

Learning of Time-Frequency Attention Mechanism for Automatic Modulation Recognition

Shangao Lin[✉], *Student Member, IEEE*, Yuan Zeng[✉], *Member, IEEE*, and Yi Gong[✉], *Senior Member, IEEE*

Abstract—Recently, deep learning-based image classification and speech recognition approaches have made extensive use of attention mechanisms to achieve state-of-the-art recognition, which demonstrates the effectiveness of attention mechanisms. Motivated by the fact that the frequency and time information of modulated radio signals are crucial for modulation recognition, this paper proposes a time-frequency attention mechanism for convolutional neural network (CNN)-based automatic modulation recognition. The proposed time-frequency attention mechanism is designed to learn which channel, frequency and time information is more meaningful in CNN for modulation recognition. We analyze the effectiveness of the proposed attention mechanism and evaluate the performance of the proposed models. Experiment results show that the proposed methods outperform existing learning-based methods and attention mechanisms.

Index Terms—Automatic modulation recognition, convolutional neural network, time-frequency attention.

I. INTRODUCTION

AUTOMATIC modulation recognition (AMR) is the task of classifying the modulation mode of radio signals received from wireless communication systems. It is an intermediate step between signal detection and signal demodulation. As a step towards understanding what type of communication scheme and emitter is present, AMR has been widely used in practical civilian and military applications, such as cognitive radio, spectrum monitoring, communications interference, and electronic countermeasures.

In the past few years, due to the great success in computer vision and natural language processing, data-driven deep learning methods have also been applied to AMR, showing great potential in improving recognition accuracy and robustness. O’Shea *et al.* generated an open modulation recognition dataset, called RadioML2016.10a, using GNU

Radio in [1], and first proposed a deep neural network (DNN) architecture for AMR in [2]. Later, various DNN architectures were introduced to improve the recognition accuracy, such as convolutional neural networks [3], recurrent neural networks [4], and graph convolutional network [5]. Recently, a few deep learning-based approaches considered the inherent properties of radio signals and communication systems in modulation recognition. Yashashwi *et al.* [6] proposed a learnable distortion correction module to shift the frequency and phase of signal according to its weights and jointly train with a CNN. In [7], high-order statistics of radio signal was computed as an additional signal representation to the CNN classifier. Zeng *et al.* [8] exploited the time-frequency analysis of modulated radio signals and proposed a CNN with the short-time discrete Fourier transform (STFT). Wang *et al.* [9] proposed a multi-cue fusion network by modelling spatial-temporal correlations from modulated signal cues. Our work further leverages time-frequency characteristics of time series during the design of attention mechanism for modulation recognition.

In addition to DNNs, attention mechanisms have also been used in a wide variety of DNN-based methods in computer vision. A neural attention module can optimize the weights of the input features by minimizing recognition errors. This can hence enhance the important information and reduce the interference caused by irrelevant information in learning-based recognition frameworks. In [10], a squeeze-and-excitation (SE) attention was proposed. It computes channel attention with the help of 2D global pooling and provides notable performance gains at a considerably low computational cost. In [11], Woo *et al.* proposed a convolutional block attention module, which sequentially implements channel attention and spatial attention to enhance important parts of the input features. In contrast, our attention mechanism attends features in channel, frequency and time dimensions for improving the modulation recognition performance of existing CNN-based models.

In this work, we propose a time-frequency attention (TFA) mechanism to learn useful features from spectrogram images in terms of channel, frequency and time, and improve the recognition performance of existing methods. The channel attention is performed first to learn weights regarding channel importance in the input feature map, and then frequency and time attention mechanisms are performed in parallel and composited using learned weights for capturing both frequency and time attention. In addition, we integrate the proposed TFA mechanism into two CNN-based AMR models to improve the performance of AMR. Moreover, we conduct ablation experiments to analyze the effectiveness of the proposed TFA mechanism, and compare the presented AMR models with

Manuscript received Month Day, Year; revised Month Day, Year; accepted Month Day, Year. Date of publication Month Day, Year; date of current version Month Day, Year. This work is supported by National Key R&D Program of China under Grant 2019YFB1802800, National Natural Science Foundation in China under Grants 62071212 and 62106095, Guangdong Science and Technology Program under Grant 2019A1515110479, Guangdong Basic and Applied Basic Research Foundation under Grant 2019B1515130003, Shenzhen Science and Technology Program under Grant JCYJ202001091414409, and Educational Commission of Guangdong Province of China under Grants 2020ZDZX3057, 2019KQNCX128 and 2017KZDXM075. The editor coordinating the review of this article and approving it for publication was C.-K. Wen. (Corresponding authors: Yuan Zeng and Yi Gong.)

S. Lin and Y. Gong are with the Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: gongy@sustech.edu.cn).

Y. Zeng is with the Academy for Advanced Interdisciplinary Studies, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: zengy3@sustech.edu.cn).

Digital Object Identifier 10.1109/LWC.2021.XXXXXXX

three existing state-of-the-art (SOTA) AMR methods in [2], [12], [13].

II. PROBLEM STATEMENT AND SPECTRUM REPRESENTATION

This paper considers a simple single-input single-output wireless communication system, where a symbol is converted and transmitted to a receiver via a communication channel. The data model of received signal $r(t)$ is given as:

$$r(t) = \mathcal{F}(s(t)) * h(t) + n(t), \quad (1)$$

where $s(t)$ denotes the transmission symbol, \mathcal{F} is a modulation function, $h(t)$ is the communication channel impulse response, and $n(t)$ is the additive white Gaussian noise. Given the received signal $r(t)$, AMR aims at decoding the modulation function \mathcal{F} . A discrete-time version of the continuous-time signal $r(t)$ can be obtained by sampling $r(t)$ for n times with a sampling rate $f_s = \frac{1}{T_s}$ i.e. $r(n) = r(t)|_{t=nT_s}$, $-\infty < n < +\infty$.

Since the time-frequency analysis of a modulated signal can reflect its frequency varies with time, which is an important distinction among different modulated signals. In this work, we exploit the insight from recent work [8] that spectrograms can achieve richer time-frequency representation of signals, and use STFT-based spectrogram to represent the signal about frequency variation trend with time. The continuous signal $r(t)$ is first converted to discrete-time signal $r(n)$ with sampling frequency f_s , and then $r(n)$ is windowed and transformed into the frequency domain by applying the STFT, that is:

$$R(m, k) = \sum_{n=mK}^{mK+(L-1)} r(n)w(n-mK)e^{-j\frac{2\pi k}{L}n}, \quad (2)$$

where m and k denote the time frame and frequency bin indices, respectively. $w(n)$ denotes the window function, L is the frame length, and K is the frame shift. The spectrogram x is given as $x = |R(m, k)|^2$, where each pixel corresponds to a point in frequency and time.

III. AUTOMATIC MODULATION RECOGNITION

A. Time-Frequency Attention

We propose a TFA mechanism for extracting meaningful channel, frequency, and time information of the spectrogram inputs for AMR. The proposed TFA aims at devoting more computing power to that small but important part of the data. The overview of the proposed TFA is shown in Fig. 1. The feature map generated from a convolutional layer is the input feature map of TFA, later the refined feature map generated by TFA is the input of next layer. The TFA contains three sub-modules, namely channel attention module (CAM), frequency attention module (FAM), and time attention module (TAM). The CAM is used to exploit the inter-channel relationship of features, and extract general information regarding channel importance in the input feature map. The FAM focuses on where is the important frequency information of the channel attention refined feature map, and TAM focuses on where is the important time information of the channel attention refined feature map.

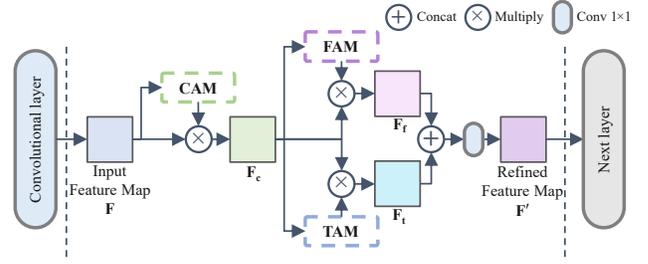


Fig. 1. Overview of the proposed TFA mechanism integrated in a convolutional layer.

Given an input feature map $\mathbf{F} \in R^{H \times W \times C}$ from previous convolutional layer, where H and W denote the height and width, respectively, and C denotes the number of channels. The TFA first uses CAM \mathbf{M}_c to generate a channel refined feature map $\mathbf{F}_c \in R^{H \times W \times C}$. Later, FAM \mathbf{M}_f and TAM \mathbf{M}_t are performed on \mathbf{F}_c in parallel. The parallel attention operations are performed to generate a frequency refined feature map $\mathbf{F}_f \in R^{H \times W \times C}$ and a time refined feature map $\mathbf{F}_t \in R^{H \times W \times C}$, respectively. After concatenation of the two refined feature maps \mathbf{F}_f and \mathbf{F}_t , a convolution layer with 1×1 -sized kernel is applied to generate the final refined feature map $\mathbf{F}' \in R^{H \times W \times C}$. Then \mathbf{F}' is treated as the input of the next layer. The overall process can be summarized as:

$$\mathbf{F}_c = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F}, \quad \mathbf{F}_f = \mathbf{M}_f(\mathbf{F}_c) \otimes \mathbf{F}_c, \quad \mathbf{F}_t = \mathbf{M}_t(\mathbf{F}_c) \otimes \mathbf{F}_c, \quad (3)$$

$$\mathbf{F}' = f^{1 \times 1}([\mathbf{F}_f; \mathbf{F}_t]), \quad (4)$$

where \otimes denotes element-wise multiplication. Multiplication process enhances the important parts of input data and fade out the rest according to the learned attention operations \mathbf{M}_c , \mathbf{M}_f and \mathbf{M}_t . $f^{1 \times 1}$ denotes a convolutional layer with 1×1 -sized kernel.

Fig. 2 shows the procedures of CAM, FAM, and TAM. The CAM first performs global max-pooling and global average-pooling on the input feature map to generate features that denote two different contexts respectively. The features are then used as the input of a shared network, which consists of a multi-layer perceptron (MLP) comprising two densely connected layers with $C/8$ and C neurons. The MLP is trained with the network with same training settings. The outputs of the shared network are element-wise added up. Then a sigmoid function is performed to generate the channel attention map. The operation \mathbf{M}_c is given as:

$$\mathbf{M}_c(\mathbf{F}) = \sigma(MLP(AvgPool(\mathbf{F})) + MLP(MaxPool(\mathbf{F}))), \quad (5)$$

where σ denotes the sigmoid function. The output channel attention map $\mathbf{M}_c(\mathbf{F}) \in R^{1 \times 1 \times C}$ is multiplied by \mathbf{F} to generate a channel attention refined feature map \mathbf{F}_c .

The FAM and TAM have similar procedures. The TAM focuses on the x-axis of input spectrogram which represent the time axes, and FAM focuses on y-axis which represent the frequency axes. The channel refined feature map \mathbf{F}_c is the input of FAM and TAM. The FAM averages \mathbf{F}_c along the time axes to focus on the frequency feature $\mathbf{F}_f \in R^{H \times 1 \times C}$, which is given by $\mathbf{F}_f = AvgPool_{1 \times W}(\mathbf{F}_c)$. Then, max-pooling

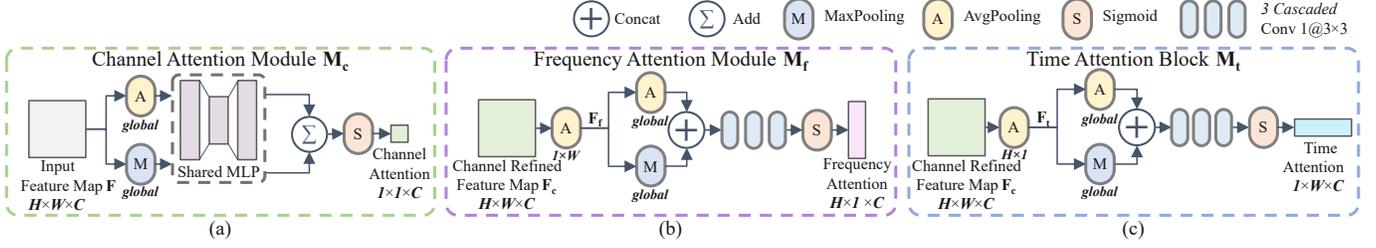


Fig. 2. Flowchart of the three submodules in TFA. (a) Channel attention module. (b) Frequency attention module. (c) Time attention module.

and average-pooling are performed and then concatenated to further extract the frequency feature. After that, three cascaded 3×3 convolutional layers and a sigmoid function are performed to generate a frequency attention map $\mathbf{M}_f(\mathbf{F}_c) \in R^{H \times 1 \times C}$. Unlike FAM, TAM averages \mathbf{F}_c along the frequency axes to focus on the time feature $\mathbf{F}_t \in R^{1 \times W \times C}$, which is given by $\mathbf{F}_t = \text{AvgPool}_{H \times 1}(\mathbf{F}_c)$, and performs the same operations as in FAM to generate a time attention map $\mathbf{M}_t(\mathbf{F}_c) \in R^{1 \times W \times C}$. The operations \mathbf{M}_f and \mathbf{M}_t are given as:

$$\begin{aligned} \mathbf{M}_f(\mathbf{F}_c) &= \sigma(f_3^{3 \times 3}([\text{AvgPool}(\mathbf{F}_f); \text{MaxPool}(\mathbf{F}_f)])), \\ \mathbf{M}_t(\mathbf{F}_c) &= \sigma(f_3^{3 \times 3}([\text{AvgPool}(\mathbf{F}_t); \text{MaxPool}(\mathbf{F}_t)])), \end{aligned} \quad (6)$$

where $f_3^{3 \times 3}$ denotes 3 cascaded convolutional layers with 3×3 -sized kernel. The frequency attention map $\mathbf{M}_f(\mathbf{F}_c)$ and time attention map $\mathbf{M}_t(\mathbf{F}_c)$ are multiplied by the input feature map \mathbf{F}_c to generate a frequency refined feature map \mathbf{F}_f and a time refined feature map \mathbf{F}_t , respectively. After concatenating \mathbf{F}_f and \mathbf{F}_t , a convolutional layer $f^{1 \times 1}$ is used to generate the final refined feature map \mathbf{F}' as shown in Fig. 1.

B. Network Architecture

We follow a CNN architecture similar to the one used in [8], called spectrum CNN (SCNN), and investigate the designed TFA block into the CNN architecture, called TFA-SCNN. Fig. 3 illustrates the overview of the framework TFA-SCNN. It consists of one input layer, 4 convolutional layers integrated with TFA, one densely connected layer, and an output softmax layer. The input of the network is a spectrogram image with the dimension of $100 \times 100 \times 3$. The convolutional layers use 3×3 -sized kernel and the number of kernels of the 4 convolutional layers is setting as 64, 32, 12, 8. The feature maps from convolutional layers integrated with TFA are followed by rectified linear unit (ReLU) [14] and a max-pooling layer with a size of 2×2 , except for the last one only followed by ReLU. Specifically, the dimension of the feature maps generated by TFA is $98 \times 98 \times 64$, $47 \times 47 \times 32$, $21 \times 21 \times 12$, $8 \times 8 \times 8$. The densely connected layer consists of 128 neurons. The output of the network is the predicted modulation mode of input. The network is trained using stochastic gradient descent to minimize the cross-entropy loss function, that is, $\mathbf{w}^* = \text{argmin}_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(\mathbf{w}; x^i, t^i)$, with the number of training examples N , the true labels t^i , and the predicted labels x^i . \mathcal{L} denotes the loss function, that is, $\mathcal{L} = -\sum_j^M \beta_j \log(q_j)$, where M denotes the number of classes, β_j is a binary indicator with $\beta_j = 1$ if x^i is t^i , otherwise $\beta_j = 0$, and q_j denotes the predicted probability of belonging to class j .

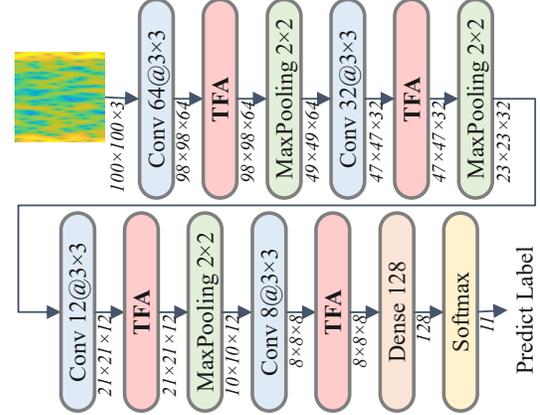


Fig. 3. Network architecture of the proposed TFA-SCNN.

IV. EXPERIMENTS

A. Dataset

We evaluate the proposed AMR framework on an open-source dataset RadioML2016.10a [1] and its larger version RadioML2016.10b. RadioML2016.10a consists of analog and digital modulation methods, including 11 commonly used modulations modes in communication systems, which are 8PSK, AM-DSB, AM-SSB, BPSK, CPFSK, GFSK, PAM4, QAM16, QAM64, QPSK, and WBFM. RadioML2016.10a includes 220000 modulated signals with 20 different signal-to-noise ratios (SNRs) ranging from -20 dB to 18 dB, and 1000 signals per SNR per modulation mode. RadioML2016.10b consists of 1200000 modulated signals with 20 different SNRs ranging from -20 dB to 18 dB and 6000 signals per SNR per mode. Each signal in the both datasets consists of complex IQ. Unlike most deep learning-based AMR methods, where IQ information is directly used as two-dimensional signals, we generate one dimensional complex signal using the given IQ information. To train, validate and test the learning-based AMR models, in our experiments, for each modulation mode per SNR, we randomly split the dataset into training set, validation set and test set, the corresponding ratio is 7:1:2.

B. Experiment Setup

We use three experiments to analyze the effectiveness of the proposed TFA mechanism and the recognition performance of the proposed models. First, we study the effect of different combination modes of the frequency and time attention mechanisms on recognition accuracy. Second, using the proposed

model SCNN with no attention as a baseline, we evaluate the performance of the proposed TFA on SCNN, called TFA-SCNN. After that, we evaluate the performance of the proposed model additionally with Gaussian filter-based noise reduction [8], called TFA-SCNN2, and compare the proposed TFA-SCNN and TFA-SCNN2 with the existing SOTAs with open source code:

- IQ-CNN [2]: a CNN-based method, which uses IQ information as the input.
- CLDNN [12]: a convolutional long short-term deep neural networks (CLDNN)-based method with optimal parameters for AMR.
- AAM-SCNN [13]: a baseline model with an adaptive attention mechanism module (AAM).

For signal representation, we convert the complex signals into spectrogram images using frame-based STFT, with a 95% overlapping Hamming window and a frame length of 40 samples. The resolution of the input spectrograms is $100 \times 100 \times 3$. We normalize all spectrograms before processing, and use root-mean-square prop (RMSprop) as the optimizer. The learning rate starts with 0.0005 and is reduced by a factor of 0.1 when validation loss does not drop within 10 epochs. The training process is terminated when validation loss does not drop within 15 epochs, and the model with the smallest validation loss is saved and used for testing. All experiments are implemented using Keras with Tensorflow backbone and NVIDIA RTX 3090 GPU platform.

C. Experiment Results

Table I shows the results of the ablation experiments on baseline model SCNN. The TFA outperforms all other variants by a significant performance improvement. Whether cascaded CAM-FAM or cascaded CAM-TAM, the recognition accuracy is improved compared to SCNN without attention, which indicates that the attention mechanism extracting meaningful frequency or time features can improve recognition accuracy. In addition, experiment results in Table I show that cascaded FAM and TAM performs worse than the proposed parallel architecture in TFA. The ablation study shows that simultaneous modelling of frequency and time importance from spectrograms in CNN improves recognition accuracy.

TABLE I
ABLATION EXPERIMENT RESULTS ON RADIOML2016.10A

Attention Variant	Accuracy			
	-8 dB	-2 dB	4 dB	10 dB
None	0.372	0.687	0.801	0.823
cascaded CAM-FAM	0.395	0.732	0.818	0.841
cascaded CAM-TAM	0.396	0.740	0.818	0.834
cascaded CAM-TAM-FAM	0.389	0.729	0.811	0.843
proposed	0.426	0.766	0.864	0.857

Fig. 4 shows the recognition accuracy comparison between SCNN and TFA-SCNN versus SNR on RadioML2016.10a and RadioML2016.10b. Compared to the baseline model SCNN, the proposed TFA-SCNN has higher recognition accuracy.

Specifically, on RadioML2016.10a, the recognition accuracy of TFA-SCNN is around 2% to 4% higher than those of the SCNN when SNR is above 10 dB, and around 5% to 8% higher than those of SCNN when SNR is between -8 dB and 10 dB. On dataset RadioML2016.10b, the recognition accuracy of TFA-SCNN is around 1% to 5% higher than those of the SCNN when SNR is above 10 dB. The TFA-SCNN gets around 3% to 9% higher accuracy than SCNN when SNR is between -8 dB and 10 dB.

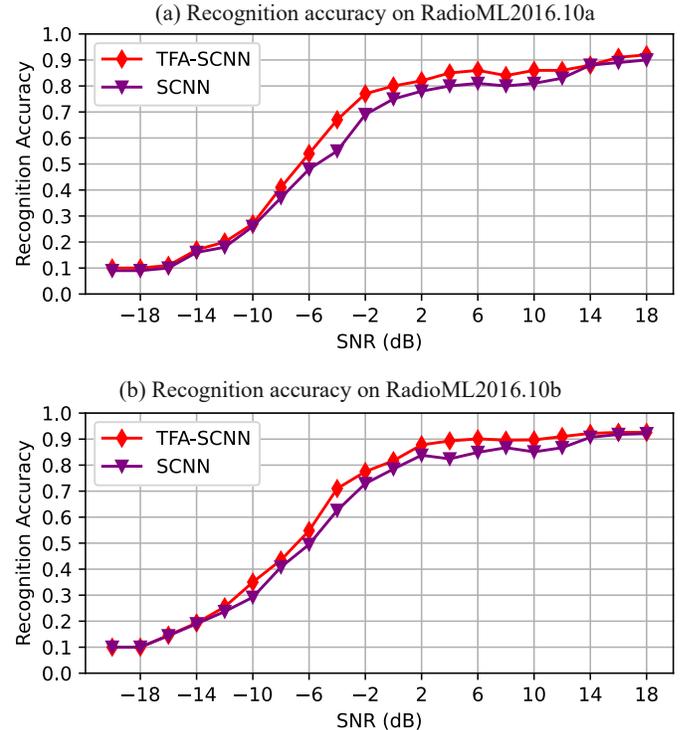


Fig. 4. Recognition accuracy of TFA-SCNN and SCNN.

Fig. 5 shows confusion matrices of TFA-SCNN and SCNN at -2 dB SNR on RadioML2016.10a. The results show that TFA is able to improve the recognition accuracy of all modulation modes, especially for modes: 8PSK, AM-DSB, and GFSK, getting around 15% to 24% performance improvement. The confusion problem between WBFM and AM-DSB is because that both of them belong to analog modulation, and the signal data were generated using the same audio source signal with silent segments, making some of their spectrogram features more difficult to distinguish. Another confusion problem is between 8PSK and QPSK, since the main difference between 8PSK and QPSK is in phase, while spectrograms are weak in representing phase information.

Next, we compare the recognition accuracy of TFA-SCNN and TFA-SCNN2 with IQCNN, CLDNN, and AAM-SCNN on RadioML2016.10a. The experiment results are shown in Fig. 6. We observe that the presented models with TFA (TFA-SCNN and TFA-SCNN2) perform better than the other methods when SNR is above -14 dB. Specifically, TFA-SNN2 performs clearly better than the other methods when SNR is above 2 dB, but the accuracy gets around 3% to 4% lower than TFA-SCNN and AAM-SCNN at 18 dB SNR. This is

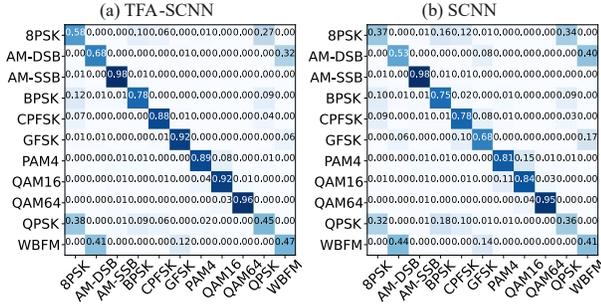


Fig. 5. Confusion matrices on RadioML2016.10a at -2 dB SNR.

consistent with the results of SCNN and SCNN2 in literature [8], since the noise reduction algorithm has limited capability to improve recognition accuracy when signals are severely distorted and close to clean. In addition, the accuracy of TFA-SCNN gets around 1% to 3% higher than that of AAM-SCNN at all SNR levels, and it achieves a maximum recognition accuracy of 92% at 18 dB SNR. This can be explained that the TFA mechanism considers the inherent characters of the time-frequency analysis and extracts important information in terms of channel, frequency and time dimensions, while the AAM mechanism pays attention to channel and spatial information.

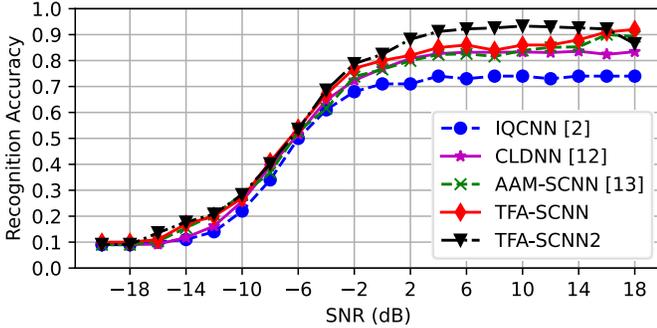


Fig. 6. Recognition accuracy comparison on RadioML2016.10a.

TABLE II
COMPUTATIONAL COMPLEXITY COMPARISON

Model	Training Time	Inference Time	Parameters
IQCNN	0.0383ms	0.0380ms	2830k
CLDNN	0.0477ms	0.0382ms	167k
AAM-SCNN	0.6163ms	0.0395ms	94k
SCNN	0.2071ms	0.0358ms	92k
TFA-SCNN	1.6798ms	0.0391ms	104k

Furthermore, we compare computational complexity between TFA-SCNN and SOTAs in terms of the average training time, the average inference time and the amount of learned parameters. The comparison results are shown in Table II. TFA-SCNN with TFA block costs much more training time (around 1.2ms) compared to the baseline model SCNN, but the inference time of TFA-SCNN increases very little (around

0.003ms). TFA-SCNN has slightly more model parameters than SCNN, but fewer parameters than IQCNN and CLDNN.

V. CONCLUSION

In this work, we presented a CNN-based framework for automatic modulation recognition with a novel TFA mechanism. The TFA is performed on input feature maps to generate attention refined feature maps by learning feature representations for explicitly attending to important channel, frequency, and time information. Experiment results demonstrated the effectiveness of modelling TFA in the CNN front-end, and the presented CNN models with TFA (TFA-SCNN and TFA-SCNN2) outperform three existing learning-based methods from literature. The proposed attention mechanism causes additional computational burden than the baseline model SCNN, but requires similar inference time as the other methods and less learned parameters than IQ-CNN and CLDNN. The performance improvement of the proposed frameworks is incremental in the presence of complex channel environment and low SNRs. As a future work, we plan to extend the method to improve recognition performance at low SNRs using deep learning-based signal enhancement techniques.

REFERENCES

- [1] T. J. O'shea and N. West, "Radio machine learning dataset generation with gnu radio," in *Proceedings of the GNU Radio Conference (GRCon)*, vol. 1, no. 1, 2016.
- [2] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *International Conference on Engineering Applications of Neural Networks (EANN)*, 2016, pp. 213–226.
- [3] R. Li, L. Li, S. Yang, and S. Li, "Robust automated vhf modulation recognition based on deep convolutional neural networks," *IEEE Communications Letters*, vol. 22, no. 5, pp. 946–949, 2018.
- [4] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 3, pp. 433–445, 2018.
- [5] Y. Liu, Y. Liu, and C. Yang, "Modulation recognition with graph convolutional network," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 624–627, 2020.
- [6] K. Yashashwi, A. Sethi, and P. Chaporkar, "A learnable distortion correction module for modulation recognition," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 77–80, 2019.
- [7] M. Zhang, Y. Zeng, Z. Han, and Y. Gong, "Automatic modulation recognition using deep learning architectures," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2018, pp. 1–5.
- [8] Y. Zeng, M. Zhang, F. Han, Y. Gong, and J. Zhang, "Spectrum analysis and convolutional neural network for automatic modulation recognition," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 929–932, 2019.
- [9] T. Wang, Y. Hou, H. Zhang, and Z. Guo, "Deep learning based modulation recognition based on multi-cue fusion," *IEEE Wireless Communications Letters*, vol. 10, no. 8, pp. 1757–1760, 2021.
- [10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.
- [11] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.
- [12] N. E. West and T. O'Shea, "Deep architectures for modulation recognition," in *IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2017, pp. 1–6.
- [13] Z. Liang, M. Tao, L. Wang, J. Su, and X. Yang, "Automatic modulation recognition based on adaptive attention mechanism and resnext wsl model," *IEEE Communications Letters*, 2021.
- [14] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011, pp. 315–323.