A Pruning Neural Network for Automatic Modulation Classification

Zherui Zhang*
Harbin Engineering University
zhangzherui@hrbeu.edu.cn

Ya Tu
Harbin Engineering University
tuya@hrbeu.edu.cn

***Abstract*—Automatic modulation classification (AMC) is a promising technology for non-cooperative communication systems in both military and civilian scenarios. Nowadays, More and more scholars apply deep learning (DL) framework to AMC. However, most of the papers do not consider that the typical deep learning model is difficult to deploy on the resource constrained devices. In this paper, a lightweight DL based Average percentage of zeros (APoZ) is used with pruning neural. We introduce a novel method of generating modulation signals called contour stellar image (CSI). We train the data through some scaling factors in convolution neural network (CNN) especially the AlexNet. It can screen out the inconsequential neurons which can be pruned. Experimental results suggest that using APoZ to prune can not only slim the network but also stabilize the average error about 1.05% compared with the original network.**

Keywords: Automatic modulation classification(AMC); Contour Stellar Image(CSI); Average percentage of zeros(APoZ); neuron pruning

I. INTRODUCTION

Automatic modulation classification (AMC) is a novel and convenient modern technology in the non-cooperative communication scenarios without agreement or authorization between the transmitters and the receivers. This technology has been extensively used in civil and military fields. N. Lay *et al* [1] proposed two classification techniques for digitally modulated signals affected by inter symbol interference (ISI). It has a great accuracy in modulation recognition. The author focused on using a criterion to measure the redundancy of convolution kernel in network and analyzing the influence of network recognition accuracy between before and after pruning.

In recent years, deep learning (DL) has been considered one of the most effective tools to solve various problems in wireless communications [2]-[6]. O'Shea *et al* [7] used the deep learning to recognize the radio signal classification, and the recognition accuracy almost ranges from 84% to 96% on the hard 24-class modulation signals. This paper also fully reflects that deep learning can also be used in the field of modulation recognition. But DL models generally have big size and complex calculation. In order to solve this problem, researchers began to learn about the lightweight network as well as ensure that the accuracy of modulation recognition can still be maintained at a high level. Wang *et al*. [8] changed pooling operation in pooling layer instead of dropout in CNN. It reduces the kernels between anterior and posterior layers. In 2016, an algorithm named average percentage of zero (APoZ), pruning the kernels of convolution layers to reduce the amount of floating-point calculation, is proposed by Hu *et al* [9]. What's more, in 2017 A. Kadav *et al* [10], the paper is proposed a methods used in studies involving weight sum.

In order to verify whether the effect of pruning is reliable in the case of high recognition rate, we use CSI data set with high recognition rate and APoZ pruning method to prove the stability of the network with high recognition rate.

The remainder of this paper is organized as follows: Section II mainly discusses generating the digital signals with CSI. Section III introduces the effect with AlexNet training. Then, an experiment is conducted in Section IV to prune the network with APoZ. Finally, a conclusion to our used method is provided in Section V.

II. RELATED WORK

Based on the paper [13], which was proposed a convolution neural network (CNN)-based AMC method applying the in-phase and quadrature (IQ) component of signals, we use constellation diagrams (CD) with feature enhancement to train network, which is called contour stellar image (CSI). The idea about CSI, generated based on the CD, is as follows: considering the impact of multiple sample points, we choose a window function to slide on the CD and compute the points' density in different regions, and then we draw different colors to distinguish this feature. Considering a rectangular window function with size $(W \times H)$, where the W is the width and the H is height of the windows. We use a metric *dot density* and a *center dot* to indicate as the center of a sampling point, how many dots are located in the rectangular window of the complex-valued plane. We set the coordinate of the center point in the complex plane as (i, j), then we can normalize the dot counts and obtain the dot density as:

$$\rho_{(W,H)}(x_i, y_i) = \frac{\sum_{m=1}^{Num} J(x_i, y_i)}{Num}, \qquad (1)$$

Where $J(x_i, y_i)$ denotes the number of dots which is in rectangular window, if the dot (x_i, y_i) falls inside the rectangular window, the function equals one, else equals zero. Num is the length of an I/Q signal frame. Next we use different color to render the CD, the dot around the place with high density is bright, with low density is dark. The entire process is depicted in Fig. 1.

We know that CD is a binary image, it just can distinguish whether this pixel has a dot or not. It can't distinguish how many dots in the same pixel. In contrast, the color in CSI will tell more details of feature about the modulation signal. In highly noisy communication environments, CSI can provide the difference of

2767-6684/21/$31.00 ©2021 IEEE
DOI 10.1109/DSA52907.2021.00029

characteristics among different modulation signals since it considers the number of overlaps in the same pixel. With color brightness as statistics, it can resist influence of some small noise on the recognition accuracy about modulation signal, and improve the recognition accuracy in low Signal to Noise Ratio (SNR).

To sum up, CSI is different from the traditional modulation recognition method, which fully compromises the advantages between the modulation recognition and CNN of the computer vision, especially convolutional feature learning. It is high potential that has to improve the signal detection and classification performance of practical systems by generalizing well and remaining sensitive to very low power signals. And it so skillfully solves the problem of binarization of CD that the dots covered by the same pixel are no longer ignored, which makes them an important factor in feature recognition.

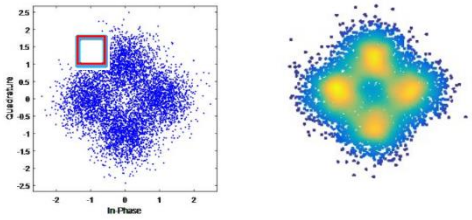


Fig. 1 CSI process

III. THE NETWORK ABOUT CNN

A. Introduction About CNN

Convolution neural network is often used in image recognition and classification. It is a kind of feedforward neural network with convolution computation and depth structure. Its main structure includes convolution layer, pooling layer, fully connected layer and dropout layer. Convolution layer is mainly used to convolute the image and extract image features. It is the most important part of CNN, and most of the floating-point operations are also in convolution layer; the main effect of pooling layer is subsampling, removing redundant information from images. It reduces the network computation to a certain extent; fully connected layer acts a classifier. It maps the implicit features extracted from the convolution layer to the sample space and achieves the effect of recognition and classification; dropout layer mainly used to improve the phenomenon of overfitting, which can make the network forget some unique features in the training image, so as to enhance the universality of the network.

B. AlexNet CNN In Experiment

In this experiment, we apply the AlexNet to train modulation signals. The AlexNet was designed in 2012, which has 60 million parameters and 6500 million neurons, five convolution layers, three fully connected layers, and the final output is 1000 channels. Because of eight kinds of modulation signals used, which are 4ASK, BPSK, QPSK, OQPSK, 8PSK, 16QAM, 32QAM, and 64QAM, we have made some corresponding adjustments, which is change the last layer into 8 channels. The following table shows the structure of AlexNet we change.

TABLE I. THE STRUCTURE OF ALEXNET

Layer(type)	*Output shape*	*Parameter*
Input	277 × 277 × 3	
Convolution	55 × 55 × 96	Core=11 × 11 Stride=4 Kernels=96
Max pooling	27 × 27 × 96	Core=3 × 3 Stride=2
Convolution	27 × 27 × 256	Core=5 × 5 Padding=2 Kernels=256
Max pooling	13 × 13 × 256	Core=3 × 3 Stride=2
Convolution	13 × 13 × 384	Core=3 × 3 Padding=1 Kernels=384
Convolution	13 × 13 × 384	Core=3 × 3 Padding=1 Kernels=384
Convolution	13 × 13 × 256	Core=3 × 3 Padding=1 Kernels=256
Max pooling	6 × 6 × 256	Core=3 × 3 Stride=2
Fully connected	1 × 4096	ReLu
Fully connected	1 × 4096	ReLu
Fully connected	1 × 8	Softmax

Tips: in this table, the parameter, padding, means supply blank pixels around the images. It is mainly used to let the convolution kernel extract the edge information of the image. The main activation functions in the fully connected layer are ReLu (The Rectified Linear Unit) and Softmax function, which is responsible for mapping the input of neuron to the output.

IV. PRUNING TECHNOLOGY EXPERIMENT

In this experiment, we apply eight kinds of digital modulation signals. we generate the sequence of modulation signals, each symbol collects eight sample complex-valued points, and a frame, which is drew on an image as a CSI, has about 625 symbols, so there are about $625 \times 8 = 5000$ points on the graph. Each modulated signal category has 10000 labeled frames for training and 1000 labeled frames for testing, which means there will be 80000 labeled training frames and 8000 labeled test frames in total.

Then we send the training set to the network for training. The training set includes nine kinds of data which have -6dB to 10dB stepping 2dB added noise is Gaussian white noise. The figure below shows some modulation signals including 4ASK, BPSK, QPSK and 16QAM between -6dB and 4dB SNR.

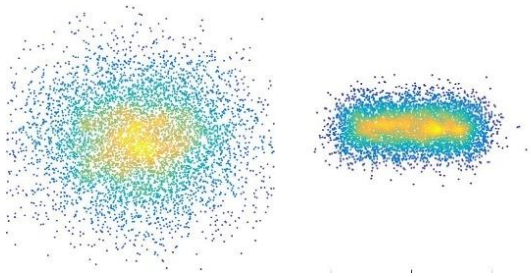


Fig. 2 4ASK between -6dB and 4dB

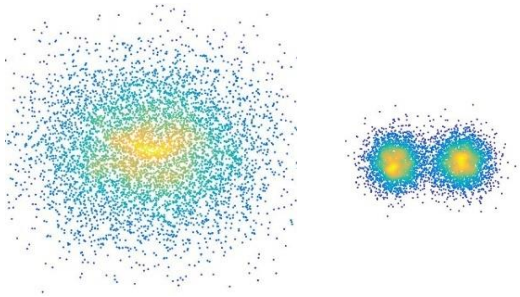


Fig. 3 BPSK between -6dB and 4dB

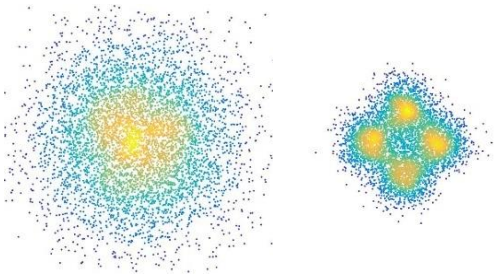


Fig. 4 QPSK between -6dB and 4dB

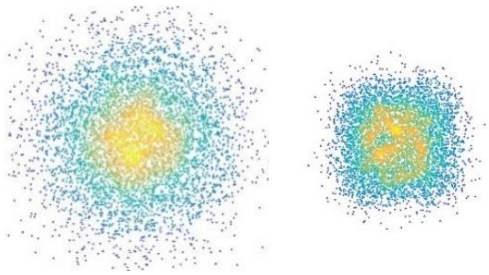


Fig. 5 16QAM between -6dB and 4dB

The gradient descent algorithm is the Stochastic Gradient Descent (SGD). This algorithm can ensure that every operation can converge in the direction of gradient descent. The correct rate after training is shown in the figure below.

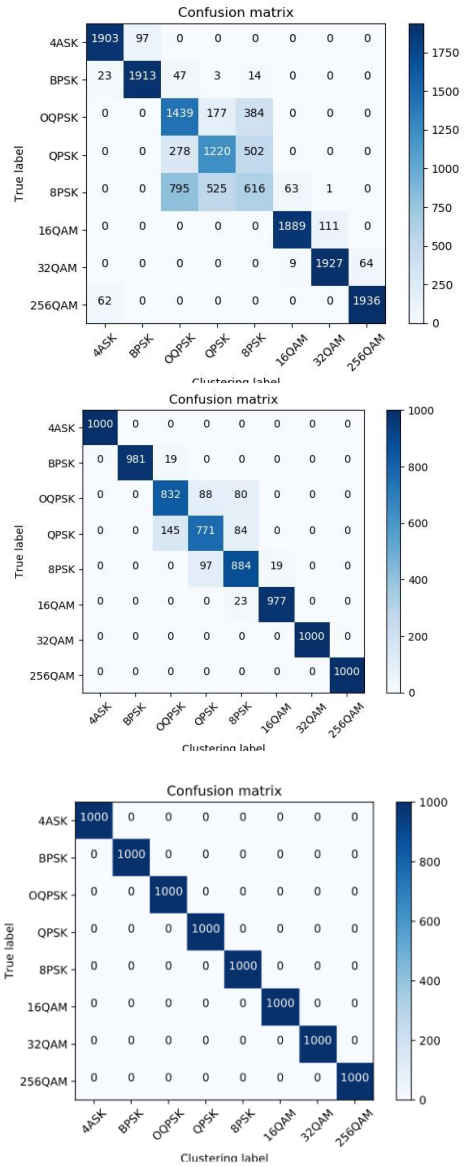


Fig. 7 some confusion matrix among -6dB, 0dB and 6dB SNR

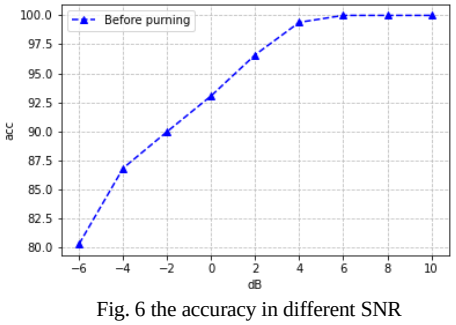


Fig. 6 the accuracy in different SNR

From the above accuracy we can see that the image effect is very good, in the case of low SNR still has 80% recognition

accuracy and high SNR has 100% accuracy. The next figure shows confusion matrix from -6dB to 10dB SNR.

Although in the previous chapters, we introduced the powerful ability of CSI image recognition in detail, the size of the network and the amount of computation still are relatively large. In this section, we mainly introduce a technology of network pruning, APoZ to prune the network and a formula for floating-point operations (FLOPs). However, pruning the network may worsen the accuracy of modulation recognition. Based on controlling the accuracy within a certain range, pruning the network is the significant topic we will introduce next.

A. *The Algorithm of APoZ*

This algorithm proposed by Hu *et al*. [14] measures each channel's output activation sparsity after ReLU mapping. The importance score of each filter can be calculated as follows:

$$APoZ(O_c^{(m)}) = \frac{\sum_a^N \sum_b^M f(O_{(c,b)}^m(a) = 0)}{MN}, \quad (2)$$

where $O_c^{(m)}$ denotes the *c*-th convolution filter of the *m*-th layer, $O_{(c,b)}^m(a)$ refers to the a validation image output corresponding b neuron in the *c*-th filter of the *m*-th layer. M is the dimension of the output $O_c^{(m)}$, and N represents the number of validation data. Function $f(\cdot)$ is an unit step function, where it equals 1 if true and equals 0 if false. The filter with higher APoZ reaches a certain threshold value, we stipulate that the filter has less activated, which has little effect on the promotion of the whole network and will be pruned. In this experiment, we set the threshold is 0.5, which means we will prune half the convolution filters.

B. *The Algorithm of FLOPs*

To measure the amount of calculation, we introduce a float point operations (FLOPs) concept. The FLOPs is a more accurate measure than measuring the instructions per second. The VTCNN2 model is primarily comprised of a convolutional layer and a fully connected layer. The FLOPs for the convolutional layer can be calculated with the following equation [12]:

$$FLOPs = 2HW(C_{in}K^2 + 1)C_{out} \quad (3)$$

where H, W and C_{in} are the height, width, and the number of the input feature map. K and C_{out} represent the kernel width, and the number of output channels. For the fully connected layer, the FLOPs can be computed as follows:

$$FLOPs = (2M - 1)N \quad (4)$$

where M is the input shape and N denotes output shape. Now we have the formula of network pruning and the algorithm of calculating floating-point numbers. Next, we show the flow of the hold algorithm.

Algorithm 1 Convolution layer pruning based on APoZ

Input: A network with full of Convolution filter;
Output: Convolution layer with pruned AlexNet Model;
Initialize: CNN parameter fixed;
1: **for** *b=1, 2, …, M* **do**
2: **for** *a=1, 2, …, N* **do**
3: Get image parameter updated from *m*-th layer
4: APoZ calculate the parameter and get a value *v*
5: **if** *v* > threshold we set
6: the convolution filter we prune
7: **end for**
8: **end for**
9: make a slight adjustment (retrain a small epoch)
10: **return** a pruned network

C. *Experiment*

We use the APoZ algorithm to prune trained AlexNet network, the next figure will show the accuracy and the confusion matrix in different SNR.

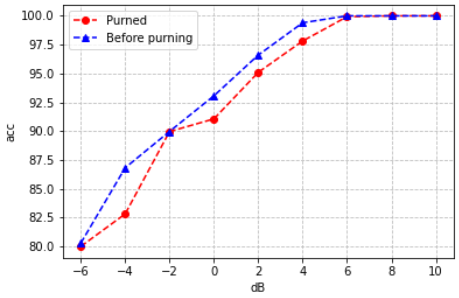
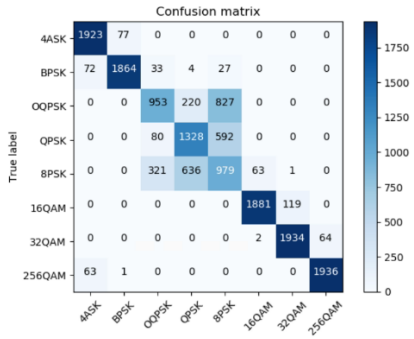
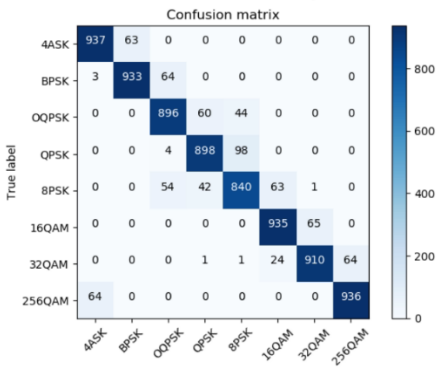


Fig. 8 the accuracy of prune network in different SNR





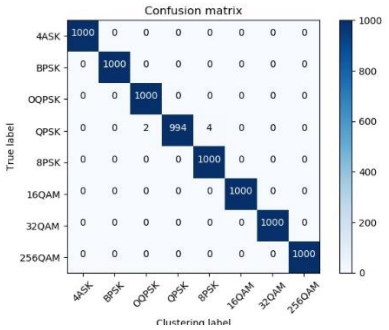


Fig. 9 confusion matrix among -6dB, 0dB and 6dB SNR

The below table will show that the bias between pruned and before pruning network.

TABLE II. COMPARING BETWEEN PRUNED AND BEFORE PRUNING

SNR	***Pruned***	***Before pruning***	***Bias***
-6	79.99%	80.27%	-0.28%
-4	82.81%	86.81%	-4%
-2	89.96%	89.97%	-0.01%
0	91.06%	93.06%	-2%
2	95.06%	96.6%	-1.54%
4	97.82%	99.4%	-1.58%
6	99.93%	100%	-0.07%
8	100%	100%	0%
10	100%	100%	0%

We can see that the pruning of the network does affect the accuracy, but most of the bias can be maintained at about 2%, which can be acceptable. The reason why we can keep the stable accuracy is that we only prune the convolution kernel which has little effect on the result in the process of clipping, a little effect on improving the accuracy, but a large floating-point operation. In other words, we prune the redundant convolution core, reducing the network size and floating-point operations. The next table will show the number of calculating floating-point and the size of AlexNet. In this table, we don't use time to measure the different before and after pruning, mainly considering that different time, place or the temperature of hardware equipment may affect the accuracy of the experiment as un-controllable factors. This experiment mainly compares the between floating-point computation and the size of the gene-rated network.

TABLE III. COMPARING BETWEEN FLOPS AND SIZE

	FLOPs	***Size***
Purned	116,609,102	455,636KB
Before purning	58,304,521	286,299KB
Bias	58,304,521	169,377KB

through the pruning network, the floating point operation of convolution layer is reduced to half of the original, and the size of network is reduced 37.16% of the original. APoZ algorithm is "slimming" for our network and it still can achieve good recognition effect.

V. CONCLUSION

In this paper, we introduced the APoZ algorithm to prune the convolution filter of network, and we compare the accuracy between before and after pruning network. Although this algorithm will slightly reduce the accuracy by 1-2% of modulation classification, the floating-point computing is reduced by 50% and network size is reduced by 37.16%. And the stability of the algorithm is also proved. There is still more space for improvement in our experiments because it is mainly in a stationary SNR for classification, it is worth to think about that training network model under the arbitrary SNR in the future.

VI. REFERENCES

[1] N. Lay, A. Polydoros. "Modulation classification of signals in unknown ISI environments," Proceedings of MILCOM '95. vol. 1, pp. 170-174, November. 1955.

[2] N. Kato, Z. Fadlullah, B. Mao, et al. "The Deep Learning Vision for Heterogeneous Network Traffic Control: Proposal, Challenges, and Future Perspective," IEEE Wireless Communications, 2017. vol. 24, pp. 146 – 153, June 2017.

[3] Z. Fadlullah, F. Tang, B. Mao, et al. "On Intelligent Traffic Control For Large Scale Heterogeneous Networks: A Value Matrix Based Deep Learning Approach," IEEE Communications Letters, 2018. vol. 22, pp. 2479 – 2482. December 2018.

[4] H. Huang, S. Huo, G. Gui, et al. "Deep Learning for Physical-Layer 5G Wireless Techniques: Opportunities, Challenges and Solutions," IEEE Wireless Communications, 2019. vol 27, pp. 214-222. February 2020.

[5] H. Huang , P. Yang , J. Yang, et al. "Fast Beamforming Design via Deep Learning," IEEE Transactions on Vehicular Technology, 2019, vol 69, pp. 1065-1069. October 2019.

[6] Hou, Changbo, Xiao Zhang, and Xiang Chen. "Electromagnetic signal feature fusion and recognition based on multi-modal deep learning." International Journal of Performability Engineering, vol. 16, no. 6, pp. 941-949, June 2020

[7] T. O'Shea, T. Roy, T. Clancy. "Over the Air Deep Learning Based Radio Signal Classification". IEEE Journal of Selected Topics in Signal Processing, vol. 12, pp. 168-179. January 2018.

[8] Y. Wang, M. Liu, J. Yang, and G. Gui, "Data-Driven Deep Learning for Automatic Modulation Recognition in Cognitive Radios," IEEE Trans. Veh. Technol., vol. 68, no. 4, pp. 4074-4077, Apr. 2019.

[9] H. Hu, R. Peng, Y. Tai, and C. Tang, "Network trimming: A datadriven neuron pruning approach towards efficient deep architectures," arXiv preprint arXiv:1607.03250, 2016.

[10] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient ConvNets," Proc. Int. Conf. Learn Representations, 2017.

[11] Y. Lin, Y. Tu, Z. Dou, et al. "Contour Stella Image and Deep Learning for Signal Recognition in the Physical Layer," IEEE Transactions on Cognitive Communications and Networking, 2020, vol. 7, pp. 34-36. September 2020.

[12] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz,"Pruning Convolutional Neural Networks for Resource Efficient Inference," Proc. Int. Conf. Learn Representations, 2017.

[13] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," IEEE Trans. Cogn. Commun. Netw., vol. 3, no. 4, pp. 563–575, Dec. 2017.

[14] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," Proc. Natl. Acad. Sci., vol. 114, no. 13, pp.3521-3526, Mar. 2017.

[15] T. O'shea, N. West. "Radio machine learning dataset generation with gnu radio,"Proceedings of the GNU Radio Conference. 2016, vol. 1, September 2016.

[16] Y. Shi, K. Davaslioglu, Y. Sagduyu, et al. "Deep Learning for RF Signal Classification in Unknown and Dynamic Spectrum Environments," 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), November 2019, pp. 1-10, doi: 10.1109/DySPAN.2019.8935684.

[17] Y. Lin, Y. Tu, Z. Dou. "An improved neural network pruning technology for automatic modulation classification in edge devices," IEEE Transactions on Vehicular Technology, vol 69, pp. 5703-5706, March 2020.

[18] S. Peng, H. Jiang, H. Wang, et al. "Modulation classification based on signal constellation diagrams and deep learning," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, pp. 718-727, July 2018.

[19] Y. Lin, Y. Tu, Z. Dou, et al. "The application of deep learning in communication signal modulation recognition," 2017 IEEE/CIC International Conference on Communications in China (ICCC). October 2017, pp. 1-5, doi: 10.1109/ICCChina.2017.8330488.

[20] F. Meng, P. Chen, L. Wu. "Automatic modulation classification: A deep learning enabled approach," IEEE Transactions on Vehicular Technology, 2018, vol. 67, no. 11, pp. 10760-10772.

[21] Y. Yang, X. Zhang. A Modified Method for Digital Modulation Recognition based on Instantaneous Signal Features[C]. 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE). IEEE, 2019, pp. 1351-1354. October, 2019.

[22] Y. Wang , J. Yang, M. Liu , et al. LightAMC: Lightweight Automatic Modulation Classification via Deep Learning and Compressive Sensing[J]. IEEE Transactions on Vehicular Technology, 2020, vol. 69, no. 3, pp. 3491-3495.