Transformer-Empowered 6G Intelligent Networks: From Massive MIMO Processing to Semantic Communication

Yang Wang[1], Zhen Gao[1], Dezhi Zheng[1], Sheng Chen, Deniz Gündüz, and H. Vincent Poor

[1]Beijing Institute of Technology

January 8, 2023

Outline

Outline

6G Requirements

- **Global Coverage**
  Provide seamless wireless coverage by multiple heterogeneous nodes, such as satellite, UAV, terrestrial, and maritime.

- **All Spectra**
  Provide much more wider band, such as mmWave and THz.

- **Intelligence**
  Provide intelligent services by big data and AI.

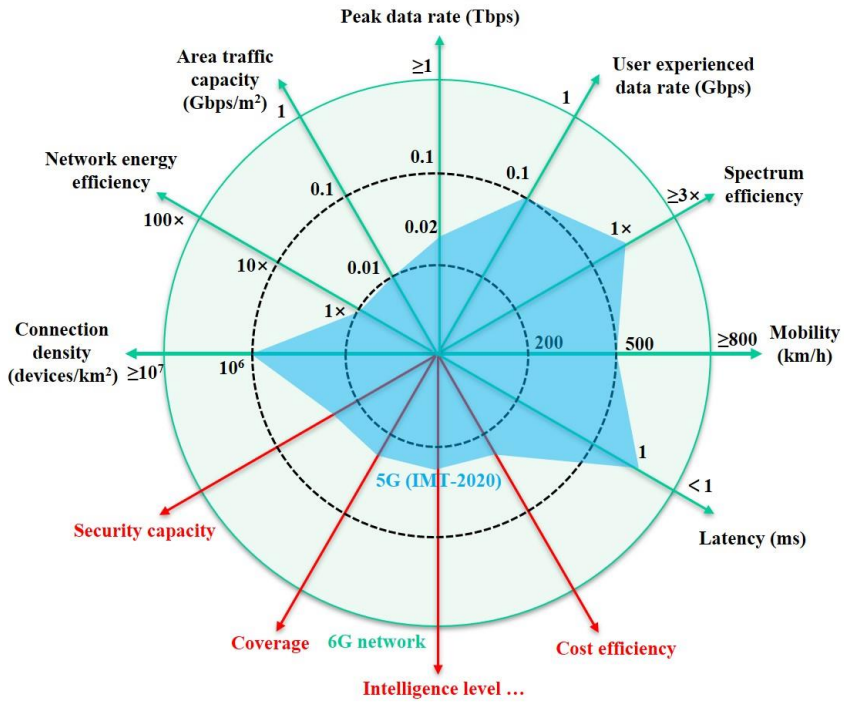


Fig. 1. The performance metrics of 6G wireless communication systems.

Challenges

Several challenges:

- **Channel Measurements and Modelling.**
  High frequency channel environment is more complex and hard to modelling.

- **Communication Resource Overhead.**
  The estimation and feedback overhead of high-dimensional channel parameters is unaffordable.

- **Algorithm Efficiency**
  Simultaneously satisfying high performance, low-complexity, and short running time are difficult for conventional mathematical optimization solution.

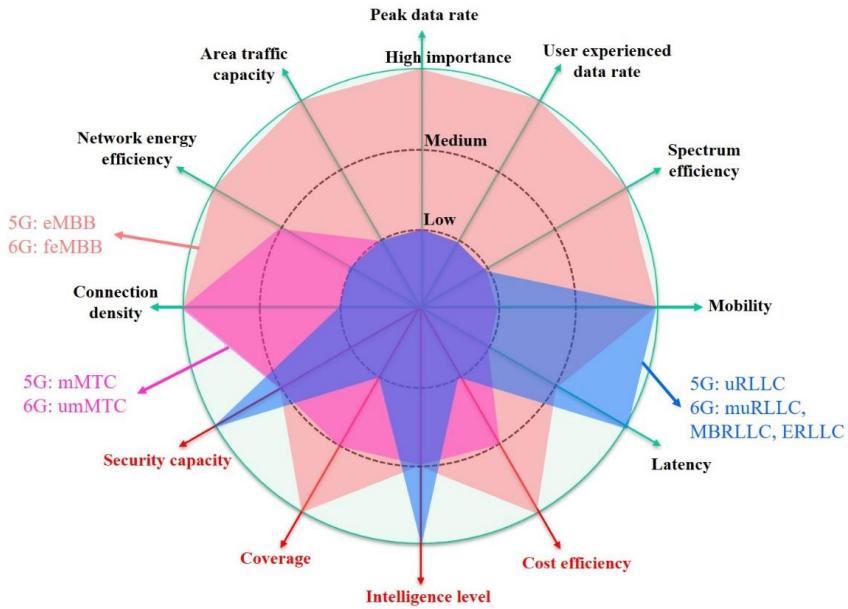


Fig. 2. Comparison of 5G and 6G key performance metrics and application scenario requirements.

In this paper, we introduce the transformer architecture and explore its application in 6G intelligent network design. We present a transformer-based architecture for 6G intelligent processing, and study its performance in various wireless communication problems.

Outline

Outline

Common DNN Architectures

Recently, the common DNN architectures have MLP, CNN, RNN, and Auto-encoder.

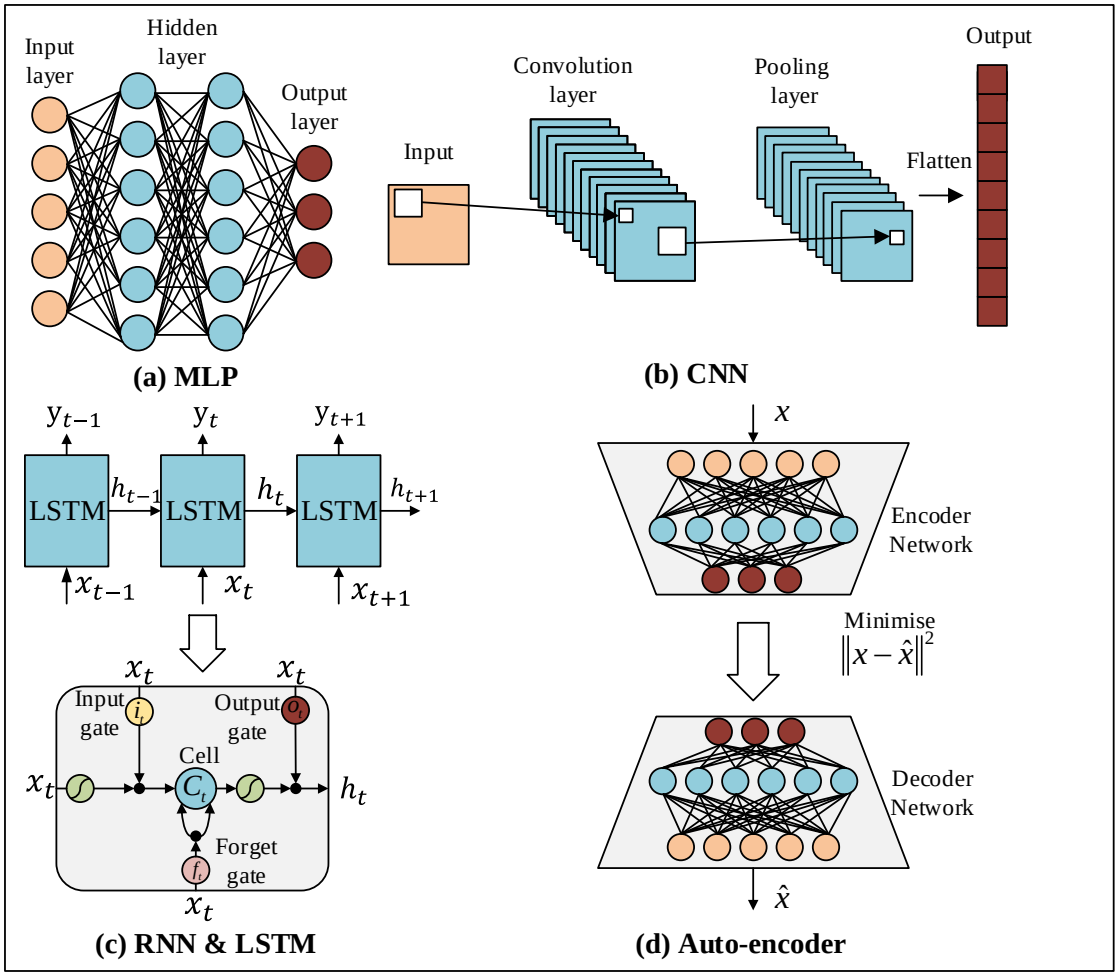


Fig. 3. Neural network structure of MLP, CNN, RNN & LSTM, and auto-encoder.

Outline

Transformer is a novel network structure, which has obtained the significant success on NLP and CV fields.

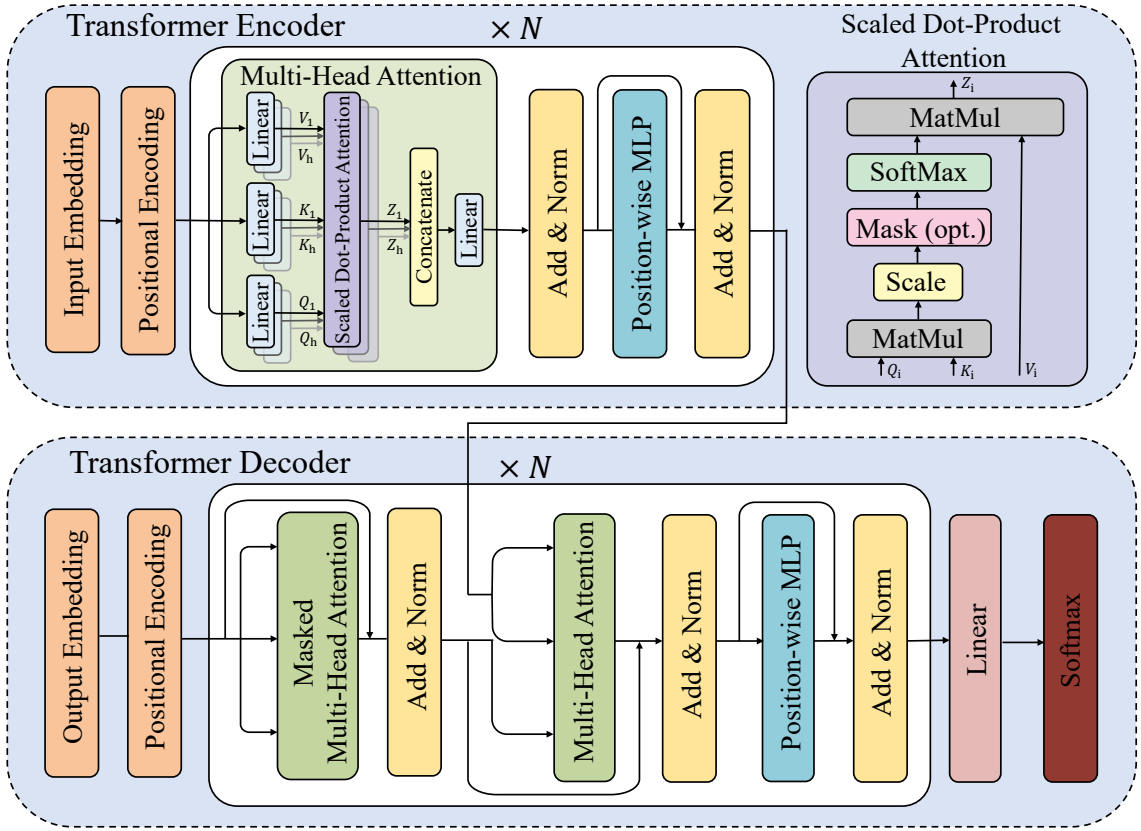


Fig. 4. Structure of the transformer network.

- **Transformer Encoder:** mainly consist of the input embedding layer, the positional embedding layer, and multiple encoder layers.
  - Multi-head self-attention sub-layer
  - Position-wise MLP sub-layer

  As shown in Fig. 4, the input sequence X_s is first transformed into three different sequential vectors: the queries, keys, and values with different learned linear projections, respectively (i.e., $\{\mathrm{Q}_i, \mathrm{K}_i, \mathrm{V}_i\} \in \mathbb{R}^{K_s \times d_m}, 1 \leq i \leq h$, where h is the number of heads and $d_m = d_{\mathrm{T}}/h$). Then, the multi-head attention operation can be expressed as

$$\begin{aligned} \mathrm{MultiHead}(\mathrm{X_s}) &= \mathrm{Concat}(\mathrm{head_i}, \cdots, \mathrm{head_h})\mathrm{W}^{\mathrm{O}}, \\ \text{where } \mathrm{head}_i &= \mathrm{softmax}\left(\frac{\mathrm{Q}_i \mathrm{K}_i^T}{\sqrt{d_m}}\right)\mathrm{V}_i, \end{aligned} \tag{1}$$

  where $\mathrm{W}^O \in \mathbb{R}^{K_s \times d_{\mathrm{T}}}$ is the linear projection matrix.

- **Transformer Decoder:** is similar to the encoder. Then, the decoder inserts a third sub-layer, which performs the masked multi-head attention over the output of the transformer.

Outline

Signal Processing Architecture

- We propose a novel 6G intelligent processing architecture employing transformer for both the massive MIMO intelligent processing blocks and the newly emerging semantic communication blocks.

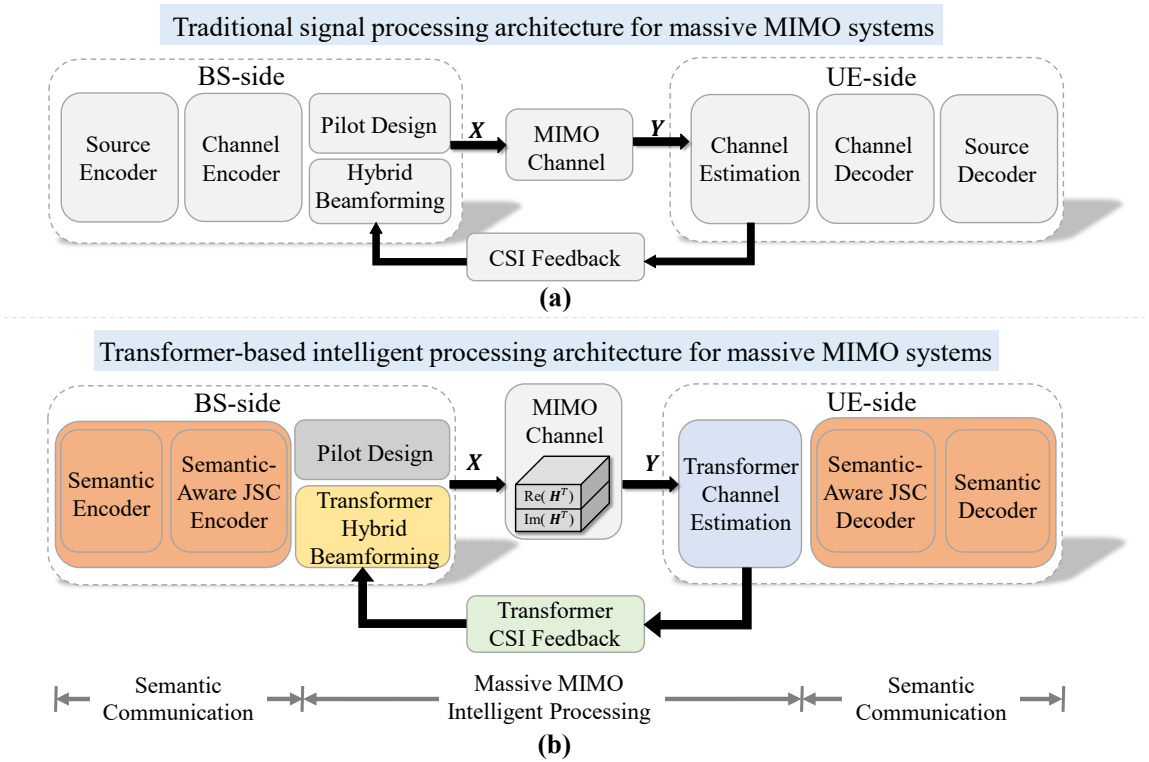


Fig. 5. Traditional and proposed transformer-based signal processing architecture for massive MIMO systems.

Outline

Channel Estimation

- Accurate CSI at the base station (BS) is critical for beamforming and signal detection in massive MIMO systems. However, it is challenging to accurately estimate high-dimensional channels with few pilots.
- Compressive sensing (CS)-based solutions: the involved matrix inversion operations and the iterative nature of CS-based techniques result in prohibitively high computational complexity and storage requirements.
- Recently, researchers have resorted to DL techniques to overcome the aforementioned challenges, such as LAMP, MMV-LAMP, DNN-based solution, and CNN-based solution.
- Here, we propose a novel channel estimator that utilizes the universal and flexible transformer architecture.

- First, we consider the downlink channel estimation problem in M successive time slots, where the BS is equipped with a UPA with N_t antennas, the UE has single-antenna, the number of OFDM sub-carriers is K. The received signals in M successive time slots can be expressed as

$$\boldsymbol{y}_k^T = \boldsymbol{h}_k^T \boldsymbol{X} + \boldsymbol{n}_k^T, \tag{2}$$

where $\boldsymbol{y}_k^T \in \mathbb{C}^{1\times M}$ is the received signal vectors on the k-th subcarrier, $\boldsymbol{h}_k^T \in \mathbb{C}^{1\times N_t}$ is the cluster-sparse channel on the k-th subcarrier, $\boldsymbol{X} \in \mathbb{C}^{N_t\times M}$ denotes the transmitted equivalent pilot signal after precoding, and $\boldsymbol{n}_k^T \in \mathbb{C}^{1\times M}$ is the complex additive white Gaussian noise (AWGN) vector.

- By stacking all the subcarriers, the received signal $\boldsymbol{Y} \in \mathbb{C}^{K\times M}$ can be written as

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{N}, \tag{3}$$

where $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_K]^T$, $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_K]^T \in \mathbb{C}^{K\times N_t}$ is frequency-spatial domain channel, and $\boldsymbol{N} = [\boldsymbol{n}_1, \boldsymbol{n}_2, \cdots, \boldsymbol{n}_K]^T$.

Transformer-Based Channel Estimation

- We propose a transformer-based channel estimation scheme.

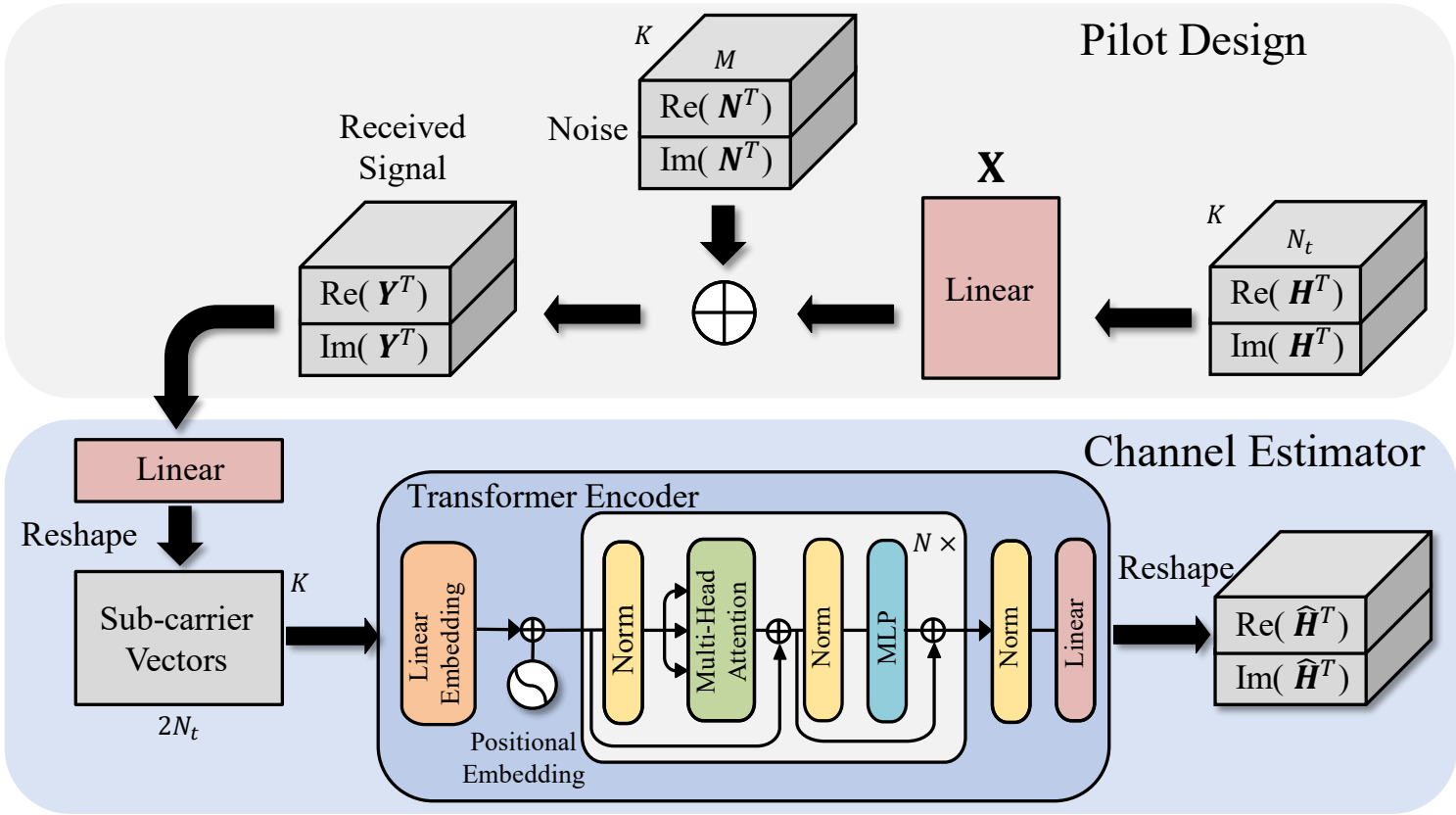
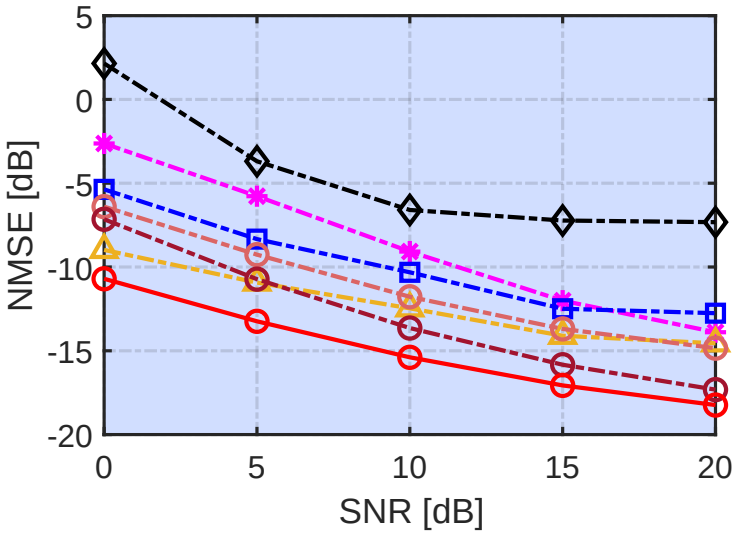


Fig. 6. The transformer-based end-to-end architecture for jointly designing the pilot signals and channel estimator.

- **Simulation Setting**
  The pilot training stage has $M = 24$ successive time slots. The BS is equipped with a uniform planar array (UPA) with $N_t = 8 \times 8 = 64$ antennas, the user equipment (UE) has single-antenna, the number of orthogonal frequency division multiplexing (OFDM) sub-carriers is $K = 32$. We consider a sparse channel scenario with $N_c = 6$ clusters, $N_p = 10$ paths per cluster, and an angle spread of $\Delta\theta = \pm 3.75^\circ$.
- **Training Setting**
  We generate training, validation, and test datasets of 100,000, 10,000, 5,000 samples, respectively. We consider the NMSE as the performance metric.



Model Name	Size	FLOPs	Runtime
SOMP	-	195 M	3.4ms
MMV-LAMP	-	132 M	12.9ms
DNN	9.26 MB	9.800 G	60.3ms
Attention-CNN	22.6 MB	24.00 G	66.4ms
Transformer-S	11.1 MB	105.4 M	3.0ms
Transformer-M	28.0 MB	346.7 M	4.9ms
Transformer-L	48.3 MB	686.8 M	9.0ms

Fig. 7. NMSE performance comparison of different channel estimation schemes vs. signal-to-noise ratio (SNR).

Outline

CSI Feedback

- The large number of antennas result in excessive feedback overhead. Similarly to channel estimation, CS-based techniques can be used to reduce the CSI feedback overhead. However, these techniques cannot fully exploit the channel structure since the channels in real systems are not exactly sparse.
- Recently, DL-based solutions have achieved impressive results for CSI feedback, such as CsiNet [10], bit-level CsiNet [11], and LSTM-based CsiNet [12].
- Herein, we present a transformer-based CSI feedback scheme to obtain more efficient quantization and compression performance compared with the prior work.
- Assuming that $\boldsymbol{H}$ is known at the UE, a general representation of the CSI quantization and feedback process can be written as

$$\hat{\boldsymbol{H}} = f_d(\mathcal{D}(\mathcal{Q}(f_e(\boldsymbol{H})))), \tag{4}$$

where $\hat{\boldsymbol{H}}$ denotes the recovered CSI at the BS, $f_e(\cdot)$ represents the preprocessing and encoding before quantization, $\mathcal{Q}$ conducts the corresponding dequantization, and $f_d(\cdot)$ represents the decoding and postprocessing function after dequantization, e.g., an inverse operation of $f_e(\cdot)$.

- We propose a transformer-based CSI feedback scheme.

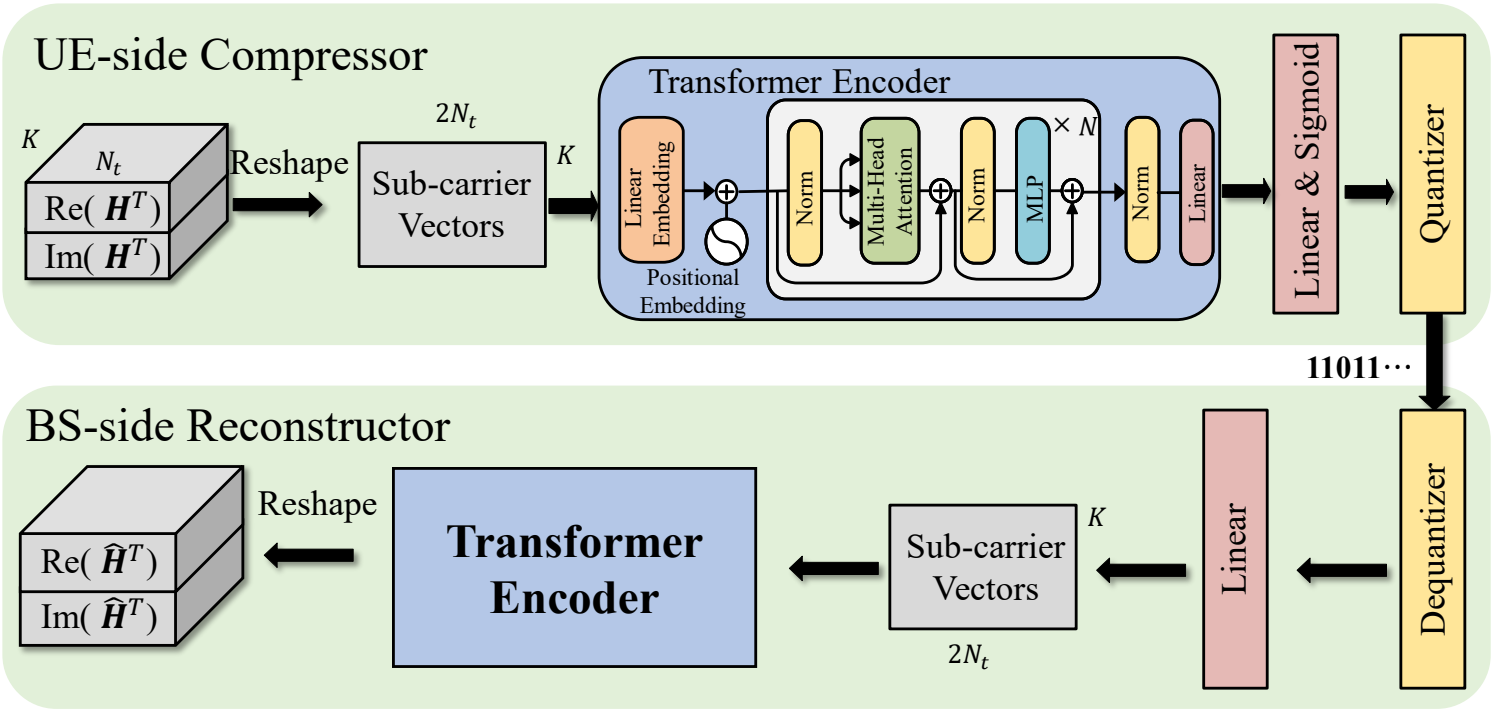
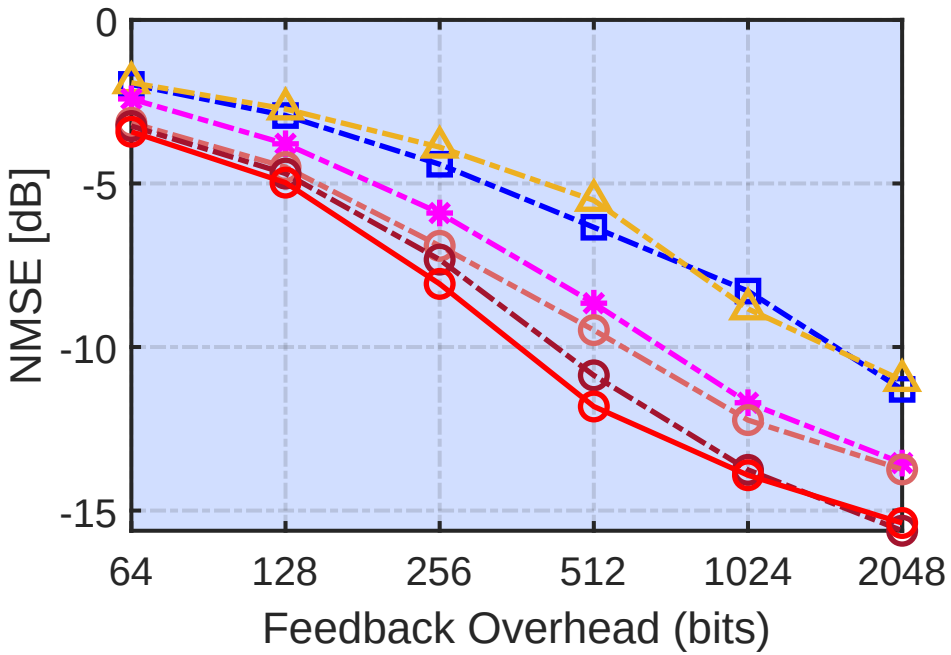


Fig. 8. The transformer-based CSI feedback architecture.

- **Simulation Setting**
  We use the same simulation parameters of channel estimation.



Model Name (512bit)	Size	FLOPs	Runtime
MLP	54.0 MB	28.34 M	2.8ms
LSTM	22.1 MB	124.8 M	18.8ms
bit-level CsiNet	21.1 MB	13.76 G	66.4ms
Transformer-S	22.6 MB	124.0 M	4.7ms
Transformer-M	50.7 MB	470.2 M	6.8ms
Transformer-L	105 MB	1377 M	17.2ms

Fig. 9. NMSE performance comparison of different CSI feedback schemes vs. feedback overhead.

Outline

Hybrid Beamforming

- The hybrid analog-digital MIMO architecture has much lower cost than that of fully-digital architecture. However, HBF optimization is significantly more challenging due to the constant modulus constraint on the analog beamformer.

- Recently, DL-inspired beamforming has been proposed, whereby prior information is captured from radio channel measurements, such as CNN-based [14], MLP-based [15] HBF schemes.

- Herein, we present a transformer-based HBF scheme to obtain more efficient spectral efficiency from limited CSI feedback bits

Problem Formulation

- In the downlink data transmission stage, the signal $y[u,k]$ received at the u-th terrestrial user on the k-th subcarrier can be expressed as

$$\begin{aligned} y[u,k] =& \boldsymbol{h}^H[u,k]\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{f}_{\mathrm{BB}}[u,k]s[u,k] \\ &+ \sum_{u' \neq u} \boldsymbol{h}^H[u,k]\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{f}_{\mathrm{BB}}[u',k]s[u',k] + n[u,k], \forall k, \end{aligned} \tag{5}$$

  where $u = 1, 2, \cdots N_u$, $k = 1, 2, \cdots K$, $\boldsymbol{h}[k,n] \in \mathbb{C}^{N_t \times 1}$ represents the downlink channel vector between the aerial BS and the u-th terrestrial user on the k-th subcarrier, and $n[u,k] \sim \mathcal{CN}\left(0, \sigma_n^2\right)$ is the additive white Gaussian noise (AWGN).

- Thus, the signal-to-interference plus-noise-ratio (SINR) of the u-th UE on the k-th subcarrier can be expressed as

$$\mathrm{SINR}[u,k] = \frac{\left|\boldsymbol{h}^H[u,k]\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{f}_{\mathrm{BB}}[u,k]\right|^2}{\sum_{u' \neq u}\left|\boldsymbol{h}^H[u,k]\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{f}_{\mathrm{BB}}[u',k]\right|^2 + \sigma_n^2}. \tag{6}$$

  Therefore, the sum rate R in the downlink multi-user transmission can be expressed as

$$R = \frac{1}{K}\sum_{u=1}^{N_u}\sum_{k=1}^{K}\log_2\left(1 + \mathrm{SINR}[u,k]\right). \tag{7}$$

Transformer-Based Hybrid Beamforming

- We propose a transformer-based hybrid beamforming scheme.

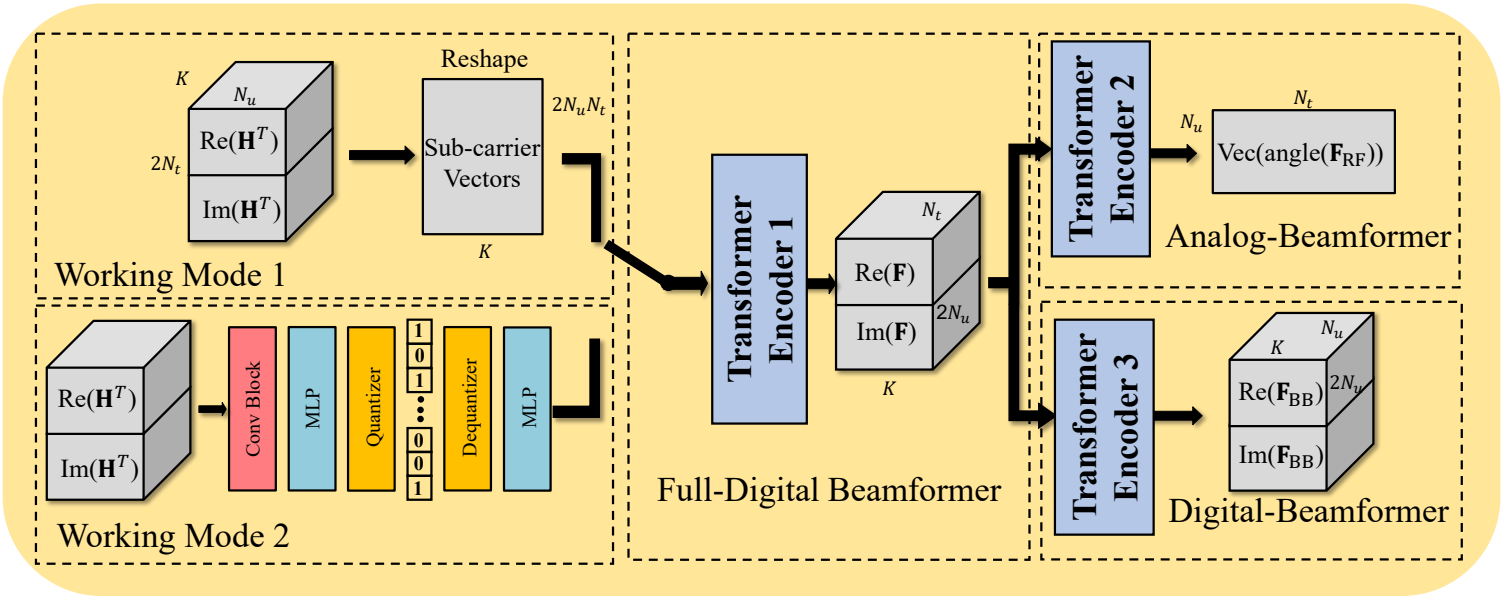
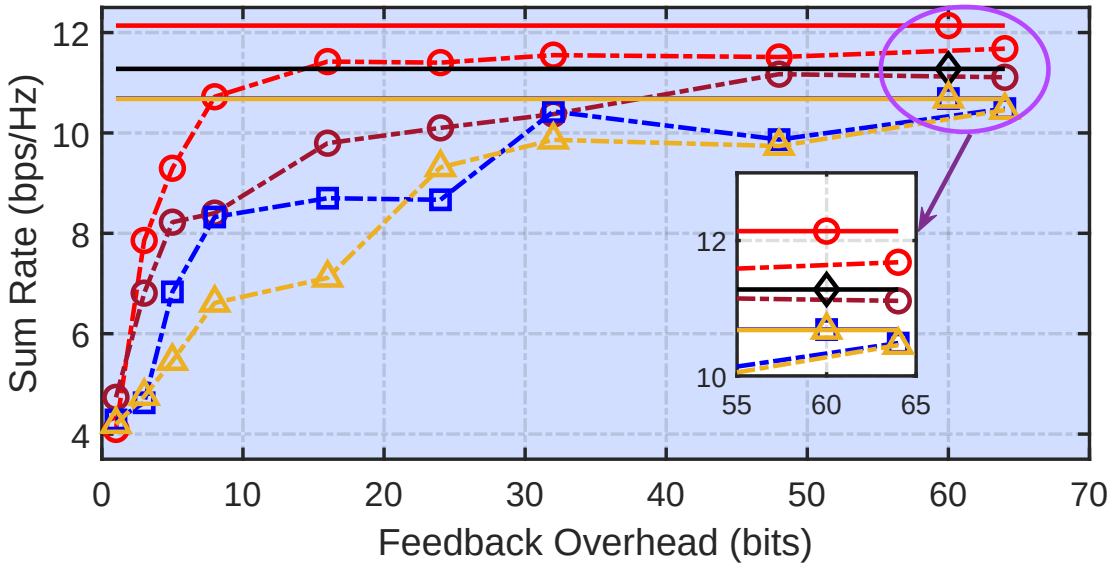


Fig. 10. The transformer-based HBF architecture.

- **Simulation Setting**
  We use the same simulation parameters of channel estimation. And we set the number of UEs to $N_u = 2$. We consider the sum rate as the performance metric.



Model Name	Size	FLOPs	Runtime
Transformer (perfect CSI at the BS)	105 MB	1.842 G	16.4ms
Transformer (limited feedback bits)	170 MB	1.866 G	21.3ms
Transformer-S (limited feedback bits)	51.4 MB	109.1 M	10.4ms
SS-HP (perfect CSI at the BS)	-	72.40 M	96.1ms
MLP (perfect CSI at the BS)	74.8 MB	39.22 M	3.20ms
MLP (limited feedback bits)	111 MB	63.38 M	6.20ms
CNN (perfect CSI at the BS)	37.7 MB	44.10 M	4.40ms
CNN (limited feedback bits)	74.0 MB	68.26 M	7.30ms

Fig. 11. Sum-rate vs. feedback overhead for different HBF schemes.

Outline

Semantic Communication

- Current communication networks do not take into account the meaning or the purpose of the delivered bits, whose interpretation and processing have been left to higher layers.
- The recently growing trend of semantic communication aims at accurately recovering the statistical structure of the underlying source signals and designing the communication system in an end-to-end fashion, similarly to joint source and channel (JSC) coding by taking the source semantics into account [4, 5, 16].
- The semantic communication transmitter includes a semantic encoder and a semantic-aware JSC encoder, and the receiver includes a semantic-aware JSC decoder and a semantic decoder.
- In general, the transmitter can perform semantic encoding on the source according to the knowledge library for obtaining highly compressed abstract semantics, followed by JSC encoder and subsequent baseband signal processing. The receiver follows the reverse steps of the transmitter, where a JSC decoder is followed by a semantic decoder based on some knowledge library.

- We show a transformer-based semantic communication scheme.

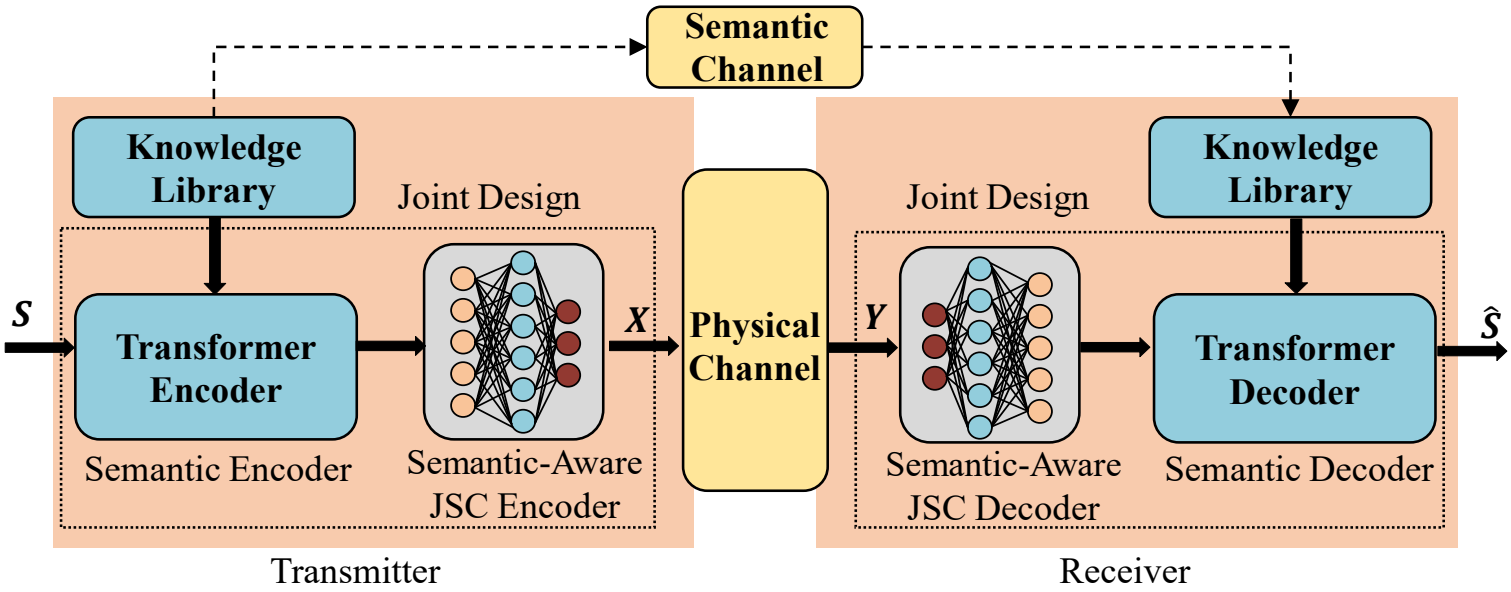


Fig. 12. The transformer-based semantic communication architecture proposed in [5].

- **Benchmarks**
  1. Huffman code followed by Reed-Solomon (RS) coding and 64-QAM
  2. Fixed-length code (5-bit) followed by RS coding and 64-QAM
  3. Huffman code followed by Turbo coding and 64-QAM
  4. 5-bit code followed by Turbo coding and 128-QAM
  5. Brotli code followed by Turbo coding and 8-QAM
  6. The JSC coding approach of [16]

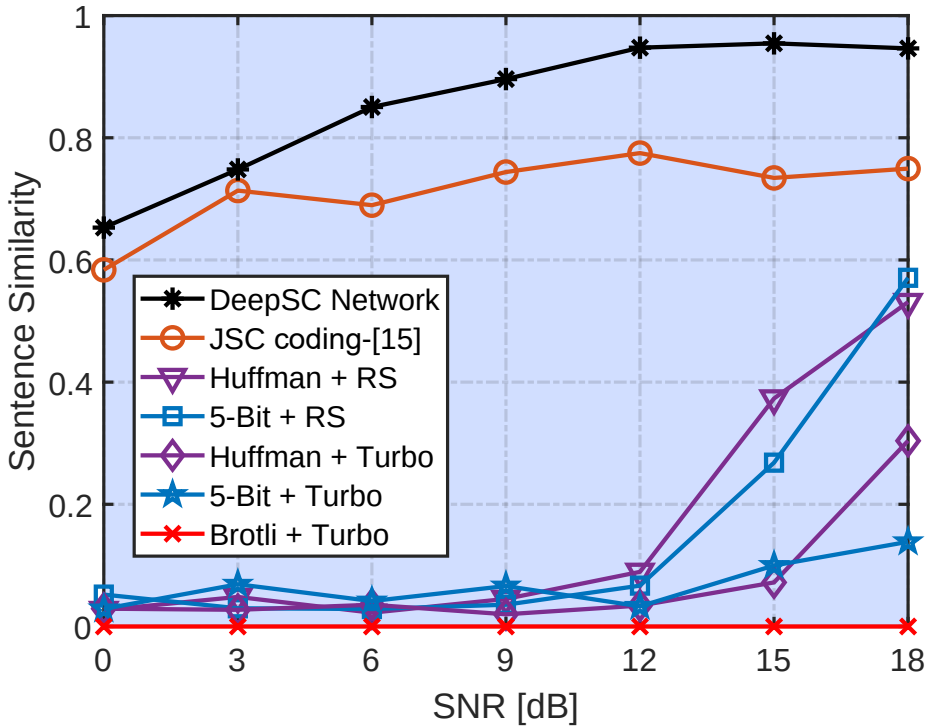


Fig. 13. Sentence similarity of various schemes vs. SNR for the same total number of transmitted symbols over the Rayleigh fading channel quoted from [5].

Outline

Challenges and Open Issues

- **Network Efficiency and Generalization**
  To successfully apply the transformer architecture to 6G networks, an important research challenge is to optimize their computational efficiency and generalization capabilities by developing effective and efficient transformer architectures targeting wireless applications.
- **Efficient Information Injection**
  How to efficiently feed the underlying input, which can include the source signal, CSI tensor, location, traffic and environment information, input the transformer architecture is one of the topics to be investigated for wireless applications.
- **Combination with Model-Driven DL**
  Integrating the transformer architecture into the model-driven framework is a promising approach to further mitigate the performance degradation caused by model inaccuracies.
- **Parallel Communication Sequential Tasks**
  Transformers are also expected to be successful in communication tasks involving temporal sequences, such as channel prediction and beam tracking.

Outline

Conclusions

- In this article, we have presented the transformer architecture and provided examples to highlight its potential benefits in addressing various challenges for 6G intelligent networks.

- Considering the applications of transformers from massive MIMO processing to semantic communication, we provided concrete examples to show their competitive performance compared to the other classical as well as recently proposed DL-based models.

- Potential research directions have also been identified to encourage efforts by the research community to further develop a transformer-based 6G intelligent network paradigm.

Outline

References I

[1] Y. C. Eldar, A. Goldsmith, D. Gündüz, and H. V. Poor, *Machine Learning and Wireless Communications*. Cambridge, UK: Cambridge University Press, 2022.

[2] Z. Qin, H. Ye, G. Y. Li, and B. F. Juang, "Deep learning in physical layer communications," *IEEE Wirel. Commun.*, vol. 26, no. 2, pp. 93-99, Apr. 2019.

[3] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, (Long Beach, CA, USA), Dec. 4-9, 2017, pp. 5998–6008.

[4] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567-579, Sep. 2019.

[5] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663-2675, 2021.

[6] M. Borgerding, P. Schniter, and S. Rangan, "AMP-inspired deep networks for sparse linear inverse problems," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4293–4308, Aug. 2017.

References II

[7] X. Ma, Z. Gao, F. Gao, and M. Di Renzo, "Model-driven deep learning based channel estimation and feedback for millimeter-wave massive hybrid MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2388-2406, Aug. 2021.

[8] X. Ma and Z. Gao, "Data-driven deep learning to design pilot and channel estimator for massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5677-5682, May 2020.

[9] M. B. Mashhadi and D. Gündüz, "Pruning the pilots: Deep learning-based pilot design and channel estimation for MIMO-OFDM systems," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 10, pp. 6315-6328, Oct. 2021.

[10] C. Wen, W. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wirel. Commun. Lett.*, vol. 7, no. 5, pp. 748-751, Oct. 2018.

[11] C. Lu, W. Xu, S. Jin, and K. Wang, "Bit-level optimized neural network for multi-antenna channel quantization," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 1, pp. 87-90, Jan. 2020.

[12] C. Lu, W. Xu, H. Shen, J. Zhu, and K. Wang, "MIMO channel information feedback using deep recurrent network," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 188-191, Jan. 2019.

References III

[13] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wirel. Commun.*, vol. 13, no. 3, pp. 1499-1513, Mar. 2014.

[14] A. M. Elbir and K. V. Mishra, "Low-complexity limited-feedback deep hybrid beamforming for broadband massive MIMO," in *Proc. IEEE 21th Int. Workshop Signal Process. Adv. Wirel. Commun.* (Atlanta, GA, USA), May 26-29, 2020, pp. 1-5.

[15] G. Zhen, M. Wu, C. Hu, F. Gao, G. Wen, D. Zheng, and J. Zhang, "Data-driven deep learning based hybrid beamforming for aerial massive MIMO-OFDM systems with implicit CSI," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 10, pp. 2894-2913, Oct. 2022.

[16] N. Farsad, M. Rao, and A. Goldsmith, "Deep learning for joint source-channel coding of text," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (Calgary, AB, Canada), Apr. 15-20, 2018, pp. 2326-2330.

[17] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey" arXiv preprint arXiv:2009.06732, 2022.

Thanks for your attention!

Q & A