWGN	None	None
Receiver (Decoder)	Dense	256	ReLU
	Dense	128	ReLU
	3×Transformer Decoder	128 (8 heads)	Linear
	Prediction Layer	Dictionary Size	Softmax
MI Model	Dense	256	ReLU
	Dense	256	ReLU
	Dense	1	ReLU

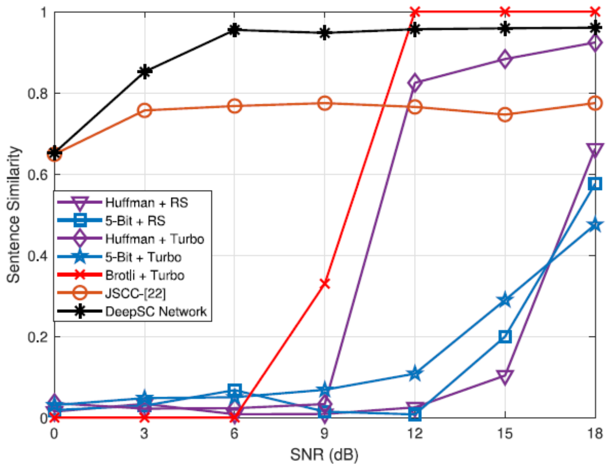
1) **Dataset**: European Parliament proceedings
2) **Channel Model**: AWGN
3) **Neural Networks and Models**: As shown in Table I, the semantic network is structured into several layers.
4) **Source-channel coding**: The network consists of Bi-Directional LSTM layers.
5) **Requirements**: Simulation is performed with a computer with Intel Core i7-9700 CPU @3.00 GHz and NVIDIA GeForce GTX 2060.
6) **Performance Metrics**: BLEU and Sentence Similarity

3) Main Results: : For purpose of the simulation, a perfect CSI was assumed for all schemes. Due to the advantage of Sentence Similarity over the BLEU score for this specific illustration, its relationship with the SNR under the same number of transmitted symbols over AWGN and Rayleigh fading channels is showed in Figure 1 and Figure 2 respectively.

Also, the performance of transfer learning aided Deep-SC for two tasks was investigated. The result is shown in Figure 5. Transfer Learning leads to faster training (quicker convergence) and better performance (higher BLEU score) compared to training from scratch, particularly in low signal-to-noise environments. This suggests that leveraging prior knowledge can significantly improve the efficiency and effectiveness of Deep-SC in various communication conditions.

C. Large Generative AI Models for Telecoms: The Next Big Thing?

1) Context Overview: The primary objective is to create an AI-native network, overcoming the limitations of Self Organizing Networks (SON). Although SON can reduce manual intervention by performing well under predefined network conditions, its efficiency diminishes when faced with unexpected changes or conditions outside of its programmed scope. Essentially, while SON is effective within its preset boundaries, it



(a) AWGN

Fig. 3. AWGN Channel: Sentence Similarity vs. SNR

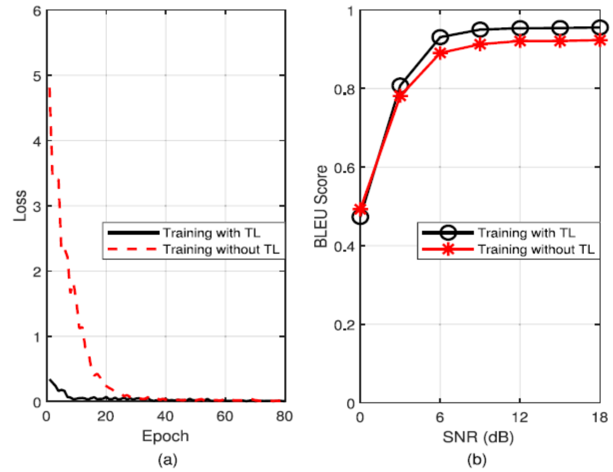


Fig. 5. Transfer learning (TL) aided Deep-SC with different background knowledge: (a) loss values versus the number of training epochs, (b) BLEU score (1-gram) versus the SNR.

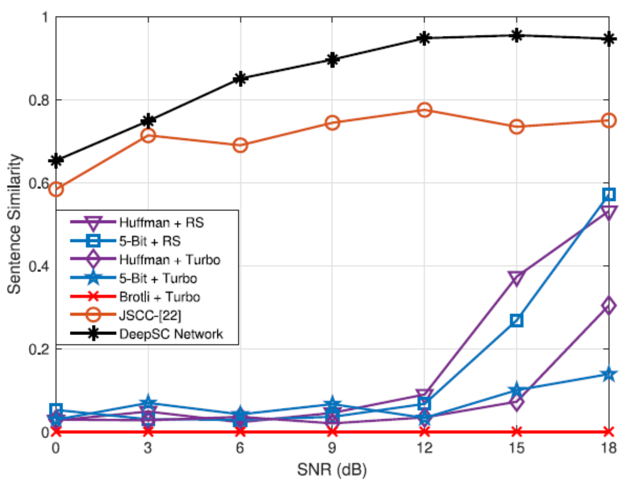


Fig. 4. Rayleigh Fading Channel: Sentence Similarity vs. SNR

lacks adaptability in more dynamic and unpredictable network environments.

[4] highlights the need for Generative AI (Gen-AI), a branch of AI that can produce new content based on patterns learned from existing data. Using transformers' self-attention mechanism and extensive training data, large models can capture intricate statistical patterns and relationships in the data, enabling them to predict and generate the necessary outcomes. The approach is structured across two key paradigms:

- Large Gen-AI Model for Wireless Applications
  - Large Language Models for Sensing: Key applications include:
    - 3D Wireless Imaging Architecture
    - Super-Resolution Localization
  - Large Language Models for Transmissions: Key applications include:
    - Multi-Modal Beamforming
    - Frequency Division Duplexing Transmission
    - Joint Source-Channel Coding (JSSC)
- Wireless Technologies for Large Gen-AI Models
  - 6G and Collective Intelligence: Key use cases include:
    - Semantic Communication: Explored in [3] and [5], this concept focuses on the extraction of meaning from transmitted data to optimize communication.
    - Emergent Protocol Learning
    - Distributed Large Gen-AI Model Powered AI Agents
  - Use Cases of Collective Intelligence: These include intent-driven network automation, efficient multi-agent communication, autonomous vehicle management, improvements in traffic flow and safety, and distributed planning using large Gen-AI models.

2) Approach and Main Results: [5] examined how large generative AI models could fundamentally transform the design, configuration, and operation of future wireless networks. It highlighted key opportunities in sensing and communication that become possible through integrating large generative AI models into wireless systems. Additionally, it analyzed how wireless networks can facilitate machine-to-machine communication using these AI models, potentially leading to networks that can evolve autonomously.

III. Part II: Critique of Paper

A. LLM4CP: Adapting Large Language Models for Channel Prediction

[2] introduced an innovative LLM-based channel prediction method designed for MISO-OFDM systems to minimize CSI acquisition overhead. Initial results affirm that the method achieves state-of-the-art performance in TDD/FDD channel

prediction, showcasing exceptional few-shot learning and generalization abilities. Also, in comparison with other selected papers for this review, it is the only paper concerned about channel estimation and prediction. However, some key concerns remain:

- The method focuses solely on moving users or objects. Can it also be effective in both indoor and outdoor environments?
- There is no clear method to guarantee that the predicted channel response will be sufficiently accurate, which could impact the effectiveness of matched-filter-based precoding.
- The approach has only been tested on MISO-OFDM channels. Will similar results be observed in other types of channels?

B. Deep Learning-Enabled Semantic Communication Systems

[3] showed that the Deep-SC system achieved best results for text transmission mostly in the low Signal-to-Noise Ratio (SNR) regime. However In its preliminary state, it is key to use this model jointly with other existing channel encoding model for various SNR regimes, especially in the high SNR regime. Its main critique are as follows;

- There is an absence of a mathematical model for semantic information theory in engineering.
- High Dimensional spaces can be difficult to work with. searching through this space can be computationally intensive, and finding the optimal representation or result can be challenging (This was addressed in [5]).

It is worthy to note that [5] differs from [3] by using a shared knowledge base as shown in Figure 6 and Figure 7. A shared knowledge base refers to a repository of common or predefined information that both the transmitter and receiver have access to and rely on during communication. This shared knowledge enables both sides to interpret and reconstruct transmitted messages more efficiently by focusing only on the residual information—the part of the message that is not already present in the shared knowledge base.

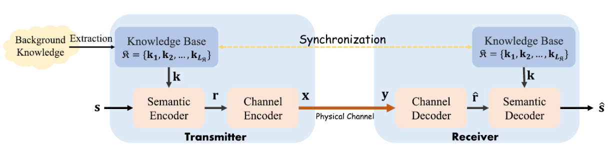


Fig. 6. Deep Learning Enabled Semantic Communication Systems (System Model)

C. Large Generative AI Models for Telecoms: The Next Big Thing?

[4] primarily focuses on the opportunities presented by large Gen-AI models in the context of self-optimizing networks. However, while it is nearly impossible to comprehensively identify all potential applications of large Gen-AI models in wireless communications, it remains necessary to explore and outline additional perspectives for their use in this field.

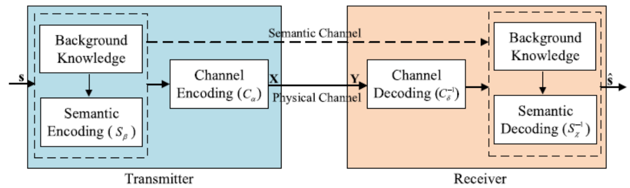


Fig. 7. Deep Learning-Empowered Semantic Communication Systems With a Shared Knowledge Base (System Model)

IV. Part III: Potential Improvements and Conclusions

LLMs have demonstrated significant potential with their wide range of applications across various aspects of wireless communications. However, a major challenge lies in the substantial computational resources and lengthy training times required to develop reliable models for routine use in industry, academia, and governmental agencies. This is likely to drive companies like NVIDIA to further innovate and lead the development of highly specialized hardware designed to meet the unique demands of this field.

References

[1] W. Yu, F. Sohrabi, and T. Jiang, "Role of deep learning in wireless communications," *IEEE BITS the Information Theory Magazine*, vol. 2, no. 2, pp. 56–72, 2022.

[2] B. Liu, X. Liu, S. Gao, X. Cheng, and L. Yang, "Llm4cp: Adapting large language models for channel prediction," *Journal of Communications and Information Networks*, vol. 9, no. 2, pp. 113–125, 2024.

[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, 2021.

[4] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, "Large generative ai models for telecom: The next big thing?" *IEEE Communications Magazine*, pp. 1–7, 2024.

[5] P. Yi, Y. Cao, X. Kang, and Y.-C. Liang, "Deep learning-empowered semantic communication systems with a shared knowledge base," *IEEE Transactions on Wireless Communications*, vol. 23, no. 6, pp. 6174–6187, 2024.