

From Multilayer Perceptron to GPT: A Reflection on Deep Learning Research for Wireless Physical Layer

Mohamed Akroud, Amine Mezghani, *Member, IEEE*, Ekram Hossain, *Fellow, IEEE*, Faouzi Bellili, *Member, IEEE*, Robert W. Heath, *Fellow, IEEE*

Abstract—Most research studies on deep learning (DL) applied to the physical layer of wireless communication do not put forward the critical role of the accuracy-generalization trade-off in developing and evaluating *practical* algorithms. To highlight the disadvantage of this common practice, we revisit a data decoding example from one of the first papers introducing DL-based end-to-end wireless communication systems to the research community and promoting the use of artificial intelligence (AI)/DL for the wireless physical layer. We then put forward two key trade-offs in designing DL models for communication, namely, accuracy versus generalization and compression versus latency. We discuss their relevance in the context of wireless communications use cases using emerging DL models including large language models (LLMs). Finally, we summarize our proposed evaluation guidelines to enhance the research impact of DL on wireless communications. These guidelines are an attempt to reconcile the empirical nature of DL research with the rigorous requirement metrics of wireless communications systems.

I. INTRODUCTION

RESEARCHERS are developing use cases where DL can potentially enhance the system performance and reduce complexity/overhead compared to classical methods (see envisioned examples for standardization in the 3GPP Release 18 [1, Section 9.2]). Since the introduction of the AlexNet network in 2012, the use of deep neural networks (DNNs) has skyrocketed within the communications community by substituting conventional optimization solvers with generative and/or discriminative DL techniques. Fig. 1 summarizes some of the key DL models applied to communications problems, starting from the multilayer perception (MLP) model [2] to the diffusion model [3]. We refer the reader to [4] for a comprehensive survey on the applications of DL for wireless physical layer design. In Fig. 1, we also include the recent generative pre-trained transformer (GPT) models [5], [6] as they are currently initiating many discussions within the communications research community about the data compression properties of large language models (LLMs) and their role in replicating digital twins (cf. Section III-B).

Because the sixth-generation (6G) networks are envisaged as multi-band, decentralized, fully autonomous, and hyperflexible user-centric systems encompassing satellite, aerial, terrestrial, underwater, and underground communications, DL techniques are expected to partially or fully substitute classical methods, their assessment metrics should comply with rigorous evaluation guidelines that equally address latency, complexity, generalization, and accuracy. In this paper, we

reflect on the last decade of research on DL for wireless communications with a focus on the physical and link layers. We start by highlighting the limitations of state-of-the-art DL methods for wireless communications. We do so by revisiting one of the first published papers [7] introducing DL-based end-to-end communications systems over additive white Gaussian noise (AWGN) channels. By doing so, we pinpoint how [7], as many of the subsequent DL papers for wireless, turn the spotlight on the accuracy of DL methods at the cost of sacrificing their generalization capabilities. We also highlight how the open literature does not draw enough attention to the important practical considerations for designing DL systems for wireless communications, such as data acquisition and adaptation to new system dimensions. We then describe the two key trade-offs in designing DL-aided wireless communications systems, namely, *accuracy versus generalization* and *compression versus latency*. These two trade-offs offer new evaluation guidelines to assess future DL research directions for wireless communications.

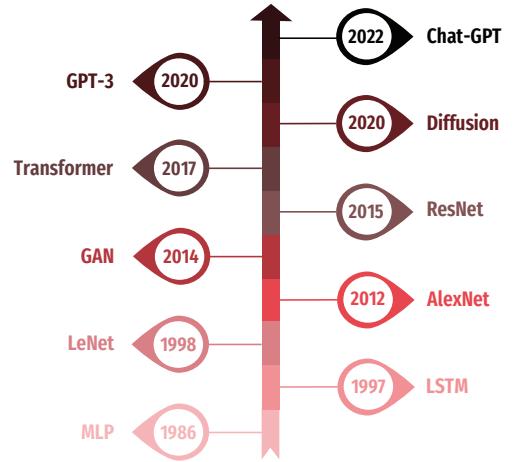


Fig. 1: Publication timeline of the key DL models used and/or being investigated by the communications research community.

II. LIMITATIONS OF THE STATE-OF-THE-ART DL TECHNIQUES FOR WIRELESS COMMUNICATIONS

In this section, we describe the limitations of neglecting the data distribution shift when evaluating DL models for wireless communications at the physical layer. Specifically, we revisit

an example from [7] to illustrate the drawbacks of blindly applying black-box DL models in a plug-and-play manner when only the model accuracy is assessed. We also describe other challenges arising from the use of DL techniques that are usually not examined in the open literature.

A. Evaluation of a Single Metric

It is always possible to beat a classical method that solves a non-closed form problem using DL techniques based on deep neural networks (DNNs) given a known model by generating training and test datasets on which DNNs are *both* trained and evaluated. For this reason, it is critical to use both the accuracy and generalization of DNNs to assess their performance on a variety of:

- in-distribution (ID) scenarios where the training and testing datasets are generated from the same distribution (e.g., the same user speed is assumed to generate mobility data to be used for training and testing).
- out-of-distribution (OOD) scenarios where a distribution shift occurs between training and testing datasets (e.g., different user speeds are assumed to generate mobility data to be used for training and testing).

In other words, DL models that are either accurate and non-robust or highly biased and robust are equally worthless for real-time physical/link layer applications.

To illustrate this idea, we revisit the study [7], which is one of the first papers promising the design of communications systems as an autoencoder for reconstruction tasks that jointly optimize transmitter and receiver components in a single end-to-end process. As shown in Fig. 2, the physical communication chain of the transmitter and the receiver are substituted by an encoder and a decoder, respectively. The channel is represented as a noisy non-parameterized layer that injects an additive Gaussian white noise at a specific energy per bit to noise power spectral density ratio, E_b/N_0 .

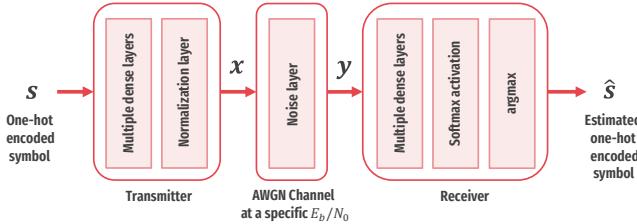


Fig. 2: An end-to-end communication system over an AWGN channel as an autoencoder: the input is the one-hot encoded transmit symbol s while the estimated one-hot symbol \hat{s} is argmax of the softmax probability distribution over all possible messages.

In [7], the training of the autoencoder was performed on a dataset generated at $E_b/N_0 = 7$ dB. However, the evaluation was conducted over a range of E_b/N_0 between $[-4$ dB, 8 dB]. By doing so, the authors obtained a lower block error rate (BLER) than the Hamming code with rate $R = 4/7$ and concluded that the autoencoder “has learned some joint coding and modulation scheme, such that a coding gain is achieved”.

From a machine learning theory perspective, this conclusion is questionable. For this reason, we train multiple autoencoders, each with a training $E_b/N_0 \in \{-4, 0, 5, 7, 8\}$ dB. We then evaluate each autoencoder on the testing range $E_b/N_0 \in [-4$ dB, 8 dB] as shown in Fig. 3. There, all autoencoders in Fig. 3 exhibit a decreasing BLER over the entire test E_b/N_0 range even though they were trained on one single E_b/N_0 value. While this fact suggests that autoencoders over AWGN channels generalize well to out-of-distribution decoding scenarios, it can also indicate that the problem at hand is easy to solve because the distribution shift between the training data distribution associated with $E_b/N_0 = 7$ dB and the test data distributions with $E_b/N_0 \in [-4$ dB, 8 dB] is minimal. To confirm this fact, we report in Table I the percentage of the area overlap between the data distribution of the received signal $y \sim \mathcal{N}\left(\mathbf{x}, \frac{1}{2R E_b/N_0} \mathbf{I}\right)$ for training and testing E_b/N_0 values¹. We observe the fact that lower testing values for E_b/N_0 compared to the training value $E_b/N_0 = 7$ dB yields a lower area overlap, or equivalently, a higher OOD shift. However, it is interesting to note the significantly high overall overlap. This suggests that the AWGN channel is a simplistic model to assess the generalization performance of DL models. One can perceive the AWGN channel model in communication as the MNIST dataset in computer vision on which any great performance is considered obsolete by the computer science community due to the intrinsic simplicity of the handwritten digit classification task.

TABLE I: The percentage of area overlap between the training ($E_b/N_0 = 7$ dB) and testing received signal distributions ($E_b/N_0 \in \{-4, 0, 5, 8\}$ dB).

Test E_b/N_0	-4 dB	0 dB	5 dB	8 dB
Area overlap with train $E_b/N_0 = 7$ dB	45.70%	62.97%	88.91%	94.43%

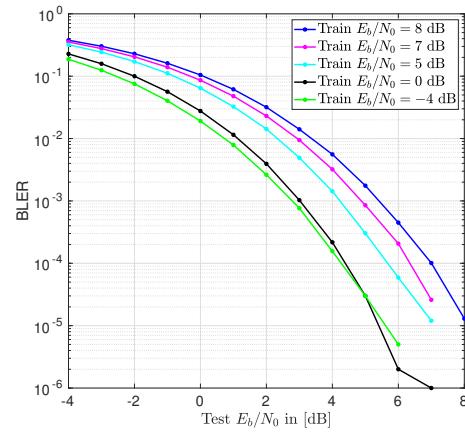


Fig. 3: BLER of the autoencoder versus test E_b/N_0 for multiple training E_b/N_0 .

¹To quantify the shift between two distributions, their area overlap is more informative than their KL-divergence because the latter only accounts for the region where both distributions are non-zero.

It is also unclear whether the chosen value $E_b/N_0 = 7$ dB is the best one to select. It is seen that training autoencoders on lower E_b/N_0 values in $[-4$ dB, 8 dB] (i.e., with higher noise levels) leads to a smaller BLER across the entire test E_b/N_0 interval. This result not only shows that the train $E_b/N_0 = 7$ dB selected in [7] is not the best choice, but also demonstrates how noisy training can be more beneficial for better generalization.

Given all the aforementioned reasons, the fact that the autoencoder outperformed the Hamming code in [7] is more justified by the simplicity of the AWGN channel model which does not shift significantly the received signal in the testing E_b/N_0 interval. As a matter of fact, adding a noise correlation to the AWGN channel or accounting for the fading effects by changing the AWGN channel to a Rayleigh one break down the decoding performance of the autoencoder. In summary, the analysis of the results as a function of the trade-off between generalization and accuracy metrics opens the door to future rigorous investigations and reveals more insights about better data generation and model training choices.

B. Is Meta-learning Sufficient for Generalization?

To sidestep the need for large data samples to train DL models, meta-learning optimizes a general model using samples from multiple tasks (a.k.a., meta-tasks) in order to adapt to new unseen tasks. By designing meta-tasks associated with specific communication scenarios using the model-agnostic meta-learning (MAML) framework [8], prior work reported better performance for meta-learning solutions compared to standard DL methods. A few attempts bypass the black-box nature of DL models and connect communication systems models with meta-learning. This enables learning a subset of the model-based parameters, thereby minimizing the search space induced by black-box DL models [9]. However, the large body of research about generalization in wireless is limited to creating multiple meta-tasks with different communication conditions following the MAML-like framework. While enumerating and generating meta-tasks do not scale for some complex communication scenarios, understanding which feature is invariant across different domains becomes critical for scaling DL techniques. By doing so, one relies on the domain knowledge in addition to the standard two-step optimization of meta-learning. For instance, variations of complex signals related to the phase generalize better than those related to the amplitude. Overall, the study of the features in the context of meta-learning (a.k.a., meta-features) for wireless problems has not been explored by the wireless communications community. Existing studies rely on back-propagation to extract suitable correlations characterizing specific meta-tasks. This is different from those in the ML community which investigate data properties that affect the learning performance, and measure similarities between datasets and meta-features using multiple criteria such as mutual information and density skewness [10].

C. Unquestioned Sources of Dataset

In conventional communications protocols, a significant portion of a transmission interval (e.g., time slot) is usually

used to send training sequences for channel estimation. The real-time computation involved during these training time slots defines the computational complexity of many communication methods. For possible future AI-aided communication protocols, it is still unclear whether they will be designed based on fully offline training procedures or by additionally relying on extra finetuning steps. The latter scenario raises the question about real-time collection of data for adaptation purposes, and the related overhead to determine the complexity of DL techniques. The existing research in the open literature disregards these practical challenges and focuses entirely on fully offline-trained DL methods, which, at the current state, cannot fulfill the adaption capabilities of wireless systems envisioned for 6G.

When accuracy is the only metric of evaluation of DL models, ignoring these practical challenges seems acceptable because offline training is enough to judge the DL performance. However, when the assessment of DL techniques accounts for their generalization issues, the availability of data sources and their properties becomes a central component of the analysis.

D. Use of Reinforcement Learning for Optimization

In the pre-DL era (i.e., before 2012), optimizing non-convex problems using gradient descent (GD) algorithms were not very popular in the wireless research community. Instead of using GD-based optimization, researchers convexified the non-convex problems in order to solve them. DL research has introduced a plethora of optimizers to train DNNs which has made GD a popular technique. This is partly due to the unique properties of flat minima characterizing DNNs' loss functions [11]. Consequently, DL techniques have been used to solve optimization problems such as channel and power allocation, beamforming/precoding, user association, and trajectory planning for unmanned aerial vehicles (UAVs) in the physical/link layers in various contexts and system models.

A common practice to tackle non-convex communication problems in the DL era is by resorting to the reinforcement learning (RL) paradigm. RL agents are trained to learn the optimal policy to act on an environment in order to maximize the sum of the instantaneous reward signals received from an environment. We refer the reader to [12] for a rigorous treatment of RL formulation in terms of Markov decision processes (MDPs). By substituting the reward signal within the RL formalism with a convex or non-convex cost function to be optimized, the RL agent can find the best policy to optimize it. For instance, one can associate the beamforming vector to the action of the RL agent and the achievable rate of the communication system to be the reward signal [13].

Because wireless communications problems must satisfy multiple constraints simultaneously (e.g., power, latency, signal-to-interference-plus-noise ratio [SINR]), prior work made use of clipping strategies to enforce constraints on the RL agent's output. While this strategy provides good results in some scenarios, it does not guarantee an optimal solution. A better choice would be to cast the communication

problem within a constrained MDP formalism [14]. However, little effort has been dedicated to properly incorporating the constraints, and the unconstrained MDP approach remains a popular data-driven approach to optimize non-convex communication problems. In addition, the discussion of RL challenges in terms of sample efficiency and generalization is usually neglected despite the fact that it is an active research area within the ML research community.

E. Fixed Structures of Deep Neural Networks

DNNs have fixed structures and it is not possible to change them anymore after initialization. This inflexibility is considered a drawback because many wireless communications problems may require different input/output sizes over time. As one example, consider the problem of channel estimation at the base station based on the SINR vector with each component being associated with a specific user. In this case, the DNN input is the SINR vector while the estimated channel matrix represents the DNN output. During multiple transmission blocks, the number of users changes and so does the size of the SINR vector. Using the dropout method [15], it is possible to randomly drop the contribution of some neurons during the DNN training. While this provides flexibility in the structure of hidden layers, the input and output layers still have a fixed size, and hence limited flexibility in practice. Actively altering the network structure by adding neurons is both a promising and challenging direction. Because DL research for wireless communications generally considers DNNs as black-box modules, this field has not attracted the attention of the wireless research community and opted for constant-padding the input and output layers to the expected maximum size. Another unexplored area in this direction is the mapping of different vector sizes to the same latent space, thereby unifying the DNN input into the same feature space. While feature extractors for computer vision and natural language processing (NLP) tasks are abundant, wireless communications problems require signal-based feature extractors, which have not been well examined by our research community yet. It is worth noting that the recent wave of NLP models has allowed DNNs

to have an *apparent* flexibility in the input and output sizes since LLMs are trained to sequentially predict the next token given the current token. This is to be opposed to the application of DL for wireless where the output of DNNs is obtained with a one-shot inference run.

III. TWO KEY TRADE-OFFS FOR DL APPLIED TO WIRELESS PHYSICAL LAYER

In this section, we highlight the importance of two fundamental learning trade-offs in the assessment of DL models, namely, *accuracy versus generalization*, and *compression versus latency*. In particular, we discuss how accuracy-generalization should guide the evaluation of smaller DL models and highlight the importance of the compression-latency trade-off for practical use cases of LLMs. Through this section, generalization does not refer to the stability of the model's performance under noise and adversarial examples (i.e., adversarial robustness), but rather to the ability of the model to generalize to unseen scenarios.

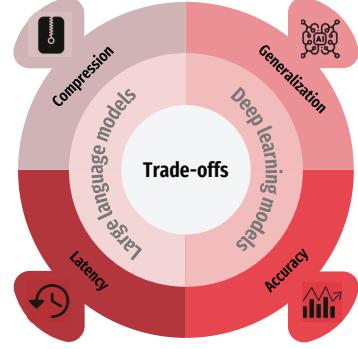


Fig. 5: The key trade-offs in assessing the relevance of small DL models and LLMs for communication use cases.

A. Accuracy Versus Generalization

In classical machine learning, the trade-off between the accuracy and the generalization pertains to the fundamental *bias-variance trade-off* which characterizes the generalization

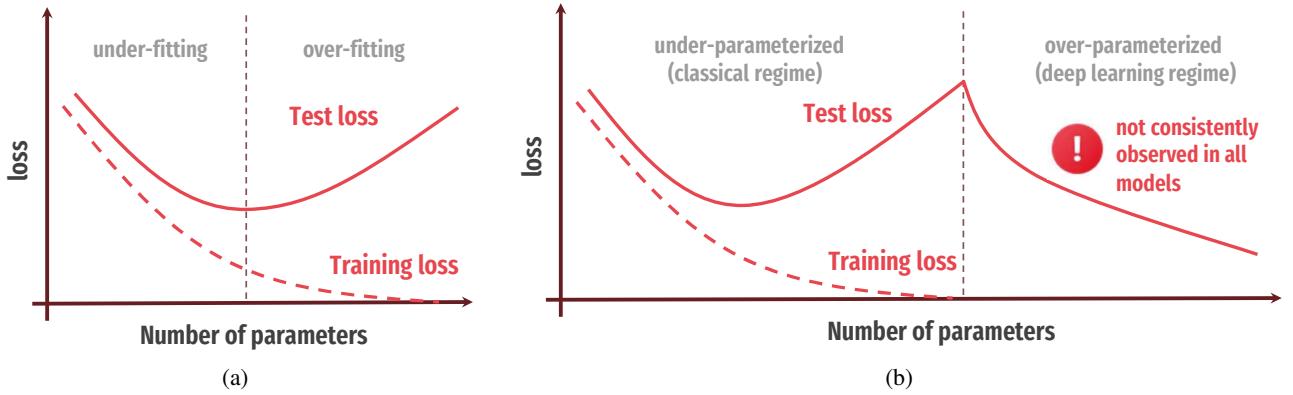


Fig. 4: Plots of training and test losses characterized by: (a) a U-shaped loss curve justified by the bias-variance trade-off, and (b) the double-descent loss curve combining the U-shaped loss curve for the under-parameterized regime and a decreasing loss for the over-parameterized regime. The latter is not a phenomenon exhibited by all deep learning models and hence is not well understood yet.

capabilities of predictive models [16]. The bias–variance trade-off stipulates that a model must have enough parameters to capture the underlying structure of the dataset without overfitting spurious patterns. Fig. 4(a) depicts the expected U-shaped variation of the test loss whose minimum represents the sweet spot in terms of the number of parameters between under-fitting and over-fitting. In DL practice, however, deeper networks with a large number of parameters are trained to interpolate between the training samples and do maintain a lower test loss accuracy on test datasets as depicted by the “double descent” curve in Fig. 4(b). While this over-parameterized regime depicted might look contradictory to classical (i.e., under-parameterized) regime, this over-parameterized regime is not well understood and is being actively investigated within the theoretical DL community [17]. This is because the double descent behavior does not consistently occur for every DNN and some of them, even very deep ones, still empirically obey the bias–variance trade-off [18]. Whether the DNN follows the variance-bias trade-off or the double descent behavior depends on the DNN’s parameters such as the number of input data points, the number of layers, and the overall number of parameters.

For the above reasons, evaluating the performance of DL models must be entirely tied to the employed DNN architecture. For example, one cannot claim that DL decoding is superior to classical methods because the mean-square error of *one specific* DNN tested on a few evaluation scenarios is lower than the one of a classical method (cf. Section II-A). Recent efforts from the communication community initiated the application of different learning paradigms including transfer learning, meta-learning, and continual learning to investigate the generalization of DNNs [19]. Some studies also showed that even the incorporation of domain knowledge in supervised learning without the explicit use of other learning paradigms can improve the DNN’s generalization. For instance, by alternating between the time-domain and frequency-domain representations of signals and using the idea of successive estimation and cancellation, it is possible to design DNNs to handle multiple sinusoid waves and improve their estimation performance on out-of-distribution samples [20].

In summary, by blinding applying DNNs without accounting for their accuracy-generalization trade-off, most research results become biased and less impactful in the long term. At the time of writing, potential leaks about the unknown GPT-4 architecture on Reddit [21] revealed that GPT-4 is not one giant monolithic lossy dataset compressor but rather an ensemble of eight 220B-parameter LLMs, each trained with different data/task distributions. This suggests that GPT-4 is a mixture of experts (unlike GPT-3.5 and GPT-3) and operates at a much smaller over-parameterized regime that challenges the myth of very large and fully end-to-end DNNs.

B. Compression Versus Latency

The study of the trade-off between lossless/lossy compression and latency has historically evolved around reaching the source coding limit established in 1948 by Shannon’s seminal work [22]. In a nutshell, it stipulates that higher compression ratios require longer block lengths and thus higher encoding and decoding time complexity. Specifically, Shannon showed that the codeword length of an optimal prefix-free code is approximately the negative logarithm of the codeword’s probability. He also proved that the expected message length of an optimal prefix-free code is close to the entropy of the message. Shannon also explored the information-theoretic relationship between compression and (next-letter) prediction by estimating the entropy of the English language [23]. Similar to the connection between next-letter predictors and data compressors, the emerging LLMs represent lossy data compressors in a data-driven fashion which exploit the inherent redundancy within the human language to embed massive datasets into a significantly smaller DNN model [24]. They are giant DNNs with billions of parameters compressing a large number of datasets by learning to predict the next token from the previous context composed of one or multiple tokens.

Given the tight information-theoretic relationship between the compression ratio and the decoding latency, it is therefore natural at this time to review some of the envisioned communication use cases of LLMs. The fascination of the general public with the quality of the output text of LLMs suggests that their compression quality will be at the cost of

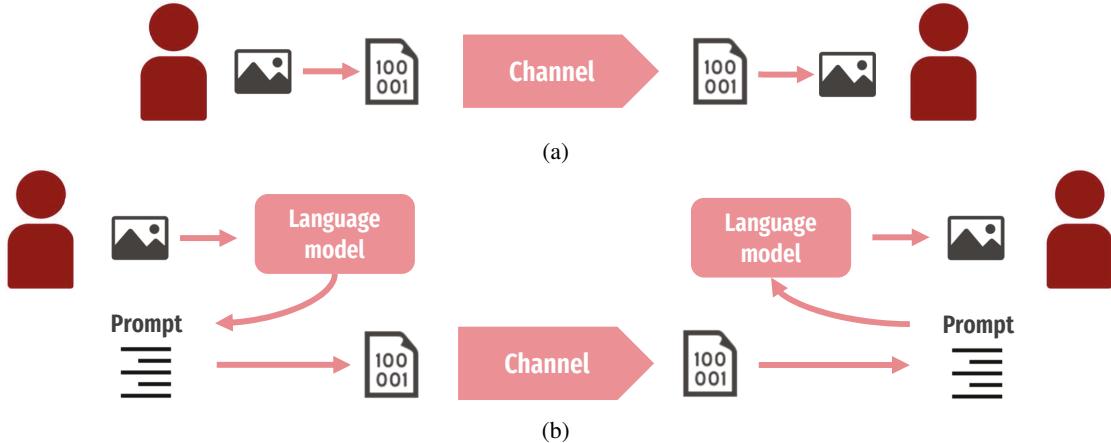


Fig. 6: Image transfer using (a) conventional communication, and (b) using a LLM as an image compressor to a prompt.

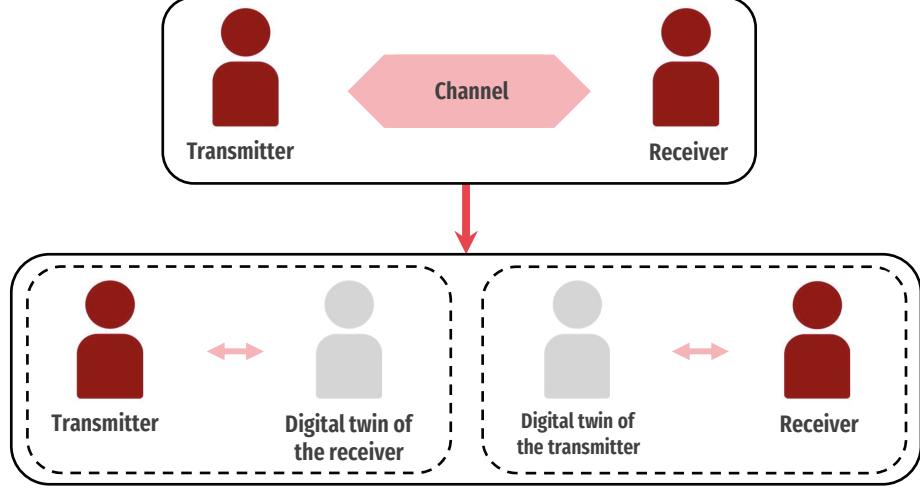


Fig. 7: The implementation of digital twins depends on finding the trade-off between latency and compression that satisfies the QoS.

their decoding latency. Indeed, an LLM with a few billion parameters requires around 100 GB of RAM. Moreover, the current optimized deployment of LLMs for efficient inference comes at the expense of high backend infrastructure costs and significant latency. Aware of these issues, important research efforts from the LLM community are actively examining the effect of quantization on the inference time of LLMs without significantly affecting their accuracies. For these reasons, the discussed applications of LLMs by the communications research community focus on latency-tolerant use cases at the application layer.

The first use case of LLMs for communication is image transfer. As shown in Fig. 6(b), a user equipped with an LLM obtains the text prompt of the image to transfer. Only the prompt is sent through the channel as a bit stream. The prompt is then recovered from the received bits to probe the LLM, which in turn outputs the image to the receiver. This is unlike the conventional data transfer where the entire bit stream of the image is transmitted through the channel as in Fig. 6(a). Here, the LLM-aided image transfer can provide a significant data compression rate under two conditions. The first one is that the parameters of the transmit and receive LLMs must be equal, or that their embeddings yield the encoded and decoded image. The second one is related to the compression-latency trade-off where both the encoding and decoding time of LLMs must provide a significant advantage over the improvement of the compression ratio between an image and a text prompt.

Another use case of LLMs is to minimize the need for communication between two users by learning a personalized LLM as a digital twin for each one of them as shown in Fig. 7. This means that an LLM must be able to reliably mimic the user interaction. From a probabilistic perspective, this corresponds to reliably learning the joint distribution of reactions, preferences, and thoughts of any user. The realizability of this scenario is still far from being a reality in the near future due to multiple reasons. This includes the data privacy and security concern, as well as the hallucination effects exhibited by LLMs as they tend to output text that appears to be

correct but is actually false or not based on the input given. In summary, short-cutting the cost of communication must come with the benefit of highly accurate digital twins. Another important limitation to implementing these use cases is the lack of rigorous metrics to assess the performance of LLMs. In fact, existing evaluation protocols are either automatic (e.g., the ROUGE metric for text summarization) exhibiting poor correlation to human judgments, or manual yielding noisy and potentially biased annotations. In summary, the adaptation of LLMs to the unique signal-based nature of communication datasets is vague and uncertain and is still in its infancy.

IV. CONCLUSION

The ever-increasing requirements for future 6G wireless applications call for the urgent need to go beyond the accuracy metric to assess DL methods for communication. We have described the importance of the accuracy-generalization and compression-latency trade-offs in shaping the future evaluation guidelines of DL techniques for wireless problems as summarized in Fig. 8.

1	Data sources	Discussion about the dataset collection process in a practical settings
2	Accuracy evaluation	Performance benchmark using the figure-of-merits
3	Generalization evaluation	Investigation of the model accuracy for different parameter values pertaining to different regimes of analysis
4	Latency evaluation	Time complexity benchmark for practical algorithmic solutions

Fig. 8: The proposed evaluation components.

These evaluation criteria should be continuously updated in light of a new understanding of the specific challenges facing the applications of deep learning models for wireless communication problems. We also have discussed how these

metrics are critical in evaluating the relevance of emerging deep learning models including large language models. We believe these trade-offs bridge the gap between the empirical nature of deep learning models applied to communication problems and the challenging technical requirements of future communication systems.

ACKNOWLEDGMENTS

The work of M. Akroud was supported by the doctoral scholarship of the Natural Sciences and Engineering Research Council of Canada (NSERC). The work of A. Mezghani, E. Hossain, and F. Bellili were supported by Discovery Grants from NSERC. The work of R. Heath was supported by the National Science Foundation (grant nos. NSF-ECCS-2153698, NSF-CCF-2225555, NSF-CNS-2147955) and in part by funds from federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program. The authors also acknowledge the insightful comments and suggestions from Osvaldo Simeone.

REFERENCES

- [1] 3GPP Release 18, “Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface RAN,” Meeting #112, Athens, Greece, Tech. Rep., 27th February – 3rd March 2023.
- [2] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [3] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [4] J. Tanveer, A. Haider, R. Ali, and A. Kim, “Machine learning for physical layer in 5g and beyond wireless networks: A survey,” *Electronics*, vol. 11, no. 1, p. 121, 2022.
- [5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [6] OpenAI, “Gpt-4 technical report,” 2023.
- [7] T. O’Shea and J. Hoydis, “An introduction to deep learning for the physical layer,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563–575, 2017.
- [8] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.
- [9] T. Raviv, S. Park, O. Simeone, Y. C. Eldar, and N. Shlezinger, “Online meta-learning for hybrid model-based deep receivers,” *IEEE Transactions on Wireless Communications*, 2023.
- [10] A. Rivolli, L. P. Garcia, C. Soares, J. Vanschoren, and A. C. de Carvalho, “Characterizing classification datasets: a study of meta-features for meta-learning,” *arXiv preprint arXiv:1808.10406*, 2018.
- [11] R. Mulayoff and T. Michaeli, “Unique properties of flat minima in deep networks,” in *International conference on machine learning*. PMLR, 2020, pp. 7108–7118.
- [12] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [13] Y. Al-Eryani, M. Akroud, and E. Hossain, “Multiple access in cell-free networks: Outage performance, dynamic clustering, and deep reinforcement learning-based design,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1028–1042, 2020.
- [14] E. Altman, *Constrained Markov decision processes*. CRC press, 1999, vol. 7.
- [15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [16] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.
- [17] M. Belkin, D. Hsu, S. Ma, and S. Mandal, “Reconciling modern machine-learning practice and the classical bias–variance trade-off,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.
- [18] S. Bubeck and M. Sellke, “A universal law of robustness via isoperimetry,” *Journal of the ACM*, vol. 70, no. 2, pp. 1–18, 2023.
- [19] M. Akroud, A. Feriani, F. Bellili, A. Mezghani, and E. Hossain, “Domain generalization in machine learning models for wireless communications: Concepts, state-of-the-art, and open issues,” *arXiv preprint arXiv:2303.08106*, 2023.
- [20] R. M. Dreifuerst and R. W. Heath, “Signalnet: A low resolution sinusoid decomposition and estimation network,” *IEEE Transactions on Signal Processing*, vol. 70, pp. 4454–4467, 2022.
- [21] “Potential GPT-4 architecture description,” https://www.reddit.com/r/LocaLLaMA/comments/14eoh4f/rumor_potential_gpt4_architecture_des/, accessed: 2023-06-25.
- [22] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [23] C. E. Shannon, “Prediction and entropy of printed english,” *Bell system technical journal*, vol. 30, no. 1, pp. 50–64, 1951.
- [24] Z. Jiang, M. Yang, M. Tsirlin, R. Tang, Y. Dai, and J. Lin, ““low-resource” text classification: A parameter-free classification method with compressors,” in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 6810–6828.