MCformer: A Transformer Based Deep Neural Network for Automatic Modulation Classification

Shahab Hamidi-Rad
InterDigital AI Lab
Los Altos, CA, USA
shahab.hamidi-rad@interdigital.com

Swayambhoo Jain
InterDigital AI Lab
Los Altos, CA, USA
swayambhoo.jain@gmail.com

***Abstract*—In this paper, we propose *MCformer* - a novel deep neural network for the automatic modulation classification task of complex-valued raw radio signals. MCformer architecture leverages convolution layer along with self-attention based encoder layers to efficiently exploit temporal correlation between the embeddings produced by convolution layer. MCformer provides state of the art classification accuracy at all signal-to-noise ratios in the RadioML2016.10b data-set with significantly less number of parameters which is critical for fast and energy-efficient operation.**

***Index Terms*—modulation recognition, cognitive radio, dynamic spectrum access, deep learning and attention mechanism.**

I. Introduction

Automatic modulation classification (AMC) pertains to the task of classifying the modulation type from the complex-valued raw radio signals with no a priori information about the signal or the channel parameters. AMC provides insight into spectrum usage and the type of transmitters present in the vicinity. Consecutively, it plays an important role in dynamic spectrum access, cognitive radios, and software defined radios [1]. Recently, there has been a series of works demonstrating the superior performance of Deep Neural Networks (DNNs) over classical approaches involving handcrafted features and likelihood-based methods. This performance of DNNs can be attributed to their ability to automatically learn feature representations for the AMC instead of hand-crafted features [2]–[6]. Most of these DNN architectures are obtained in the field of computer vision, natural language processing, speech recognition etc. However, the latency and energy efficiency requirement for successful application of DNNs on raw radio signals in wireless communication are much more stringent. Therefore, in the area of wireless communication AMC has become an important problem through which research community is trying to find an appropriate DNN architecture suited for raw radio signals.

Motivated by excellent performance of DNNs in AMC in this paper we propose the MCformer - a transformer-based DNN for AMC. MCformer uses a convolution layer to obtain a high dimensional embedding for each in-phase (I) and quadrature (Q) sample pair of complex-valued radio signal. Each component of the embedding is obtained by different learnable filter. This convolution layer transforms the complex-valued raw radio signals to sequence of embeddings which are subsequently fed to the transformer encoder layers to obtain embeddings used by the dense layers for final modulation type classification. The transformer encoder layer uses the self-attention mechanism which was first proposed in context of sequence of vector embeddings arising in natural language processing [7], [8]. Self attention mechanism allows efficient computation of long term correlation between the sequence embedding vectors. In past few years it has become one of the most efficient ways of modeling sequences. It significantly outperforms long short term memory (LSTM) based recurrent neural networks (RNNs) as they allow successful exploitation of correlations between distant sequence samples [8], [9]. In context of AMC, the temporal correlations have been explored in existing literature using LSTM based RNN [5] and convolution LSTM DNN (CLDNN) [6]. However, none of these architectures perform better at all SNR values. The proposed MCformer architecture builds upon the success of self-attention mechanism in effectively modeling long term correlations in sequences and provides significantly better performance at all SNRs. To the best of our knowledge MCformer is the first DNN that leverages self-attention mechanism for AMC.

Through extensive numerical evaluations we demonstrate that the proposed MCformer architecture provides state of the art performance as compared to existing techniques at all signal-to-noise ratios (SNRs) on the RadioML2016.10b dataset [10]. We also perform a detailed experimental study to understand contribution of various components of MCformer architecture on the performance. While application of self-attention is marred with significant computational complexity to our surprise the MCformer achieves superior performance with significantly fewer number of parameters. We study two variations of MCformer architecture: MCformerLarge and MCformerSmall. The MCformerLarge has $72,810$ parameters whereas MCformerSmall has just $10,050$ representing a memory efficient variation of the MCformer architecture. MCformerLarge significantly outperforms previous state of art ResNet architecture which has around $150,000$ parameters [6]. MCformerSmall performs similar to ResNet. This implies that MCformerLarge is slightly more than 2 times and MCformerSmall is around 15 times more parameter efficient than ResNet. Therefore, MCformer is a strong alternative that is fast and energy-efficient as compared with existing DNN architectures for complex-valued raw radio signals.

II. Outline

The rest of this paper is organized as follows. We begin by providing a background on the automatic modulation classification problem in Section III which is followed by a brief survey of existing works in Section IV. The MCformer architecture is described in Section V. Section VI provides details of the RadioML dataset on which performance of MCformer is evaluated. Detailed experimental evaluation is discussed in Section VII and Section VIII concludes this paper with summary of our findings and future research directions.

III. Automatic modulation classification Background

The automatic modulation classification is defined as the problem of classifying modulation type of wireless signals whose baseband complex envelope is given by

$$r(t) = s(t;\mathbf{z}_i) + n(t), \tag{1}$$

where the analytical expression for $s(t;\mathbf{z}_i)$ is given by

$$s(t;\mathbf{z}_i) = a_i e^{j2\pi\Delta f t} e^{j\theta} \sum_{k=1}^{K} e^{j\Phi_k} s_k^{(i)} g\left(t-(k-1)T-\epsilon\right), \tag{2}$$

where $0 \leq t \leq KT$ is the noise-free baseband complex envelope of the received signal, $n(t)$ is the instantaneous channel noise at time t, a_i is the unknown signal amplitude, Δf is the carrier frequency offset, θ is the time-invariant carrier phase, ϕ_k is the phase jitter, $\left\{s_k^{(i)}, 1 \leq k \leq K\right\}$ denotes K complex symbols, T represents the symbol period, ϵ is the normalized time offset between transmitter and signal receiver, $g(t) = P_{\text{pulse}}(t) \otimes h(t)$ is the composite effect of the residual channel with $h(t)$ denoting the channel impulse response and $\otimes$ denoting mathematical convolution, and $P_{\text{pulse}}(t)$ is the transmit pulse shape. Here, $\mathbf{z}_i = \left\{a_i, \Delta f, \theta, g(t), \{\phi_k\}_{k=1}^{K}, \left\{s_k^{(i)}\right\}_{k=1}^{K}\right\}$ is the high dimensional vector containing all deterministic unknown signal or channel parameters for the i^{th} modulation type. The automatic modulation classification problem is defined as the detection of modulation type i from the received signal $r(t)$.

IV. Related Works

Classical approaches to AMC involve various likelihood-based methods [11]–[16] and expert handcrafted feature design based machine learning methods [17]–[22]. These approaches require precise estimation of various signal parameters such as carrier frequency and signal power with manual calibration of threshold. Typically, these approaches classify only a small subset of modulation types.

Multilayer Perceptron (MLP) based on handcrafted features alleviates the problem with calibration of threshold and provides classification for a broader class of modulation types. More recently, significant improvement in AMC was shown using modern deep neural networks (DNNs) which alleviates the problem of handcrafted feature design and perform classification directly from raw signals [2], [3]. Another significant milestone was the release of RadioML datasets that allowed for consistent bench-marking of various AMC techniques and led to rapid improvement in DNN architectures for AMC. Starting with basic convolution neural network (CNNs) various architectures such as residual networks (ResNets) [3] and densely connected convolution network (DenseNets) [4] have been proposed in literature. Other DNNs which leverage the temporal aspect of the radio signals are long short-term memory (LSTM) based recurrent neural networks (RNN) [5] and convolution LSTM DNN (CLDNN) [6]. In a recent extensive experimental study, it was shown that CLDNN and ResNet performed best at lower SNRs whereas at high SNRs LSTM and ResNet perform best [6]. The LSTM based DNN demonstrated the advantage of modeling signals as a sequence of vectors but the superior performance is limited to high SNR values.

In past few years self-attention mechanism has emerged as significantly better way to model sequences especially in the field of natural language processing [8], [9]. The proposed MCformer architecture builds upon this and leverages self-attention mechanism in conjunction with convolution layers to provide significantly better performance at all SNRs. To the best of our knowledge MCformer is the first DNN to leverages self-attention mechanism for complex-valued raw radio signals.

V. MCformer: A transformer based deep neural network

We propose a transformer based DNN architecture MCformer for modulation classification as shown in Figure 1. The input to MCformer is such that I and Q components form a one-dimensional image of depth two. The input is fed to a convolution layer with kernel size $k \times 1$ with same-padding and n_c^o number of output channels. We use convolution layer because due to the lack of synchronization the captured I/Q samples start and end at random locations of the received signal. We want to be able to extract features in a location independent way and convolution layer is the best known method for this purpose. The output of convolution layer is fed to Scaled Exponential Linear Unit (SELU) non-linearity [23] as it has self-normalizing properties which is important for dealing with AMC at different SNR values. The output of SELU is then reshaped to an $n_c^o \times 128$ matrix. DNN till this point essentially converts the two-dimensional vector at each time sample to an n_o^c dimensional feature representation obtained using different kernels of the convolution layer.

Next, this matrix is passed through N self-attention based transformer encoder layers with hidden size n_c^o and h heads. Each transformer encoder layer takes a $128 \times n_c^o$ matrix and outputs a matrix with the same shape by using self-attention based processing. The high level block diagram for the transformer encoder layer is shown in Figure 2. This layer was first proposed in [8]. The resulting $128 \times n_c^o$ matrix obtained after N transformer encoder layers is down-sampled

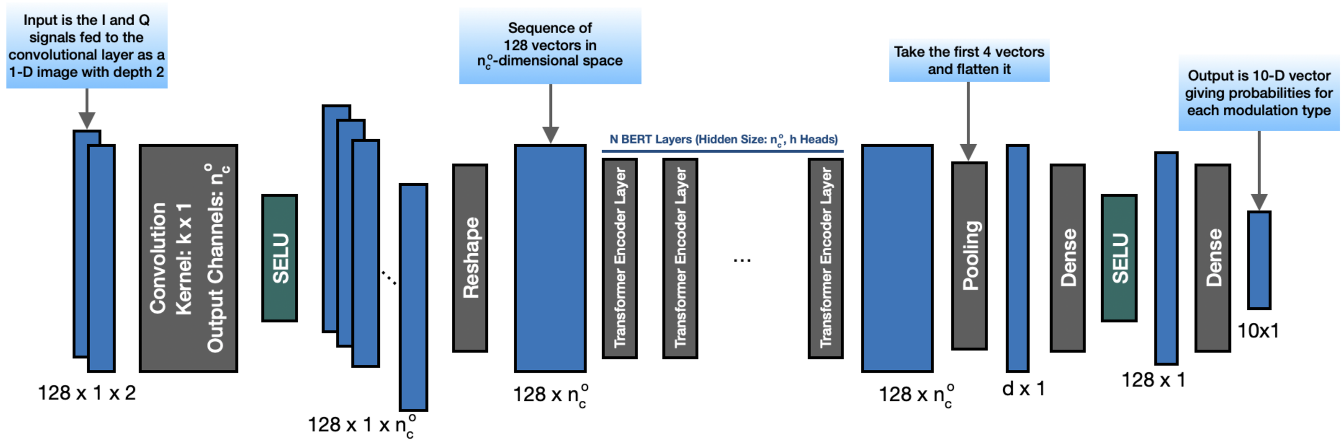

Input is the I and Q signals fed to the convolutional layer as a 1-D image with depth 2
128 x 1 x 2
Convolution
Kernel: k x 1
Output Channels: n_c^o
SELU
128 x 1 x n_c^o
Reshape
Sequence of 128 vectors in n_c^o-dimensional space
128 x n_c^o
N BERT Layers (Hidden Size: n_c^o, h Heads)
Transformer Encoder Layer
Transformer Encoder Layer
...
Transformer Encoder Layer
128 x n_c^o
Take the first 4 vectors and flatten it
Pooling
d x 1
Dense
SELU
128 x 1
Dense
Output is 10-D vector giving probabilities for each modulation type
10x1

Fig. 1. The MCformer - model architecture.

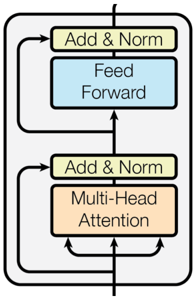


Fig. 2. Transformer encoder layer [8].

by the pooling layer which selects the first 4 rows of this matrix and reshapes it into a $d = 4n_c^o$ dimensional vector. This d-dimensional embedding vector is fed to a dense layer followed by another dense layer with softmax non-linearity which outputs a 10-dimensional vector whose components provide probabilities of the modulation types.

A. Computational complexity

The total number of parameters is proportional to the computational complexity of the MCformer. The convolution layer has $2kn_c^o + n_c^o$ parameters. Whereas each transformer encoder layer has $n_c^o(12n_c^o + 13)$ parameters. The dependence on number of heads h does not show up as it gets canceled due the way number of heads and internal dimension in multi-head attention are choosen [8]. The first fully connected layer has $128d + 128$ parameters and last layer has 1290 parameters. The total number of parameter in MCformer has quadratic dependence on n_c^o.

B. Comparison with self-attention used in natural language processing

As the proposed MCformer DNN is primarily motivated by the success of self-attention in natural language processing here we compare and contrast its usage in AMC. There are some similarities between natural language processing tasks and AMC task. For example the Stanford Sentiment Treebank (SST) task from the GLUE [24] benchmark involves classification of sentences to "Positive" and "Negative". The input sentences are a sequence of words that are fed to the model as embeddings. This is similar to the AMC task where we convert sequences of complex samples to embeddings and feed them to the model to classify the type of modulation. Also, a temporal correlation between different elements of the sequences exists in both cases.

Unlike the NLP models where the sequences (i.e. sentences) have a deterministic start and end, the input samples for AMC tasks are usually taken at random times. This means the embedding process for AMC must be able to extract the features in a location-independent manner. Another difference is the type of data in the sequences, for NLP tasks, there is often a finite number of possible sequence elements (i.e. words) in a vocabulary. But for AMC, each item in a sequence is a complex number. This difference is very important and it affects the way we embed the sequence elements into a high dimensional space in which transformer models operate. Typically, in NLP there is a need of additional positional encoding to encode positions of embeddings in a sequence however for the AMC task we don't want this as there is a natural order in samples of complex-valued radio signal. Our initial experiments showed that we get better performance without these positional encodings.

VI. RadioML Dataset

The task of automatic modulation classification is typically evaluated on the RadioML2016.10b dataset [10]. This dataset consists of 128 dimensional complex vectors obtained by sampling wireless baseband signals of ten modulation types [eight digital and two analog modulation] at SNR values uniformly distributed from -20 dB to $+18$ dB, with a step size of 2 dB, i.e, $\{-20 \text{ dB}, -18 \text{ dB}, -16 \text{ dB}, \cdots, +16 \text{ dB}, +18 \text{ dB}\}$ The digital modulation types consist of BPSK, QPSK, 8PSK, QAM16, QAM64, BFSK, CPFSK, and PAM4; and analog modulations consist of WBFM and AM-DSB. The dataset is

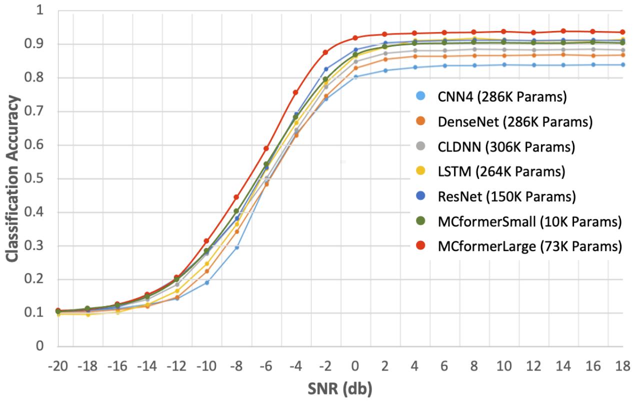


Fig. 3. SNR vs. classification accuracy plots for various DNNs. The MCformerLarge provides significantly better performance with only $72,810$ parameters compared to previous state of the art ResNet which has around $150,000$ parameters at all SNR values. MCformerSmall performs similar to ResNet but with just $10,050$ parameters.

labeled with SNR values and modulation types. The dataset is generated by simulating various channel imperfections such as thermal noise, multi-path fading, and hardware related noises such as sample rate and frequency offset etc. Detailed information about the dataset generation can be found in [25]. The final dataset consists of a total of $1,200,000$ samples such that for each SNR value there are $60,000$ samples with $6,000$ samples from each of ten modulation types.

VII. Experiments

We used RadioML2016.10b dataset for measuring performance of the proposed model. The dataset was randomly split into 45% for training, 5% for validation and 50% for testing following the experimental setup proposed in [6]. We evaluate our model against CNN, DenseNet, CLDNN, LSTM, and ResNet architecture considered in [6]. The MCformer was trained with batch size of 128 for 100 epochs with exponentially decaying learning rate from 0.002 to 0.0002. For all experiments, we used TensorFlow machine learning platform running on a Linux machine with 24 core Intel (R) Core(TM) i9-7920X @ 2.9 GHz CPU, 128 Gigabyte RAM, and 2 Titan-XP NVIDIA GPU cards.

Figure 3 shows the classification accuracy of trained MCformer on test data at different SNR values. The result for baselines are taken from [6]. We see that for all DNN architectures the classification accuracy increases with increasing SNR. We consider two variations of the proposed MCformer architecture: MCformerLarge and MCformerSmall. The MCformerLarge uses kernel size 65×1 with $n_o^c = 32$, $N = 4$ self-attention based transformer layers with $h = 4$ heads and hidden size 32, followed by two dense layers of size 128×128 and 128×10. The MCformerSmall has kernel size 65×1, $n_o^c = 8$, $N = 4$ self-attention based transformer layers with $h = 4$ heads and hidden size 8, followed by two dense layers of sizes 32×128 and 128×10.

We observe that MCformerLarge outperforms the baselines at all SNR values. MCformerSmall provides competitive performance with respect to existing baselines. While MCformerLarge outperforms MCformerSmall the performance gain comes with increased computational complexity as MCformerLarge has $72,810$ parameters as compared to $10,050$ parameters in MCformerSmall. However, as compared to baselines the number of parameters are significantly less in both architectures. This implies MCformer is a better DNN architecture for AMC task.

In order to understand the performance on individual modulation types we show confusion matrix for MCformerLarge in the Figure 4. We observe that analog modulation types WBFM and AM-DSB are particularly difficult to identify. The confusion matrix suggests while the transformer is able to identify the analog modulation from digital modulation but it is not able to identify exact type of analog modulation as the majority of times WBFM is classified as AM-DSB. In existing literature these modulation types are known to be challenging as they are obtained by sending acoustic voice speech with some periods of silence during which the modulation classification is difficult.

In order to better understand the impact of various parameters on the MCformer architecture we perform experiments wherein we vary kernel size k of the convolution layer, the hidden dimension size n_c^o, and number of transformer encoder layers N. In all the experiments we use the MCformerLarge and vary the desired parameter.

Figure 5 shows the classification accuracy at various SNR values for the kernel sizes $k = 9, 33, 65$. We observe that while the performance is slightly better at SNR values in range -10 dB to 2 dB for larger kernel size, the performance gain is not significant.

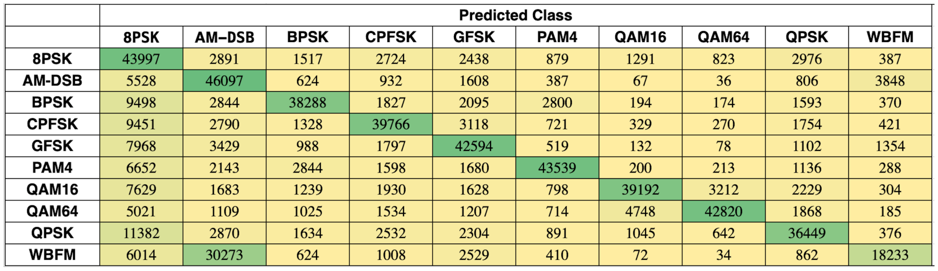
In Figure 6 we show the performance for the hidden sizes $n_c^o = 8, 16, 32$. We observe performance gain is significant for SNR values greater than -10dB as we increase the hidden size from 8 to 16. However, relative gain when we increase the hidden size from 16 to 32 is marginal. As discussed earlier that the number of parameters in MCformer increase in quadractic fashion with hidden size this result highlights the trade-off between performance and computational complexity.

Figure 7 shows the performance with varying number of transformer based encoder layers N. We observe that similar to hidden size while the performance increases significantly as we increase number of layers from 1 to 2, performance gain is marginal as we go from 2 to 4 layers.

VIII. Conclusion and Future works

We proposed a novel transformer based DNN - MCformer for AMC on complex-valued radio signals. MCformer leverages convolution layers and self-attention mechanism that allows for state of the art performance with significantly less number of parameters. We studied two MCformer based DNNs - MCformerSmall and MCformerLarge. Through numerical evaluation on RadioML2016.10b dataset we demonstrated superior performance of MCformer based architectures.

While this paper provides encouraging results for MCformer architecture there are many possible future research directions.



	Predicted Class									
	8PSK	**AM–DSB**	**BPSK**	**CPFSK**	**GFSK**	**PAM4**	**QAM16**	**QAM64**	**QPSK**	**WBFM**
8PSK	43997	2891	1517	2724	2438	879	1291	823	2976	387
AM-DSB	5528	46097	624	932	1608	387	67	36	806	3848
BPSK	9498	2844	38288	1827	2095	2800	194	174	1593	370
CPFSK	9451	2790	1328	39766	3118	721	329	270	1754	421
GFSK	7968	3429	988	1797	42594	519	132	78	1102	1354
PAM4	6652	2143	2844	1598	1680	43539	200	213	1136	288
QAM16	7629	1683	1239	1930	1628	798	39192	3212	2229	304
QAM64	5021	1109	1025	1534	1207	714	4748	42820	1868	185
QPSK	11382	2870	1634	2532	2304	891	1045	642	36449	376
WBFM	6014	30273	624	1008	2529	410	72	34	862	18233

Fig. 4. Confusion matrix for MCformerLarge. The $(i, j)^{\mathrm{th}}$ entry denotes number of test examples of modulation type i that are classified as modulation type j.

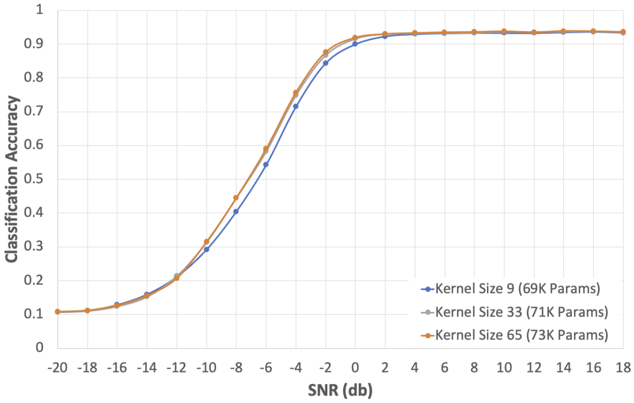


Fig. 5. SNR vs. classification accuracy plot for MCformerLarge architecture with different kernel sizes k.

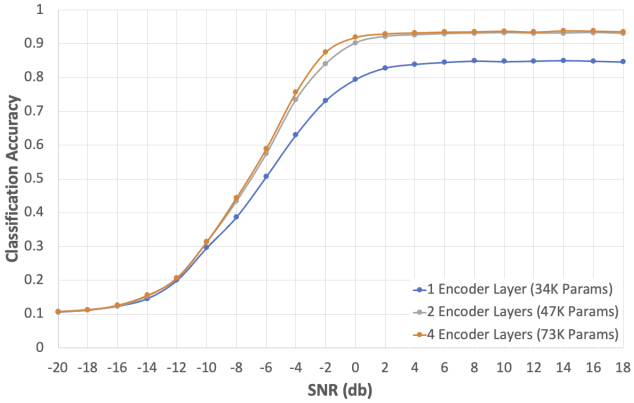


Fig. 7. SNR vs. classification accuracy plot for MCformerLarge architecture with different number of transformer based encoder layers.

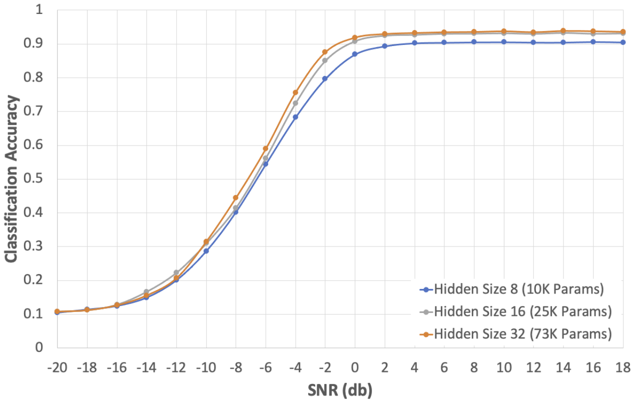


Fig. 6. SNR vs. classification accuracy plot for MCformerLarge architecture with different hidden sizes.

A study on the robustness of MCformer architecture to adversarial attacks is critical to real life applications. Another direction could be oriented towards improving the computational efficiency using DNN model compression techniques or by down-sampling the input. It remains to be seen if MCformer type architecture could be leveraged for tasks other than AMC.

References

[1] Z. Zhu and A. K. Nandi, *Automatic modulation classification: principles, algorithms and applications*. John Wiley & Sons, 2015.

[2] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Springer International Conference on Engineering Applications of Neural Networks*, 2016, pp. 213–226.

[3] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 168–179, 2018.

[4] J. Krzyston, R. Bhattacharjea, and A. Stark, "High-capacity complex convolutional neural networks for i/q modulation classification," *arXiv preprint arXiv:2010.10717*, 2020.

[5] G. Tao, Y. Zhong, Y. Zhang, Z. Zhang *et al.*, "Sequential convolutional recurrent neural networks for fast automatic modulation classification," *arXiv preprint arXiv:1909.03050*, 2019.

[6] S. Ramjee, S. Ju, D. Yang, X. Liu, A. E. Gamal, and Y. C. Eldar, "Fast deep learning for automatic modulation classification," *arXiv preprint arXiv:1901.05850*, 2019.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[9] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," *arXiv preprint arXiv:1702.00887*, 2017.

[10] "Radioml2016.10b," https://www.deepsig.ai/datasets.

[11] A. Polydoros and K. Kim, "On the detection and classification of quadrature digital modulations in broad-band noise," *IEEE Transactions on Communications*, vol. 38, no. 8, pp. 1199–1211, 1990.

[12] B. F. Beidas and C. L. Weber, "Modulation classification of mfsk signals using the higher-order correlation domain," in *Proceedings of IEEE Military Communications Conference*, vol. 1, 1995, pp. 186–191.

[13] P. Sapiano and J. Martin, "Maximum likelihood psk classifier," in *Proceedings of IEEE Military Communications Conference*, vol. 3. IEEE, 1996, pp. 1010–1014.

[14] J. A. Sills, "Maximum-likelihood modulation classification for psk/qam,"

in *Proceedings of IEEE Military Communications*, vol. 1, 1999, pp. 217–220.

[15] P. Panagiotou, A. Anastasopoulos, and A. Polydoros, "Likelihood ratio tests for modulation classification," in *Proceedings of IEEE Military Communications*, vol. 2, 2000, pp. 670–674.

[16] L. Hong and K. Ho, "Antenna array likelihood modulation classifier for bpsk and qpsk signals," in *Proceedings of IEEE Military Communications*, vol. 1, 2002, pp. 647–651.

[17] S. S. Soliman and S.-Z. Hsue, "Signal classification using statistical moments," *IEEE Transactions on Communications*, vol. 40, no. 5, pp. 908–916, 1992.

[18] L. Mingquan, X. Xianci, and L. Leming, "Cyclic spectral features based modulation recognition," in *Proceedings of IEEE International Conference on Communication Technology*, vol. 2, 1996, pp. 792–795.

[19] E. E. Azzouz and A. K. Nandi, "Modulation recognition using artificial neural networks," in *Automatic Modulation Recognition of Communication Signals*. Springer, 1996, pp. 132–176.

[20] L. Hong and K. Ho, "Identification of digital modulation types using the wavelet transform," in *IEEE Military Communications Conference Proceedings*, vol. 1, 1999, pp. 427–431.

[21] A. Swami and B. M. Sadler, "Hierarchical digital modulation classification using cumulants," *IEEE Transactions on Communications*, vol. 48, no. 3, pp. 416–429, 2000.

[22] G. Hatzichristos and M. P. Fargues, "A hierarchical approach to the classification of digital modulation types in multipath environments," in *IEEE Conference Record of Thirty-Fifth Asilomar Conference on Signals, Systems and Computer*, vol. 2, 2001, pp. 1494–1498.

[23] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *arXiv preprint arXiv:1706.02515*, 2017.

[24] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[25] T. J. O'shea and N. West, "Radio machine learning dataset generation with gnu radio," in *Proceedings of the GNU Radio Conference*, vol. 1, no. 1, 2016.