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

Yang Wang, Zhen Gao, Dezhi Zheng, Sheng Chen, Fellow, IEEE, Deniz Gündüz, Fellow, IEEE, and H. Vincent Poor, Life Fellow, IEEE

Abstract—6G wireless networks are foreseen to speed up the convergence of the physical and cyber worlds and to enable a paradigm-shift in the way we deploy and exploit communication networks. Machine learning, in particular deep learning (DL), is going to be one of the key technological enablers of 6G by offering a new paradigm for the design and optimization of networks with a high level of intelligence. In this article, we introduce an emerging DL architecture, known as the transformer, and discuss its potential impact on 6G network design. We first discuss the differences between the transformer and classical DL architectures, and emphasize the transformer’s self-attention mechanism and strong representation capabilities, which make it particularly appealing in tackling various challenges in wireless network design. Specifically, we propose transformer-based solutions for massive multiple-input multiple-output (MIMO) systems and various semantic communication problems in 6G networks. Finally, we discuss key challenges and open issues in transformer-based solutions, and identify future research directions for their deployment in intelligent 6G networks.

I. INTRODUCTION

The sixth generation (6G) of wireless cellular networks are expected to cover the cyber and physical worlds by allowing humans to seamlessly interact with each other and with a variety of devices in a mixed reality metaverse through connected intelligence. These new and fascinating applications impose challenging requirements and constraints on communication networks, including ultra-high reliability, ultra-low latency, extremely high data rate, substantially high energy and spectral efficiency, ultra-dense connectivity, and a high level of intelligence [1]. These stringent demands of 6G have driven researchers to look for sophisticated physical layer techniques that would go beyond the cycle of incremental improvements. Current wireless networks have been largely designed as a combination of dedicated processing blocks, such as channel estimation, equalization, coding/decoding blocks, where each block is designed separately on the basis of mathematical models that define the statistical behavior of the wireless channels and the underlying data traffic. However, this model- and block-based design approach is facing increasing challenges in the complex and diversified scenarios the future 6G networks will operate in. The diversity of the devices and hardware technologies, increasing co-existence requirements, and the variety of traffic and service demands make such modeling approach difficult and inaccurate. In addition, with the deployment of ultra-massive multiple-input multiple-output (MIMO) systems and large-scale reconfigurable intelligent surfaces (RISs), the optimization of physical layer functionalities based on rigid mathematical models and solutions become prohibitive due to the computational complexity and associated control overheads. Therefore, conventional mathematical models and solutions cannot provide the required dramatic enhancement in the capacity and performance of wireless networks.

While DL-based solutions are appealing, the actual deployment is still very difficult since they have limited performance gain at the expense of training efficiency and a large dependence on hyperparameter optimization. Therefore, proposing a more efficient and widely applicable DL architecture is essential for solving complex communication problems. A novel DNN structure, called the transformer, has emerged recently, and achieved remarkable success in a variety of natural language processing (NLP) and computer vision (CV) tasks [5]. The transformer architecture is built upon the self-attention mechanism, which relates different parts of a data sequence for a more accurate representation of the sequence. Self-attention layers in the transformer architecture enable a global receptive field, and the multi-head mechanism ensures that the network can pay attention to multiple discriminative parts of the inputs. By highlighting the transformer’s multi-model fusion and feature representation capabilities, this article explores its application to 6G intelligent network design, and proposes a new transformer-based intelligent processing architecture, particularly focusing on massive MIMO and semantic communication applications. On the other hand, we expect to see the transformer architecture to find applications in many other components of future data-driven solutions for 6G networks, particularly in end-to-end designs [6].

The rest of the article is organized as follows. The following section briefly introduces the connection between DL and wireless communications. Then, we provide the important basis for the novel DNN structure called transformer. Next,
we design and propose a transformer-based architecture for 6G intelligent processing. We then discuss open research issues in transformer-empowered 6G intelligent networks and conclude the article.

II. OVERVIEW OF DEEP LEARNING AND THE TRANSFORMER ARCHITECTURE

DL is a powerful computational tool to understand complex data representations and patterns, and as such, offers a new paradigm to tackle complicated problems in 6G intelligent network design. In this section, we briefly provide some background on popular DNN architectures and their applications in wireless communications.

A. Common Neural Network Architectures

Classic neural network architectures include multi-layer perception (MLP), convolutional neural network (CNN), recurrent neural network (RNN), and stacked autoencoder (SAE).

MLP is an artificial neural network that consists of at least three layers of fully-connected neurons, which requires to configure a substantial number of connection weights. MLP-based solutions have been developed to address various wireless communication problems, such as channel estimation and beamforming [3]. It has been observed that deeper networks typically provide better generalization; however, training deep fully-connected networks suffers from high complexity and low convergence efficiency.

To reduce the training complexity, CNNs employ a set of locally connected kernels, rather than fully-connected layers, to capture local correlations between different data regions. Compared with MLP, CNN reduces the number of model parameters significantly and maintains the affine invariance by leveraging three important ideas: sparse interactions, parameter sharing, and equivariant representations. By treating the channel matrices as two-dimensional images, CNNs have shown great potential for tasks such as channel estimation, channel state information (CSI) feedback, beamforming [3], as well as semantic communications [4].

RNNs constitute another class of DNN architectures that exploit sequential correlations between samples. At each step, it produces the output via recurrent connections between hidden units. However, the traditional RNN architecture is slow to train, and suffers from vanishing and exploding gradients. Long short-term memory (LSTM) architecture mitigates these problems by introducing a set of gates, which allows memory to be restored across longer sequences. Recently, there have been several works utilizing LSTMs to extract temporal correlations of data, (e.g., in channels with memory) for communication system design [3].

SAE architecture consists of hierarchically connected multiple autoencoders. Its basic component, autoencoder, contains two parts: an encoder that acquires a low-dimensional representation of input and a decoder that reconstructs the input from the compressed vector. SAE is widely used to extract effective features and patterns that contain essential and compressed information about data. From a learning perspective, the entire communication system can be viewed as an end-to-end SAE, and its multiple sub-modules can also be viewed as SAEs, including pilot design and channel estimation, CSI feedback, and semantic communications [3], [4]. Thus, SAE is a core DNN structure for many of the current DL-based communication system components.

B. Self-Attention and Transformer

Although MLP, CNN, RNN, and SAE have been widely utilized in intelligent processing for communication systems with some success, efforts continue to push the boundaries of DL models in practical communication systems. Recently, the evolution of DNN architectures in NLP has led to a prevalent architecture known as the transformer [5]. We argue
that the transformer holds a great potential also in the design of intelligent communication systems.

As shown in Fig. 1, the transformer is a sequence-to-sequence DNN model and consists of an encoder module and a decoder module with several encoder/decoder layers of the same architecture. The input and output sequences are converted to vectors of dimension $d_{\text{model}}$ by embedding and positional encoding layers. Each encoder/decoder layer has the same structure, and is mainly composed of a self-attention sub-layer and a position-wise MLP sub-layer, while each decoder also contains a masked attention sub-layer before the self-attention sub-layer. For building a deep model, a residual connection is employed around each sub-layer, followed by a layer normalization module.

Self-attention, also called intra-attention, is an attention mechanism relating different positions in a single sequence to compute a representation of the sequence, which can also be regarded as a non-local filtering operation. In the single-head self-attention layer, the input sequence $X \in \mathbb{R}^{n \times d_{\text{model}}}$ is first transformed into three different sequential vectors: the query $Q \in \mathbb{R}^{n \times d_k}$, the key $K \in \mathbb{R}^{n \times d_k}$ and the value $V \in \mathbb{R}^{n \times d_v}$ by three different linear matrices, where $n$ is the length of the sequential vectors, $d_k$ and $d_v$ are the dimensions of query, key, and value subspace, respectively. Subsequently, as shown in Fig. 1, the scale dot-production attention operation generates the attention weights by aggregating the query and the corresponding key. The resulting weights are assigned to the corresponding value, yielding the output vectors. Instead of performing single-head self-attention with query, key, and value, multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. Specifically, different heads use different three group linear matrices, and these matrices can project the input vectors into multiple feature subspaces (i.e., $\{Q_i\}_{i=1}^h$, $\{K_i\}_{i=1}^h$, and $\{V_i\}_{i=1}^h$, where $h$ is the number of heads) and processes them by several parallel attention heads (layers). The resulting vectors are concatenated and mapped to the final output.

The position-wise MLP sub-layer is a fully-connected feedforward module that operates separately and identically on each position. This module consists of two linear transformations with a ReLU activation, where the parameters are shared across different positions. Since the transformer does not introduce recurrence or convolution, it has no knowledge of positional information (especially for the encoder). Thus, additional positional information is introduced through positional encoding in order to model the relative positions of the input sequences.

Compared with CNN/RNN models, the transformer makes few assumptions about the underlying structure of data, which makes it a universal and flexible architecture. The non-sequential nature of the transformer allows it to capture long-range dependencies in the input through self-attention. Its remarkable success in the fields of NLP and CV has inspired communication researchers to investigate its applications to various communication problems. Not surprisingly, transformers have also shown remarkable success in certain communication tasks, specifically, for channel estimation and semantic communication [7], [8].

III. TRANSFORMER FOR 6G INTELLIGENT PROCESSING

Massive MIMO is an essential physical layer technology to accommodate the exponential growth of mobile data traffic. Fig. 2(a) illustrates a generic communication system, divided into two parts: the MIMO processing part and source & channel coding part. The former includes pilot design, channel estimation, CSI feedback, and hybrid beamforming (HBF). The latter is composed of source coding and channel coding.
Based on the advanced DL model introduced earlier, we seek to expand the applicability of the transformer to serve as a general-purpose backbone for these crucial modules. Herein, as illustrated in Fig. 2 (b), we propose the novel 6G intelligent processing architecture employing transformer, including the massive MIMO intelligent processing blocks and the newly emerging semantic communication blocks.

A. Channel Estimation

Accurate CSI at the base station (BS) is critical for beamforming and signal detection in massive MIMO systems. However, CSI acquisition overhead of conventional orthogonal pilot approaches increases linearly with the number of antennas. To reduce the pilot overhead, existing 5G NR standard pilot approaches increases linearly with the number of antennas. However, it is challenging to accurately estimate the high-dimensional channels from low training overhead. By exploiting the sparsity of the channels in the angular domain and/or delay domain, compressive sensing (CS)-based channel estimation solutions have been proposed to overcome this issue. Nevertheless, since the dimension of the CSI to be estimated is extremely large, the involved matrix inversion operations and the iterative nature of CS-based techniques result in prohibitively high computational complexity and storage requirements.

Researchers have resorted to DL techniques to solve the aforementioned problems. A multiple-measurement-vector learned approximate message passing (MMV-LAMP) network was proposed in [9] to reconstruct the spatial-frequency channel matrix by exploiting the channel’s structured sparsity. The authors of [10] proposed an end-to-end DNN architecture to jointly design the pilot signals and channel estimator. Moreover, a CNN module combined with non-local attention layer is employed in [11] to exploit longer range correlations in the channel matrix.

Nevertheless, most existing DL-based channel estimation solutions are based on the MLP and CNN architectures. Here, we propose a novel channel estimator that utilizes the universal and flexible transformer architecture, as illustrated in Fig. 3 (a). Specifically, the proposed transformer-based solution includes a dimensionality reduction network for pilot design and a reconstruction network for channel estimation. Similarly to [11], we consider frequency-selective MIMO channel matrix as the input of the network, which allows us to simultaneously utilize the channel’s frequency-domain correlation and the angle-domain sparsity to significantly improve the channel estimation performance. If the number of radio-frequency (RF) chains is at least twice the total number of data streams, the HBF structure can realize any fully-digital beamformer exactly [10]. Thus, we exploit the fully-connected linear layer to represent the pilot sequences. More importantly, in our channel estimation module, the encoder part of the transformer is exploited to reconstruct the channel for enhancing the estimation accuracy.

To evaluate the performance of the proposed transformer-based channel estimator, we investigate the downlink channel estimation problem in $M$ successive time slots, where 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 $K = 32$, and the channel estimation compression ratio is $\rho = \frac{M}{N_t} = \frac{3}{8}$. We consider a sparse channel scenario with $N_c = 6$ clusters, $N_p = 10$ paths per cluster, and an angle spread of $\Delta \theta = \pm 3.75^\circ$. We generate the training, validation and test sets of 100,000, 10,000, 5,000 samples, respectively. We choose the normalized mean square error (NMSE) as the performance metric.

To illustrate the advantages of our proposed channel estimator, denoted as ‘Transformer’ in Fig. 3, we choose four benchmark channel estimators. The first one is the traditional simultaneous orthogonal matching pursuit (SOMP) based estimator, denoted as ‘SOMP’. The second and third are the conventional DL-based channel estimators, namely, the MMV-LAMP based estimator [9] and the DNN-based estimator [10], denoted as ‘MMV-LAMP’ and ‘DNN’, respectively. We also utilize a state-of-the-art attention-CNN based channel estimator [11],
cosine functions of different frequencies to represent the layer to linearly embed the channel data and use sine and cosine functions of different frequencies to represent the relative positions of the sub-carriers. Then, the encoder of the transformer extracts the features from the channel data embedded with the positional information. Next, the features are vectorized, and a fully-connected linear layer is used to generate a real-valued compressed codeword. The codeword is then converted to the feedback bit stream through a quantization layer. Since the whole network structure corresponds to the compression recovery task, the decoder adopts the same structure as the encoder. Thus, the feedback bits received by the BS are input into a dequantization layer, and then the decoder module is used to reconstruct the complete CSI at the BS.

B. CSI Feedback

CSI feedback is necessary to reconstruct the downlink CSI at the BS in frequency-division duplex (FDD) systems. For time-division duplex (TDD) systems, by exploiting channel reciprocity, the transmitter may estimate the downlink CSI from the uplink CSI. But such reciprocity relies on many ideal factors, including the accurate calibration of the transceiver RF chains at both the BS and UE. For massive MIMO, the perfect uplink and downlink reciprocity is difficult to achieve, and the BS has to rely on CSI feedback for both FDD and TDD operations. However, the large number of antennas results in excessive feedback overhead. Similarly to channel estimator, 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 compression and recovery has achieved good results in CSI feedback. In [12], CsiNet, an autoencoder architecture, was proposed to reduce feedback overhead in massive MIMO systems. It has been shown that CsiNet remarkably outperforms traditional CS-based methods in terms of both compression ratio and recovery accuracy [12]. Subsequent studies expanded the original scope of the network and designed various network models based on CNN and LSTM architectures to handle different CSI feedback problems [3].

Herein, we present a transformer-based CSI feedback scheme to obtain more efficient quantization and compression performance compared with the aforementioned methods. As illustrated in Fig. 4 (a), we utilize the fully-connected linear layer to linearly embed the channel data and use sine and cosine functions of different frequencies to represent the

<table>
<thead>
<tr>
<th>UE-side Compressor</th>
<th>BS-side Reconstructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Re}(H^T) )</td>
<td>( \text{Re}(H^T) )</td>
</tr>
<tr>
<td>( \text{Im}(H^T) )</td>
<td>( \text{Im}(H^T) )</td>
</tr>
<tr>
<td>Reshape</td>
<td>Reshape</td>
</tr>
<tr>
<td>Sub-carrier Vectors</td>
<td>Sub-carrier Vectors</td>
</tr>
<tr>
<td>Transformer</td>
<td>Transformer</td>
</tr>
<tr>
<td>Linear &amp; Sigmoid</td>
<td>Linear &amp; Sigmoid</td>
</tr>
<tr>
<td>Dequantizer</td>
<td>Dequantizer</td>
</tr>
<tr>
<td>Linear</td>
<td>Linear</td>
</tr>
<tr>
<td>Norm</td>
<td>Norm</td>
</tr>
<tr>
<td>EmbeddingLinear</td>
<td>EmbeddingLinear</td>
</tr>
<tr>
<td>Positional Embedding</td>
<td>Positional Embedding</td>
</tr>
<tr>
<td>Reshape</td>
<td>Reshape</td>
</tr>
<tr>
<td>Vectorization</td>
<td>Vectorization</td>
</tr>
</tbody>
</table>

Fig. 4. (a) The transformer-based CSI feedback architecture; and (b) NMSE performance comparison of different CSI feedback schemes versus feedback overhead.

<table>
<thead>
<tr>
<th>Feedback Overhead (bits)</th>
<th>NMSE [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>-16</td>
</tr>
<tr>
<td>128</td>
<td>-14</td>
</tr>
<tr>
<td>256</td>
<td>-12</td>
</tr>
<tr>
<td>512</td>
<td>-10</td>
</tr>
<tr>
<td>1024</td>
<td>-8</td>
</tr>
<tr>
<td>2048</td>
<td>-6</td>
</tr>
</tbody>
</table>

We use the same simulation parameters of Subsection III-A to generate the CSI data for evaluating the proposed transformer-based CSI feedback, denoted as ‘Transformer’. Three benchmark schemes are adopted for comparison. The first one is the MLP-based CSI feedback, ‘MLP’, where the encoder and decoder consist of three fully-connected layers, respectively. The second one is the ‘CsiNet’ scheme in [12], while the third is the ‘Enhanced-CsiNet’ scheme, which improves the CsiNet with a deeper residual network. We again use the NMSE metric for performance evaluation. Fig. 4 (b) shows that the transformer-based CSI feedback scheme dramatically outperforms the other three benchmarks. In a nutshell, by utilizing the transformer, the implicit features in the CSI are better extracted and fewer feedback bits are needed to reconstruct the CSI at the BS, which enhances the CSI reconstruction accuracy and reduces the feedback overhead and latency.

C. Hybrid Beamforming

Conventional massive MIMO with fully-digital architecture requires a dedicated RF chain for each antenna, which imposes excessive power consumption and extremely high RF hardware cost. To this end, HBF, which utilizes a limited number of RF chains to connect the digital baseband precoder/combiner and the analog RF precoder/combiner, has emerged as a viable solution with affordable hardware costs and power consumption. However, the HBF optimization is very difficult due to the constant modulus constraint on analog beamformer [13].
Many model-based solutions have been proposed to tackle this challenge. For instance, the authors of [13] proposed spatial sparse hybrid precoding (SS-HP) to achieve a near fully-digital performance by exploiting channel sparsity.

However, model-based HBF algorithms require time consuming optimization iterations to obtain near-optimal solutions. Moreover, they demand either perfect downlink CSI or a codebook with an accurate sparse basis, which are difficult to acquire in practice. To overcome these issues, DL-inspired beamforming has been proposed. Specifically, DL can approach the optimal solution for supporting adaptive and real-time massive MIMO beamforming, whereby the prior information is captured from the radio channel measurements. In [14], the authors proposed a CNN-based HBF architecture that can be trained to maximize the spectral efficiency with imperfect CSI.

To the best of our knowledge, the existing DL-based HBF schemes mainly adopt the MLP or CNN architectures, and there is still a large gap between their performance and the optimal one. To exploit the powerful transformer architecture, we propose a transformer-based HBF scheme, composed of three transformer modules. According to [13], analog RF precoding and digital baseband precoding can be optimized to approach the optimal fully-digital precoding. Motivated by this principle, each transformer in Fig. 5 (a) implements a part of the HBF optimization. More specifically, the first, second, and third transformers represent the fully-digital precoding, the analog RF precoding, and the digital baseband precoding, respectively. By introducing the structural prior information of traditional optimization methods, combined with the transformer’s feature extraction ability based on the self-attention mechanism, we can achieve better performance than traditional as well as existing DL-based methods.

To illustrate the superior performance of our transformer-based HBF, we use the channel parameters similar to those of Subsection III-A. We set the number of UE to $N_u = 2$. As shown in Fig. 5 (a), we consider both the perfect CSI and the compressed CSI feedback as the input to HBF, respectively. We choose three benchmarks for comparison, namely SS-HP from [13], CNN-based HBF of [14], and a MLP-based HBF, which is obtained by replacing the transformer layers in our solution with fully-connected layers.

The sum rate comparison of different HBF schemes is depicted in Fig. 5 (b). It can be seen that the transformer-based HBF scheme significantly outperforms SS-HP and other DL-based HBF schemes with both perfect and limited CSI feedback. The performance gains over the benchmarks are particularly noticeable at low feedback overhead of 3 to 24 bits. Moreover, the proposed scheme with limited feedback bits even outperforms SS-HP with perfect CSI, when the feedback overhead is greater than 24 bits. This demonstrates the effectiveness of the proposed transformer-based HBF architecture, particularly under the practical scenario of limited feedback.

D. Semantic Communication

Our communication networks have been traditionally conceived and designed as bit pipes; that is, the goal has been to deliver as many bits as possible with the highest reliability. 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. To meet the requirements of 6G wireless networks, however, it is important to propose more efficient information acquisition and delivery methods. The recently growing trend of semantic communications aims at accurately recovering the statistical structure of the underlying information of the source signal 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], [8], [15]. Fig. 6 (a) shows the general framework of a semantic communication model, where the transmitter includes a semantic encoder and a semantic-aware JSC encoder, and the receiver includes a semantic-aware JSC decoder and a semantic decoder. Specifically, the transmitter performs semantic encoding on the source according to the knowledge library for obtaining highly compressed abstract semantics, and then performs 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.

Semantic communication is particularly effective for complex information sources, such as text, speech, image, or video, where the reconstruction quality depends on the source semantics, and is often difficult to measure through traditional measures of bit error rate or mean square error. In [15], the authors proposed an LSTM-based model to extract the
semantic information of sentences through JSC coding for text transmission. However, due to the lack of a separate semantic coding module, JSC coding can only implicitly utilize the semantic information, which has difficulty to represent specific semantics. Recently, a DL-enabled semantic communication (DeepSC) scheme was proposed in [8], where a separate semantic coding network is utilized to better extract accurate semantic information. Specifically, as shown in Fig. 6 (a), a transformer encoder is utilized as the semantic encoder and an MLP is used as the JSC encoder. The Rayleigh fading channel is interpreted as an untrainable layer in the model. Correspondingly, the receiver consists of an MLP-based JSC decoder followed by a transformer-based semantic decoder for text reconstruction. The whole network is trained in an end-to-end fashion to simultaneously minimize the sentence similarity and maximize the mutual information. Fig. 6 (b) compares the performance of the DeepSC network [8] in transmitting text over a Rayleigh fading channel with the following benchmarks: Huffman code followed by Reed-Solomon (RS) coding and 64-quadrature amplitude modulation (QAM), fixed-length code (5-bit) followed by RS coding and 64-QAM, Huffman code followed by a Turbo coding and 64-QAM, 5-bit code followed by a Turbo coding and 128-QAM, Brotli code followed by a Turbo coding and 8-QAM, and the JSC coding approach of [15]. The details of this experiment can be found in [8]. The simulation results demonstrate that thanks to the powerful transformer architecture, the sentence similarity performance of DeepSC [8] is far superior than all traditional approaches based on separate compression followed by channel coding, as well as the JSC coding approach [15]. Hence, we foresee that semantic-aided communication is another important challenge, where the transformer architecture is likely to have an impact on the future communication system design as it will more effectively learn and adapt to the statistics of complex signals, such as text, image or video.

IV. CHALLENGES AND OPEN ISSUES

We hope that the above examples have convinced the readers of significant potential of the transformer architecture for various components of future 6G intelligent networks. In addition to these examples, we expect that the transformers will find applications in waveform design, localization, RIS-assisted terahertz communication systems, as well as more advanced channel estimation techniques exploiting other information sources such as LiDAR or cameras. We would like to highlight that the transformer architecture was invented only in 2017. Although it has received significant attention in the last years thanks to its superior performance, the research on transformer-based communication system design is still in its infancy, and many key issues are still open. In this section, we discuss several potential directions for future study.

Network Efficiency: An important limitation of the transformer architecture is its high computation and memory complexity, mainly due to the self-attention module. This results in a significant increase in the training time and energy, and even considerable inference complexity, particularly for long sequences. This can prevent its implementation on resource-limited devices such as mobile phones. Hence, a promising research direction to successfully apply transformers to future intelligent networks is to tame its complexity and memory requirements by developing highly effective and efficient transformer architectures for mobile devices. Achieving better network performance usually is at the expense of higher computational complexity, and determining an appropriate balance between the transformer-enabled network performance and efficiency should be a topic of further study.

Network Generalization: In high-speed time-varying scenarios, such as high-speed railway and low orbit satellites, the model mismatch problem is difficult to deal with due to the large target dynamic range. Moreover, since the transformer makes few assumptions on the structural bias of the input data, the network cannot perform real-time parameter retraining to address these model mismatches. One potential solution is to utilize very large-scale data for pre-training, so that the transformer-based model can learn the knowledge covering wide range of communication scenarios from the data. Then the transformer-based network can be fine tuned online based on the small-scale data obtained in the actual scene to meet the actual communication needs. Another idea can be to introduce structural biases or regularization based on prior information about the communication channel or information sources to accelerate the fine-tuning process.

Network Adaptation: The air-space-ground integrated network is a widely recognized solution to achieve global coverage for 6G networks. Furthermore, a universal network struc-
ture is a prerequisite for the association of heterogeneous communication platforms. However, most of the existing DL-based solutions for communication systems are designed to handle specific communication tasks on individual platforms. The literature on transformers have demonstrated their capability in dealing with multiple tasks in a single model. We believe that more communication tasks across heterogeneous platforms, such as channel estimation, beamforming, signal detection and so on, can be carried out by a single transformer model. Hence, unifying all signal processing and communication tasks (e.g., compression, coding and modulation) in a grand unified transformer-based model is an exciting research challenge for our community.

V. 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. We have considered the applications of the transformer from massive MIMO processing to semantic communication, and provided concrete examples to show its competitive performance compared to the other classical as well as recently proposed DL-based models, hence demonstrating its great potential in designing the AI-native future communication systems. Potential research directions have also been identified to channel the efforts of the research community to the transformer-based 6G intelligent network paradigm.

REFERENCES