Augmented Convolutional Neural Networks with Transformer for Wireless Interference Identification

Pengyu Wang, Yufan Cheng, Binhong Dong
National Key Laboratory of Science and Technology on Communications
University of Electronic Science and Technology of China
Chengdu, China
e-mail: 201911220507@std.uestc.edu.cn, {chengyf, bhdong}@uestc.edu.cn

***Abstract*—As electromagnetic environments are more and more complex, wireless interference identification (WII) is becoming vital for non-cooperative communication systems in both civilian and military scenarios. With the enormous success of deep learning (DL), methods that optimize convolutional neural networks (CNNs) for WII have been proposed. However, due to the intrinsic characteristics of CNN, the existing networks are difficult to capture long-range feature dependencies, causing the low recognition accuracy and the high computational complexity. Motivated by the success of transformers in natural language processing (NLP) domain, we propose an augmented convolutional neural network with transformer (ACNNT), which combines both the advantages of CNNs and transformers to simultaneously strengthen locality and establish long-range dependencies. Specifically, the ACNNT has multiple stages, and every stage consists of a convolutional layer and a transformer module to model local and long-range dependencies of context, respectively. At the end of the network, a classification token is used for classification. A channel attention (CA) module is proposed to further improve the expressive ability of the transformers. Extensive experiments demonstrate that the proposed method leads to performance improvement as compared to conventional DL based methods.**

***Index Terms*—Wireless interference identification (WII); convolutional neural networks; transformer**

I. Introduction

Wireless communication technology is in an era of rapid development and dramatic evolution. With the dramatic growth of number and diversity of wireless electronic equipments, radio frequency spectrum is becoming increasingly scarce. Multiuser communication may suffer from interference among users in non-cooperative communication systems, in which different users compete for limited resources [1]. In addition, hostile interference poses a huge threat to wireless communication systems [2]. It is critical and essential for intensifying the cognition capabilities of the communication systems. Wireless interference identification (WII) is a promising technology for understanding complex electromagnetic environments via recognizing types of interference signals, and it is widely used in the civilian and military domains.

Due to distinguished characteristics of different interference signals, WII is modeled as a pattern recognition problem based on feature extraction. Reference [3] combined high order cumulant and joint diagonalization to separate and analyze interference. The results obtained satisfactory recognition accuracy for amplitude modulation (AM) interference and frequency modulation (FM) interference. In [4], a feature extraction based WII algorithm was proposed. The features of interference signals included spectral bandwidth, power spectrum flatness, fractional degree of aggregation, etc. The experiments showed that the proposed method could lead to better performance in classification accuracy. In [5], the different characteristics in the amplitude undulation, high order cumulant and bispectrum were compared and analyzed to identify deception interference. Although the feature-based methods are effective, expert knowledge with hand-crafting engineering of features is required.

Recently, following the success in advancing natural language processing and computer vision [6]–[8], deep learning (DL) is expected to bring revolutionary changes to wireless communication [9]. At the same time, many researchers have introduced DL into WII. Reference [10] proposed the first WII approach based on convolutional neural networks (CNNs) for classifying various types of interference siganls. The experiments showed that the method outperformed traditional methods. Reference [11] investigated that three independent deep neural networks (DNNs) recognized cooperatively interference signals. The results demonstrated that utilization of DNNs in sensor system could improve the classification accuracy for WII. To address the problem of the degeneration of recognition accuracy under low interference-to-noise ratio (INR) conditions, reference [12] incorporated residual blocks and asymmetric convolution blocks into plain CNNs to strengthen representative power of the model. Compared with traditional methods, the proposed method achieved better and stable recognition performance. The authors in [13] applied CNNs to identify active jamming, and results demonstrated that CNNs have strong ability to distinguish active jamming. In [14], a CNN-based siamese network was presented for WII to solve the problem of limited training samples. The sufficient experiments verified the effectiveness of proposed methods.

However, for WII task, most of the existing works adopt CNNs, which tend to focus on capturing local context while ignoring long-range dependencies. The intrinsic characteristic of CNNs inevitably causes the loss of recognition performance. Inspired by [8], to overcome this problem, we present an augmented convolutional neural network with transformer (ACNNT) to exploit both long-range and local contexts. To the best of our knowledge, this is the first research to introduce

transformers for WII. Specifically, our proposed model has several stages and every stage is composed of CNN layers and the transformer module. The transformer module consists of a multi-head self-attention (MHA) layer and a multi-layer perceptron (MLP) layer. We also design channel attention (CA) for the transformer to further improve the recognition performance for WII.

II. Preliminaries

A. Signal Model

In this paper, we consider the interference signal model under the additive white Gaussian noise (AWGN) channel, which can be formulated as,

$$r(t) = e^{j(2\pi f_c t + \Phi(t))} s(t) + n(t) \tag{1}$$

where $r(t)$ is the received signal, $s(t)$ is the unknown interference signal, and $n(t)$ is Gaussian white noise. The carrier frequency of $s(t)$ is f_c, and the initial phase is referred as $\Phi(t)$.

The objective of WII is to recognize types of interference $s(t)$ without any prior information. Generally speaking, the methods of WII try to seek a mapping relationship between $r(t)$ and $s(t)$. The maximum-a-posterior (MAP) can be adopted to solve this problem. Formally, the posterior probability density function is denoted as $P(s(t) = c|\, y(t))$, where $c \in C$ is one of interference type candidates, and C is all types of interference in task. According to MAP criterion, the final prediction c^* can be obtained as follows,

$$c^* = \arg\max_{c \in C}(P(s(t) = c|\, y(t))) \tag{2}$$

B. Time-Frequency Distribution

Considering that the form of the received signal has a significant influence on the identification of the interference signal $s(t)$, we need to apply the appropriate transformation to $y(t)$. We carry out short-time Fourier transform (STFT) for $y(t)$ to simultaneously model the time-frequency domain characteristics of interference in this paper. Interference signals in this paper include eight types: sing-tone (ST), binary phase shift keying (BPSK), amplitude modulation (AM), noise amplitude modulation (NAM), sinusoidal frequency modulation (SFM), liner frequency modulation (LFM), quadrature frequency shift keying (4FSK), binary frequency shift keying (2FSK). Fig. 2 shows time-frequency images (TFIs) of these interference signals when the INR is 8 dB.

III. The Proposed Scheme

Our proposed ACNNT, in which a stack of convolutional layers is interleaved with transformer module, allows every sample to establish long-range and local dependencies of context. Before presenting our proposed methods for WII, we briefly review the recent mainstream CNN-based WII approaches.

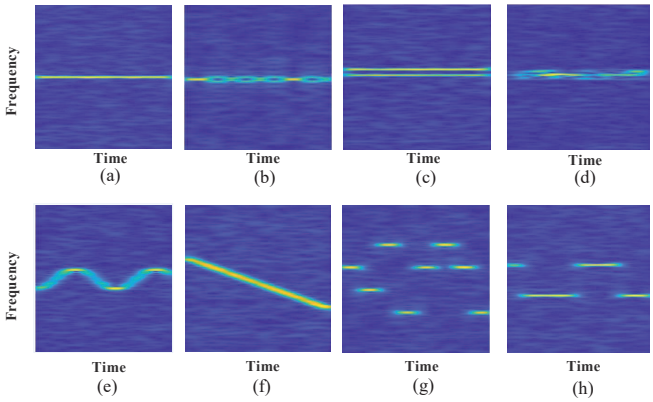

Fig. 1. TFIs of (a)ST, (b)BPSK, (c)AM, (d)NAM, (e)SFM, (f)LFM, (g)4FSK, (h)2FSK.

A. Revisiting CNN-based WII

The WII methods can be modeled as classification problems. CNNs take TFIs or power-spectrum data as input I, and corresponding labels Y are prepared at same time. The CNNs can automatically learn features of input through self-optimization during data-driven supervised learning process. Finally, CNNs map the relationship from I to Y by minimizing the following risk function,

$$R(f) = R_e(f) + \lambda_J J(f), \tag{3}$$

where $R_e(f)$ is empirical risk function, and cross entropy loss is often used for this term. $J(f)$, representing the complexity of the model, is also added to the risk function to avoid over-fitting. λ_J is a hyperparameter to balance $R_e(f)$ and $J(f)$. However, CNNs have the intrinsic characteristic, $e.g.$, focusing on capturing the local context while missing long range context for exploring. To this end, we propose a novel model, which can attend to not only the local but also global context.

B. The Proposed Method

Motivated by the success of transformers in establishing long-range dependencies [8], we aim to improve recognition accuracy of WII by combining the advantages of CNNs and transformer. The overall pipeline of our proposed model is presented in Fig. 2. The proposed ACNNT has M stages and every stage has a similar architecture, which is composed of a transformer encoder layer and a convolutional block. To simplify the description, we take the first stage as an example to elaborate the proposed model.

We denote input of ACNNT as $x \in R^{H \times W \times C}$, where H and W are the height and weight of the original size of the input, C is the number of channels. Firstly, x is processed by a convolutional block. The convolutional block consists of several convolution layers followed by batch normalization (BN). The convolution kernel of each convolution layer is $k \times k$ with k strides. This operation is calculated as:

$$x_c = f_{conv}(x, \theta) \tag{4}$$

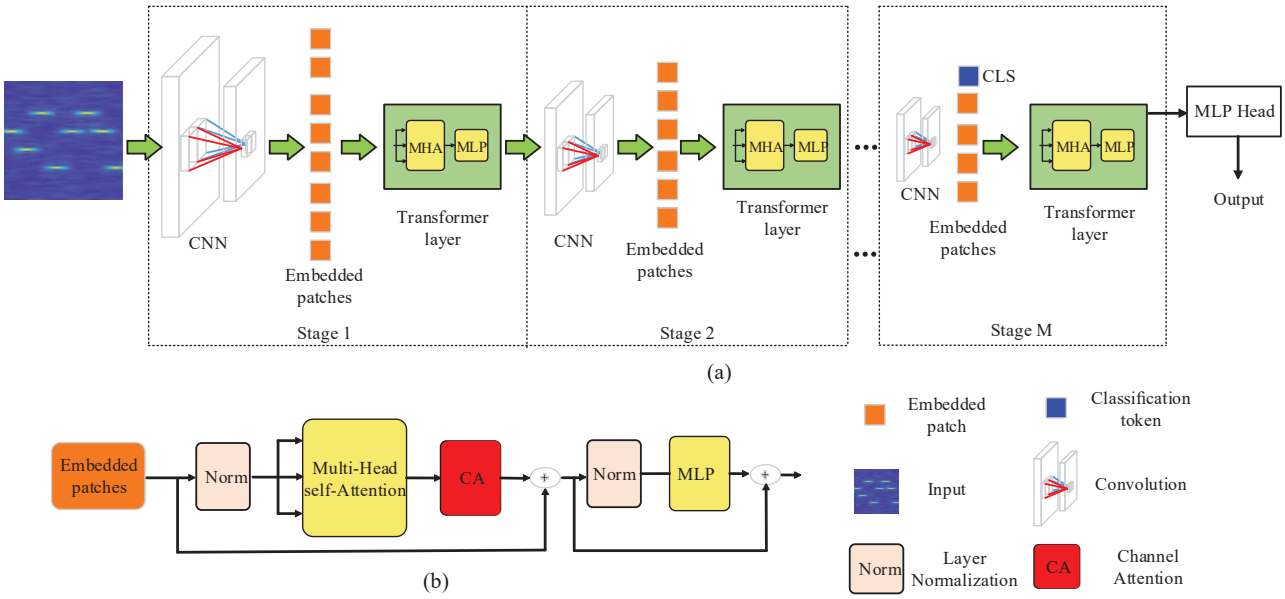

Fig. 2. The architecture of our proposed ACNNT for WII. The model has M stages and every stage is composed of a transformer layer and a convolutional block to simultaneously capture global and local dependencies. (a) the overall pipeline, (b) the transformer layer in this paper.

where $f_{conv}(\cdot,\theta)$ represents the convolutional block with parameters θ. The $x_c \in R^{H'\times W'\times C'}$ is output, where (H', W') is the size of x_c, and the number of channels is C'.

Next, the x_c will be processed by the transformer to capture long-range dependencies. The feature map x_c is needed to be sequentialized because the transformer layer requires a 1D sequence of tokens. Specifically, we set size of every patch to 1×1, so x_c is evenly divided into N patches and then flatten to patch embeddings $\mathbf{x}_p \in R^{N\times C'}$, where $N{=}H' \times W'$ is the total number of these patches. N trainable position embeddings $PE \in R^{N\times C'}$ is added to every feature embedding of patches. The sum of two embeddings is referred as $X \in R^{N\times D}$, where $D = C'$ and X becomes the input sequence of the transformer. The transformer layer consists of multi-head self-attention (MHA) layer and the multi-layer perceptron (MLP). A matrix of query $Q{=}XW_Q$, a matrix of key $K{=}XW_K$ and a matrix of value $V{=}XW_V$ are three inputs of MHA layer, where $X \in R^{N\times D}$ mentioned before is the input sequence, $W_Q \in R^{D\times d}$, $W_K \in R^{D\times d}$, $W_V \in R^{D\times d}$ are trainable parameters. The $d = \frac{D}{h}$ is vector size, and h represents the number of attention heads of MHA. Details of the self-attention can be formulated as,

$$Att(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{5}$$

The h heads are added to self-attention for constructing MHA as follows,

$$\mathrm{MHA}(Q,K,V) = \mathrm{concat}(\mathrm{hd}_1, \mathrm{hd}_2, ..., \mathrm{hd}_h)W \tag{6}$$

$$\mathrm{hd}_h = Att(Q_h, K_h, V_h) \tag{7}$$

where W is a learnable Fully connected (FC) layer, $\mathrm{hd}_h \in R^{N\times d}$ represents the h-th head of Att self-attention.

The output of MHA is $X' \in R^{N\times D}$. Finally, MLP sublayer converts X' to $X'' \in R^{N\times D}$. The MLP module is composed of two FC layers with Gaussian Error Linear Unit (GELU) activation function, which can be formulated as,

$$\mathrm{MLP}(X'){=}\ \mathrm{GELU}(X'W_1{+}b_1)W_2 + b_2 \tag{8}$$

where $W_1 \in R^{D\times d_m}$ and $W_2 \in R^{d_m\times D}$ are weights of two FC layers, and $b_1 \in R^{d_m}$ and $b_2 \in R^D$ are bias vectors. In short, we summarize these processes as follows,

$$\begin{aligned} &[x_p^1; x_p^2; ...; x_p^N] = flat(split(f_{cov}(x,\theta)) \\ &X = [x_p^1; x_p^2; ...; x_p^N] + PE \\ &X' = MHA\,(\mathrm{LN}(X)) + X \\ &X'' = MLP\,(\mathrm{LN}(X')) + X' \end{aligned} \tag{9}$$

where LN is layer normalization, and $X' \in R^{N\times D}$ and $X'' \in R^{N\times D}$ are output of MHA and MLP of transformer layer, respectively. Before processing by next stage, $X'' \in R^{N\times D}$ is converted to $R^{H'\times W'\times D}$. Since each stage has strided convolutions, the number of patch embeddings is gradually decreasing, which makes the latter transformer layers consume less computational resources. In the final stage, we add classification token x_{cls} for classification. Unlike ViT [8], we added x_{cls} in the last transformer layer.

C. Channel Attention for Transformer Layers

We propose channel attention (CA) module for the transformer in this paper. The CA module can exploit the relationship among channels and focus on more important and meaningful channels. We insert CA between the MHA and MPL layer, as shown in the Fig. 2 (b).

The output of MHA is $X' \in R^{N\times D}$, and we transpose X' as $F \in R^{D\times N}$. We compute channel statistics by applying globe average pooling (GAP) and globe max pooling (GMP) along

the spatial dimension. The motivation of applying two pooling is to obtain finer channel-wise attention. This process can be formulated as:

$$F_a{=}GAP(F), \tag{10}$$

$$F_m{=}GAP(F), \tag{11}$$

where $F_a \in R^{D\times 1}$ and $F_m \in R^{D\times 1}$ denote average pooled features and max pooled features, respectively. Channel attention can be obtained by implementing a trainable linear projection on F_m and F_a. In this paper, the linear projection is composed of two FC layers. To reduce computation and memory, the r reduction ratio is introduced. Specifically, The linear projection can be expressed as $LP = W'(W''(\cdot))$, where W' and W'' represent the weights of the two FC layers with neurons of D/r and D. The GELU activation function is followed by W''. Finally, the output of CA is expressed as:

$$\overline{F}{=}(\sigma(LP(F_m) + LP(F_a)) \times F)^T, \tag{12}$$

where T is transpose operation, $\overline{F} \in R^{N\times D}$ is output of CA, and σ is sigmoid function. During multiplication, the attention values $\sigma(LP(F_m){+}LP(F_a)) \in R^{D\times 1}$ is needed to broadcast along the spatial dimension.

D. Loss Function

The training TFIs and corresponding labels are denoted as (x_i, y_i) ,$i = 1, 2, ..., I$, where $x_i \in R^{60\times 60\times 3}$ is i-th sample with 60×60 resolution size and three channels. The y_i represents its corresponding label. The cross entropy loss is applied for the loss function in this paper, which can be expressed as:

$$L = -\sum_{i=1}^{I} \mathbf{q}(x_i) log(\mathbf{p}(x_i)), \tag{13}$$

where $\mathbf{p}()$ is the predicted probability distribution, while $\mathbf{q}()$ is the real probability distribution of samples.

IV. Experiments

We evaluate the proposed methods for WII in this section. The eight kinds of interference signals (ST, BPSK, AM, NAM, SFM, LFM, 4FSK, 2FSK, mentioned in section II) are generated in Matlab. The dataset comprises 500 TFIs of every INR per interference category for training, and 100 samples every INR per interference category for testing, where the INR is from -14 dB to 8 dB with 2 dB as an interval. We conduct the experiments in python with GPU.

A. Model Structures

In order to perform comprehensive comparisons, the CNN [6] and the Resnet [7], which are widely used for WII, are applied to evaluate. In addition, ViT [8] is tested simultaneously for the same task of WII. The architectures of CNN and Resnet in this paper are shown in Table I and Table II, respectively. Fig. 3 shows the structure of ResBlock. Considering that the resolution of input TFIs is only 60×60, we adopt both the shallow CNN and the Resnet. The proposed ACNNT is showed as Table III. It has M stages and each stage is composed of a convolutional layer and a transformer layer.

B. Recognition Accuracy

For fair comparisons, we will compare the recognition accuracy of these models under approximately equal computational complexity. The floating-point operations per second (FLOPs) is widely used to measure of computational complexity. Firstly, we fix structures of the Resnet and the CNN and adjust the number of channels to keep FLOPs of these models close to one Million. Specifically, the channels of the two convolution layer (U_{c1} and U_{c2}) of CNN are both set to 8. The channels of the first convolutional layer and the residual block of resnet are both 4, namely $U_{r1} = U_{r2} = 4$. For the ViT, we slice input TFIs into 15×15 patches of 4×4 pixels, embed them into vectors of dimension 32. The d_m in Eq. (8) of MLP is 32 for ViT. The number of the transformer layer and attention heads is both set to 4. For proposed ACNNT, we set M to 4, r reduction ratio to 1 for CA, and U_{t1} and d_m to 32. The kernel size k is set to be 3, 2, 2, 1 for 4 stages, respectively.

We give FLOPs and parameters of these models in Table IV. From the table, we can find that the FLOPs of all the model is close to one Million. ACNNT-CA represents incorporating CA into ACCNT. The proposed models (ACNNT and ACNNT-CA) have less FLOPs than other models. In addition, we observe that ViT, ACNNT and ACNNT-CA have more parameters than the CNN and the Resnet. This is because ViT and the proposed models have more layer normalization and FC layer, which occupy more storage resources. The introduction of CA only adds few FLOPs and parameters to ACNNT.

The classification accuracy is shown in Fig. 4 and Table V.

TABLE I
STRUCTURE OF CNN IN THIS PAPER FOR WII

NO.	Type	structure
-	-	Input (3,60,60)
1	Conv	Conv2D(U_{c1},3×3) + BN +ReLU +Maxpool(2,2)
2	Conv	Conv2D(U_{c2},3×3)+BN+ReLU +Maxpool(2,2)
3	FC	Dense(8)+softmax

Conv2D(U_{c1},3×3) represents a 3×3 convolution kernel with U_{c1} channels, and U_{c2} represent the number of channels for second convolution. ReLU refers to activation function, Maxpool(2,2) denotes max pooling with 2×2 kernel size, and Dense(8) denotes FC layer with 8 neurons.

TABLE II
STRUCTURE OF RESNET IN THIS PAPER FOR WII

NO.	Type	structure
-	-	Input (3,60,60)
1	Conv	Conv2D(U_{r1},3×3) + BN +ReLU +Maxpool(2,2)
2	Residual block	ResBlock (U_{r2},3×3)
3	FC	Dense(8)+softmax

U_{r1} and U_{r2} represent the number of channels, and ResBlock(U_{r2},3×3) represents a 3×3 residual building block with U_{r2} channels. Fig. 3 shows the structure of ResBlock.

It can be seen from the Figure and the Table that (i) the models with the transformer structure (like ViT, ACNNT and ACNNT-CA) have better performance than the models without the transformer structure (like Resnet or CNN) under

TABLE III
STRUCTURE OF ACNNT

NO.	Type	structure	
-	-	Input (3,60,60)	
1	Conv	Conv2D(U_{t1},$k \times k$)	$\times M$
2	Transformer layer	MHA+MLP	
3	FC	Dense(8)+softmax	

U_{t1} represents the number of convolution channels, $k \times k$ denotes kernel size of the convolution.

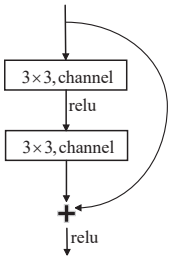

Fig. 3. The structure of residual block (ResBlock) in Table II.

approximate FLOPs. For instance, when the INR is equal to -8 dB, the recognition performance of ACNNT is about 22% and 56% better than that of CNN and Resnet, respectively. This is because structure of transformer layer has global receptive field, and this property allow the model to provide more accurate and superior predictions when compared to CNNs. (ii) Due to the combination of the advantages of CNN and Transformer, the proposed ACNNT has better recognition performance than ViT. From the Table, compared to ViT, ACNNT has a 1.6% performance improvement on the average recognition rate under all INRs. (iii) Due to strengthening the important channels, the proposed ACNNT-CA have brought 0.8% performance improvement on the average recognition accuracy for ACNNT.

For comprehensive comparisons, we have compared the recognition performance of these models when the FLOPs are near 20 Million.

Specifically, the channels U_{c1} and U_{c2} are both set to 44 for the CNN. For the Resnet, we set $U_{r1} = 16$ and $U_{r2} = 17$. In addition, for the ViT, we divide input TFIs into 12×12 patches of 5×5 pixels, embed them into vectors of dimension 512. The number of the transformer layer and attention heads is both set to 8. The d_m of MLP is 1024. For our proposed ACNNT, M is set to 8, r reduction ratio to 8 for CA, attention heads to 8, channels U_{t1} to $(64, 64, 128, 128, 256, 256, 512, 512)$ and d_m to $(512, 512, 1024, 1024, 1024, 1024, 2048, 2048)$ for every stage, respectively.

FLOPs and parameters of these models is given in Table VI. From the table, we can find that the FLOPs of all the models are close to 20 Millon as expected. Due to more layer normalization and FC layers, we observe that parameters of ViT, ACNNT and ACNNT-CA are significantly higher than CNN and Resnet. The parameters of ACNNT and ACNNT-CA are significantly less than those of ViT. This is because the proposed network gradually reduces the number of embedded patches by introducing convolutional layer of every stage, while ViT keeps the number of embedded patches constant for all transformer layers.

The classification accuracy is shown in Fig. 5 and Table VII when FLOPs is around 20 Million.

It can be seen from the Figure and the Table that (i) due to increasing computing resources, all models improve the recognition performance for WII. (ii) Here again, the ACCNT has better performance than Resnet, CNN and ViT. For instance, when the INR is equal to -8 dB, the recognition performance of ACCNT is about 15%,16% and 4% better than that of CNN, Resnet, and ViT, respectively. (iii) The proposed CA brings 0.9% performance improvement on the average recognition accuracy for ACCNT.

TABLE IV
PARAMETERS AND FLOPS FOR DIFFERENT MODELS

structure	CNN	Resnet	ViT	ACNNT	ACNNT-CA
FLOPs	1.36M	1.48M	0.99M	0.92M	0.94M
Parameters	912	460	0.14M	0.03M	0.05M

M represents million.

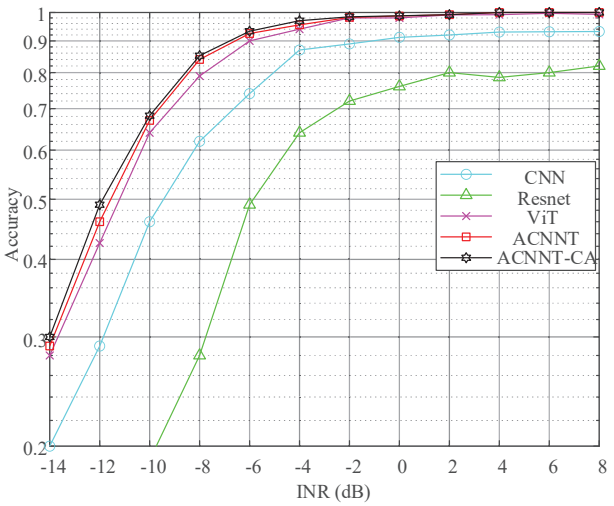

Fig. 4. The accuracy for different models under FLOPs near 1 Millon.

TABLE V
ACCURACY FOR DIFFERENT MODELS UNDER FLOPS NEAR 1 MILLON

Method	CNN	Resnet	ViT	ACNNT	ACNNT-CA
Accuracy	72.4%	54.6%	82.5%	84.1%	84.9%

TABLE VI
PARAMETERS AND FLOPS FOR DIFFERENT MODELS

Method	CNN	Resnet	ViT	ACNNT	ACNNT-CA
FLOPs	20.36M	20.95M	19.35M	19.9M	20.3M
Parameters	0.02M	0.006M	16.85M	9.02M	9.36M

TABLE VII
ACCURACY FOR DIFFERENT MODELS UNDER FLOPS NEAR 20 MILLON

structure	CNN	Resnet	ViT	ACNNT	ACNNT-CA
Accuracy	78.7%	80.6%	85.3%	86.9%	87.8%

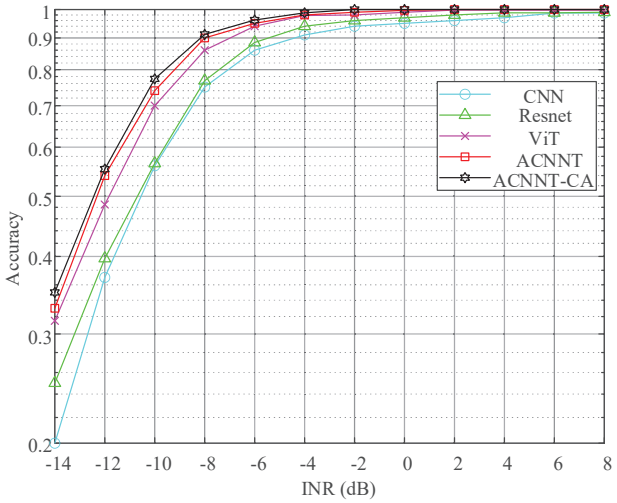

Fig. 5. The accuracy for different models under FLOPs near 20 Millon.

C. Convergence Curve

We present the training loss of different models in Fig. 6 when FLOPs is near one Million.

From the Figure, we can see that with the increasing of epoch, the training loss of each model gradually decreases. Among them, the training loss of the ACNNT-CA is lower than that of the other models. This process mainly lies in the high training efficiency with the aid of the ability at modelling both long-range and short-range dependencies by ACNNT-CA.

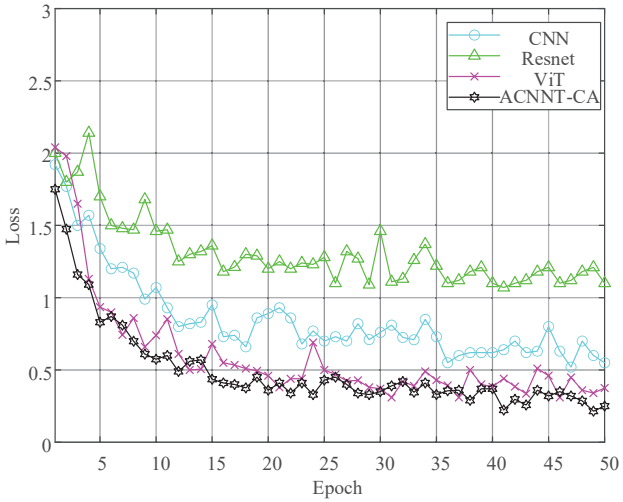

Fig. 6. The training loss for different models under FLOPs near 1 Millon.

V. CONCLUSION

In this paper, we have proposed ACNNT for WII. Our ACNNT is composed of a stack of convolutional layers and transformer modules, which has the powerful ability to model long-range and local dependencies of context. In addition, we improve the performance of ACNNT by inserting CA module into transformer module. Experimental results show that the proposed methods are better than the traditional methods for WII when the computational complexity (FLOPs) of these models is approximately close.

ACKNOWLEDGMENT

This work was supported by the National Nature Science Foundation of China under Grant U19B2014 and the National Key R&D Program of China (Grant No. 254).

REFERENCES

[1] A. Y. Zakariya, A. F. Tayel and S. I. Rabia, "Comments on "Optimal Target Channel Sequence Design for Multiple Spectrum Handoffs in Cognitive Radio Networks"," in IEEE Transactions on Communications, vol. 63, no. 8, pp. 3021–3024, Aug. 2015.

[2] C. Li, P. Qi, D. Wang and Z. Li, "On the Anti-Interference Tolerance of Cognitive Frequency Hopping Communication Systems," in IEEE Transactions on Reliability, vol. 69, no. 4, pp. 1453–1464, Dec. 2020.

[3] D. Wei, S. Zhang, S. Chen, H. Zhao, and L. Zhu, "Research on anti-jamming technology of chaotic composite short range detection system based on underdetermined signal separation and spectral analysis," IEEE Access, vol. 7, pp. 42298-C42308, 2019.

[4] G. Wang, Q. Ren and Y. Su, "The interference classification and recognition based on SF-SVM algorithm," 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN), Guangzhou, 2017, pp. 835–841.

[5] Li Jian-xun, Shen Qi and Yan Hai, "Signal feature analysis and experimental verification of radar deception jamming," Proceedings of 2011 IEEE CIE International Conference on Radar, Chengdu, China, 2011, pp. 230–233.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", Proc. Int. Conf. Learn. Representat. pp. 1-14, 2015.

[7] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby," An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," In International Conference on Learning Representation (ICLR), 2021.

[9] G. Gui, M. Liu, F. Tang, N. Kato, and F. Adachi, "6G: opening new horizons for integration of comfort, security, and intelligence," *IEEE Wireless Commu.*, vol. 27, no. 5, pp. 126–132, Oct. 2020.

[10] M. Schmidt, D. Block and U. Meier, "Wireless interference identification with convolutional neural networks," 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), Emden, 2017, pp. 180-185.

[11] J. S. Patel, F. Fioranelli, M. Ritchie, and H. D. Griffiths, "Fusion of deep representations in multistatic radar networks to counteract the presence of synthetic jamming," IEEE Sensors J., vol. 19, no. 15, pp. 6362C6370, Aug. 2019.

[12] Q. Qu, S. Wei, S. Liu, J. Liang and J. Shi, "JRNet: Jamming Recognition Networks for Radar Compound Suppression Jamming Signals," in IEEE Trans. Veh. Technol., vol. 69, no. 12, pp. 15035–15045, Dec. 2020.

[13] Y. Wang, B. Sun and N. Wang, "Recognition of radar active-jamming through convolutional neural networks," in The Journal of Engineering, vol. 2019, no. 21, pp. 7695-7697, 11 2019.

[14] G. Shao, Y. Chen and Y. Wei, "Convolutional Neural Network-Based Radar Jamming Signal Classification With Sufficient and Limited Samples," in IEEE Access, vol. 8, pp. 80588-80598, 2020.