IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 6, 2019, accepted April 20, 2019, date of publication April 30, 2019, date of current version May 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2913945

Deep Neural Network Compression Technique Towards Efficient Digital Signal Modulation Recognition in Edge Device

YA TU AND YUN LIN

College of Information and Communication Engineering, Harbin Engineering University, Harbin 15001, China

Corresponding author: Yun Lin (linyun_phd@hrbeu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771154, and in part by the Fundamental Research Funds for the Central Universities under Grant HEU-CFG201830 and Grant GK2080260148.

ABSTRACT Digital signal modulation recognition is meaningful for military application and civilian application. In the non-cooperation communication scenario, digital signal modulation recognition will help people identify communication target and have better management over them. In order to the classification accuracy, deep learning is widely used to complete this task. However, current papers have not considered the deployment of deep learning in compute capability and storage limited edge equipment. In this paper, we utilize neural network pruning techniques to reduce the convolution parameters and floating point operations per second (FLOPs), which will pave a wide way to deploy signal classification convolution neural network (CNN) in edge equipment. We set the Average Percentage of Zeros (APoZ) criterion for convolution layers. Compared to original CNN, the experiment result shows that light CNN convolution layer could use only 1.5%∼5% parameter and 33%∼35% time without losing significant accuracy.

INDEX TERMS Modulation classification, deep neural network, network prune, edge device.

I. INTRODUCTION

With the increasing demand for already congested wireless bandwidth of radio spectrum, it is important to find a smart way to spectral allocation and interference mitigation [1], [2].To achieve efficient data transmission and avoid unnecessary signal interference, a lots of modulation techniques, such as amplitude shift keying (ASK), frequency shift keying (FSK), phase shift keying (PSK), Quadrature Amplitude Modulation (QAM) has been used to encode data on different carrier frequencies [3]. Therefore, effective modulation classification method has attracted many researchers' attention. In traditional way, people will consider expert features, such as maximum power spectral density, standard deviations amplitude, frequency, phase, and variance of the zero-crossing [4]. Reference [5] proposes a subtle feature extraction and recognition algorithm for radiation source individual signals based on multidimensional hybrid features. Reference [6], [7] takes modulation classification in the satellite communication scenario and Waveform Design. Reference [8] uses dimensionality reduction and classifier in machine learning to identify Radio frequency (RF) fingerprint in wireless device. Reference [9] uses powerful kernel-based learning to process statistical signal. Reference [10] applies Hilbert transform and principal component analysis to generate the RF fingerprint of Device-to-device (D2D) device, and use SVM and CV-SVM to be the classifier, simulation results proves that the proposed method is effective for D2D device recognition. Reference [8] uses a combination of robust principal component analysis (RPCA) and random forests and improves wireless device authentication security protection. Reference [11] presents boosting algorithm as an ensemble frame to achieve a higher accuracy than a single classifier. The results show that the Fisherface algorithm is effective in reducing the feature dimension of digital signal in automatic modulation recognition. Based on information entropy features and Dempster-Shafer evidence theory, [12] proposes a novel automatic modulation recognition methods to obtain a higher recognition result in low SNR. However, those simple features that are considered in ideal environment, is heavily depended on prior knowledge, human experts and leads to limited classifier performance in

The associate editor coordinating the review of this manuscript and approving it for publication was Guan Gui.

noisy environment. Therefore, it is imperative to explore a signal feature self-extraction method to confront with increasingly complexed electromagnetic environment [13].

Deep learning (DL) is part of a broader family of machine learning methods due to its learning data representations, which use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Nowadays, the world has witnessed a lot of successful DL application in many fields, such as Computer Vision, Speech Recognition, and Nature Language Processing. In 2012, Krizhevsky *et al.* [14] firstly proposed a large, deep convolutional neural network to classify the 1.2 million high-resolution images and get the top in the competition of ImageNet large scale visual recognition challenge (ILSVRC). In 2016, He *et al.* [15] presented a residual learning framework to ease deeper neural network training and the residual nets framework won top place on the ILSVRC 2015 classification task. In 2018, Zhou *et al.* [16] proposes a new framework termed transfer hashing with privileged information (THPI) to address this so-called data sparsity issue in hashing.

With the development of DL, many researchers also exploited DL potential in signal processing field. O'Shea1 *et al.* [1] firstly publishes an open access evaluation dataset consists of modulated signal sampled from GNU Radio. Wang *et al.* [3] adopt dropout in CNN instead of pooling operation to achieve higher recognition accuracy. Gui *et al.* [17] propose a novel and effective DL-aided NOMA system, in which several NOMA users with random deployment are served by one base station. Peng *et al.* [4] firstly use high-resolution image, Signal Constellation Diagrams, to perform signal modulation classification and they has showed promising result due to great learning power of AlexNet and GoogLeNet. In our related work [18], we use dots density to recover deep level signal statistical information in Signal Constellation Diagrams at lower signal-to-noise ratio (SNR) and create Contour Stellar Image (CSI), our method achieved average 7% $\sim$ 8% higher accuracy at lower SNR compared to [4]. Further, we also use CSI dataset to conduct data augmentation on GANs [18] and semi-supervised learning on GANs [19]. In, Liu *et al.* [20] propose a multi-objective resource allocation (MORA) scheme for the UAV-assisted Het-IoT.

Despite of the great success in DL, a typical deep model is hard to be deployed on resource-constrained devices due to the limitation of compute capability and storage. It is a widely-recognized that most of deep neural network are over-parameterization [21]. To solve this, network prune technique has been identified as an effective way to combat with high computational cost and high memory cost issues. Current network pruning technique mainly includes three step:

1) Training from Scratch: the deep neural network will find the way to extract data feature from dataset and adjust neural weight to get a better result.
2) Network Pruning: This is core step of deep neural network. According to the different pruning granularity, we could choose a weight connection or a convolution filter. Then we will evaluate whether it is an important component of accuracy due to different evaluation criterion. Different method has different evaluation criterion.
3) Finetune: during network slimming processing, the pruning operation will unavoidable undermine influence network accuracy. At this time, fine-tuning is a necessary step to recover the generalization ability.

Nowadays, mainstream pruning processing [22], [23] can be depicted in Fig. 1:

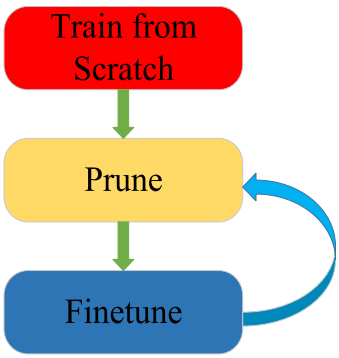

FIGURE 1. Current network pruning process.

The rest of this paper is organized as follows. We first give a briefly introduce to our dataset in section II. Section III at-tempts to explain how our pruned network will behavior. Section IV gives the details of our prune criterion on convolution layer channel level. Latter, we will post the network slimming evaluation criterion and experiment result in Section V. Lastly, we give the conclusion about our original intention.

II. CONTOUR STELLAR IMAGE

We introduce the conception of CSI in our related work [18]. Inspired by [4], we believe CNN could perform better if deeper level feature in Constellation Diagrams (CD). In CD, a dot in the image derives from a sample point in signal wave and CD carries the amplitude and the phase information about this sample point.

In CSI, we consider dot density. Sliding on CD, CSI will utilize a square window function to count how many dots are in the window in different area. Then we will use (2) to calculate normalized dots density:

$$\rho(i,j) = \frac{\sum_{i=m_1}^{m_2}\sum_{j=n_1}^{n_2} dots(i,j)}{\sum_{m_1=W_0}^{W_1}\sum_{n_1=H_0}^{H_1}\sum_{i=m_1}^{m_2}\sum_{j=n_1}^{n_2} dots(i,j)} \quad (1)$$

where m_1, n_1 is top-left corner of square windows function currently coordinate, m_2, n_2 is bottom-right corner of square windows function currently coordinate, W_0, H_0 denotes for CD top-left corner coordinate, W_1, H_1 denotes for bottom-right corner coordinate of CD. $\rho(i,j)$ is Relatively Point

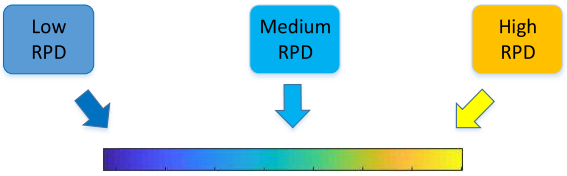

FIGURE 2. Transfer from CD to CSI color bar.

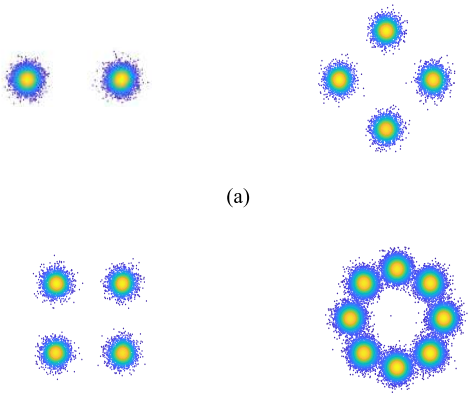

FIGURE 3. (a) CSI of BPSK at SNR equals to 14dB and CSI of QPSK at SNR equals to 14dB (b) CSI of OQPSK at SNR equals to 14dB and CSI of 8PSK at SNR equals to 14dB.

Destiny (RPD), $dots(i, j)$ are:

$$dots\,(i,j) = \begin{cases} 1 & if\ (i,j)\ exists\ a\ dot \\ 0 & if\ (i,j)\ exists\ no\ dot \end{cases} \tag{2}$$

After calculating all RPD on the CD, we will map RPD into color, the color bar is as followed:

From the perspective of signal processing, CSI could indicate deep statistics information. At higher SNR, both CD and CSI could reveal modulation signal statistical information, such as Gaussian noise, Non-coherent single frequency interference, Phase noise, Amplifier compression in the signal waveform. At lower SNR, due to heavily perturbation from noise, CD modulation signal statistical information will be disguised and CSI will use dot density to recover statistical information by different color mapping from RPD.

III. NETWORK PRUNE BASED ON APoZ CRITERION

There exists some heuristic criteria to score the importance of each filter in the literature. In this paper, we apply APoZ to convolution layer.

Convolution Layer calculation time master the most of forwards propagation time, due to numerous matrix most of multiplication operation compared to fully connected layer. After pruning convolution layer, the CNN will run faster.

In [24], APoZ was proposed to calculate the sparsity of each channel in output activations as its importance score:

$$\mathrm{APoZ}\left(O_C^{(i)}\right) = \frac{\sum_k^N \sum_l^M \delta(O_C^{(i)}(\mathrm{k}))}{MN} \tag{3}$$

where $O_C^{(i)}$ denotes the output of *c-th* channel neuron in *i-th* layer, $O_C^{(i)}(k)$ denotes the corresponding k validation image output of *c-th* channel n in *i*-th layer. δ is Dirac delta function. M means the dimension of output channel of $O_C^{(i)}$, and N is the total number of validation examples.

Now, given a set of m training examples $\{\widehat{x_i}, \widehat{y_i}\}$, and validation set $\{\overline{x_i}, \overline{y_i}\}$, we denotes the i-th neuron in l-th layer activation to $a_{i,j}$, the convolution filter level pruning algorithm based on APoZ can be defined as the following:

For convolution filter level APoZ pruning, the process can be also depicted in Fig. 4:

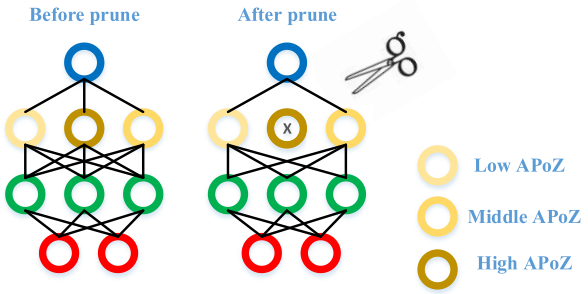

FIGURE 4. Convolution filter level pruning base on APoZ.

We take AlexNet at 6dB and 0dB as an example and consider those filter channels with APoZ higher than 0.9 to be the less activated convolution filters. The less activated convolution filters proportion can be depicted in Fig. 5:

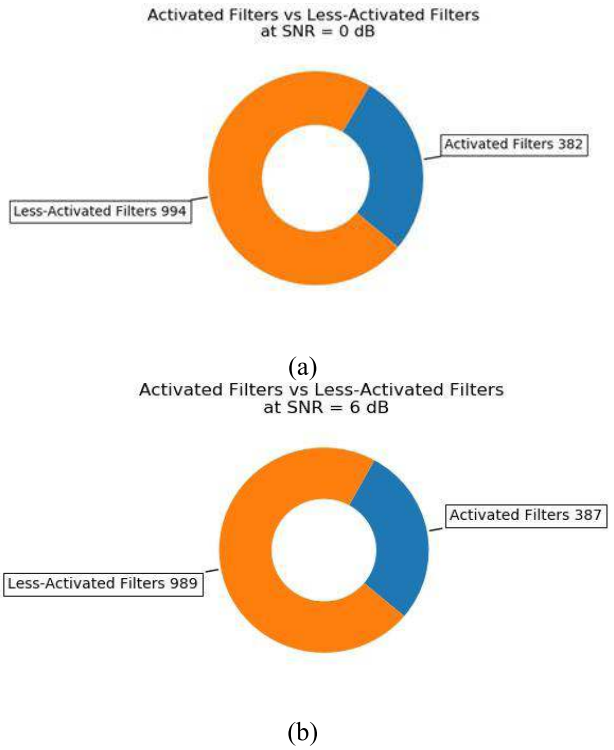

FIGURE 5. (a)Activated filters and less activated filters in convolution layer at SNR = 6dB (b) Activated filters and less activated filters in convolution layer at SNR = 0dB.

Algorithm 1 Convolution Layer Channel Level Pruning Base on APoZ

Input: Select i-th channel in l-th layer;
Training set $\{\widehat{X}_i, \widehat{Y}_i\}$; The number of Channel C; Activation Theshold H; The neuron unit a; The tolerate accuracy loss P; Validation examples numbers N; Output feature map dimension M
Output: Convolution Layer Pruned CNN Model

initialize: CNN Parameter fixed;
$APoZ(i, l) = 0$
for $t = 1, 2, \ldots, N$ **do**
  for $j = 1, 2, \ldots, M$ **do**
    Feed CNN with t-th validation examples
    Get the output of i-th channel
    if $a_{i,j} = 0$ **then**
      $APoZ(i, l) = APoZ(i, l) + 1$
    else
      Continue
    end if
  end for
end for
$APoZ(i, l) = \frac{APoZ(i,l)}{M \times N}$
if $APoZ(i, l) > H$ **then**
  Remove Channel (i, l) from CNN
else
  Continue
end if
Fine tune CNN with $\{\widehat{x}_i, \widehat{y}_i\}$

From Fig. 5, we could see (a) and (b) that less activated filters makes up of 71.9 % of convolution filter in AlexNet at 6dB and 72.2 % of convolution filter in AlexNet at 0dB, it shows that original AlexNet trained on CSI at 6dB and 0dB obtains a lot of redundancy parameters in convolution layer, we may prune them.

IV. EXPERIMENT

In this section, we will firstly introduce the experiment implement detail, secondly, we will evaluate network slimming result from FLOPs and storage perspective. Lastly, we deploy original CNN and pruned CNN into different device to verify our work in reality situation.

A. IMPLEMENT DETAIL

Thanks to Keras user friendliness, modularity and easy extensibility feature, in our experiment, we choose AlexNet, which is implemented in Keras 2.2.4, to conduct signal modulation classification and network slimming. AlexNet consist of 5 layers' convolution layers and 2 fully connected layers, with 60 million parameters, and 727 million FLOPs. We slightly modify softmax layers, change 1000 neurons to 8 neurons to satisfied classification task while keeping the rest layers unchanged. In addition to the changes of model architecture, we still choose the default parameter according to [14]. After that modification, AlexNet parameters shrinks to 21 million parameters and FLOPs declines to 43.20 million.

For dataset, we choose CSI dataset ranges from −6dB to 6dB, with 8000 training set, 8000 validation set and 8000 test set per SNR. Modulation signal category includes: 4ASK, BSPK, QPSK, OQPSK, 8PSK, 16QAM, 32QAM, 64QAM.

For Validation device, we choose NVIDIA Jetson TX2 Module to simulate computability limited environment.

In this paper, we use the concept of APoZ to measure the importance of convolution filters. We experimented different ways to prune the convolution filter according to the APoZ measurements. It was found that pruning specified convolution filter or neurons too many time in one step will cause serious and irreversible damage to accuracy. However, an iterative route has been adopted to slim the network. The strategy we use is firstly keep pruning Conv1 then pruning Conv2 and Conv3, Conv4, Conv5 repeatedly by APoZ criterion until we discover the loss of accuracy after fine tune. Having done many experiment, we found the best APoZ threshold is:

$$\mathrm{T} = \overline{\mathrm{APoZ}} + \mathrm{std(APoZ)} \times \mathrm{C} \tag{4}$$

where $\overline{\mathrm{APoZ}}$ denotes the mean value of APoZ, std(APoZ) denotes the standard deviation value of APoZ, C denotes a hyper parameter to adjust reject threshold, T is the threshold. In this paper, we set C to 1 and those neurons which get APoZ higher than T will be pruned.

For dataset, [4] uses exponential decay model to convert CD into 3-channel image, named enhanced CD, and we choose dot density to convert CD into 3-channel image, named CSI. To prove CSI method behave better than [4], in this paper, we pick the same CNN architecture, AlexNet and GoogLeNet [25], and the same category modulation signal to [4], which is 4ASK, BSPK, QPSK, OQPSK, 8PSK, 16QAM, 32QAM, 64QAM. We just use 1/10 training set and conduct the comparison experiment.

B. ACCURACY COMPARISON

In this part, we will evaluate accuracy pruned AlexNet and original AlexNet. In traditional view, there always exists tradeoff between accuracy between compute complexity and we will figure out how severely network pruning will damage the accuracy. Since our test set is a balance dataset, we directly use average accuracy to be our criterion. To compute average accuracy, we use:

$$\mathrm{Acc} = \frac{N_{correct}}{N} \times 100\% \tag{5}$$

where $N_{correct}$ is the number of correctly classified test samples, N is the total number of test samples, Acc is average accuracy. Then we give accuracy comparison between original AlexNet and Pruned AlexNet and [4] method in Fig. 6:

According to [26], DNN with fewer neurons may lost capacity to present the data accurately and DNN with more neurons will not generalize well. In other word, there exists

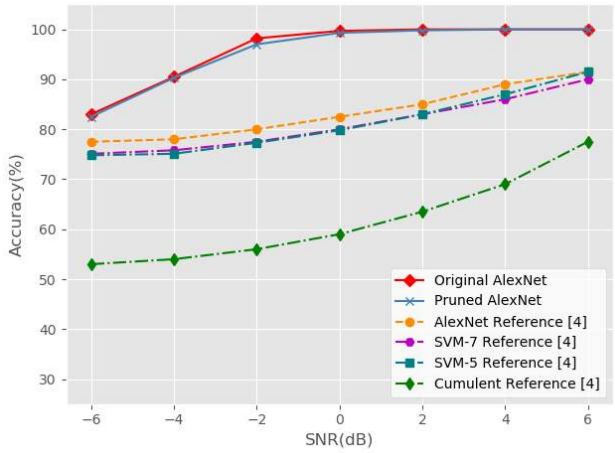

FIGURE 6. Average accuracy comparison between Pruned AlexNet and Original AlexNet and [4] method.

trade-off between train error and the network complexity. It is a little surprise to find that pruned AlexNet performance worse than original AlexNet just about 0.2 % $\sim$ 1.2%. In essence, the pruning process based on APoZ has served as regularity, which will pick out those filters or neurons does not have generalizability weights. What is more, CNN and enhanced CD achieves higher classification compared to shallow machine learning method with expert feature, since CNN will extract the feature that can contributes more to classification accuracy. CNN and CSI method surpasses CNN and enhanced CD method in precision, which indicates CSI is a better choice for CNN based modulation classification dataset.

We also give the confusion matrix for pruned AlexNet and original AlexNet at 0dB:

Compared with (a) and (b), we could find that the accuracy decreases in OQPSK and 32QAM after pruned. From Fig. 7, we could see that accuracy start to decline due to the different category signal CSI feature gets close to others, AlexNet will become more sensitive to the change of architecture.

C. FLOPS COMPRESS

Flops is one of the most prevalent ways to estimate the amount of calculation in DNN model [23]. Flops will take amount of calculation in convolution layer and fully connected layer into consideration. This index will point how well the pruned model will behave in computability limited device. In this part, we inspect how many Flops have been reduced. To compute parameters compression ratio, we use:

$$\mathrm{FCR} = \frac{F_{original}}{F_{pruned}} \tag{6}$$

where $F_{original}$ denotes the amount of FLOPs belongs to original AlexNet, and F_{pruned} denotes the amount of FLOPs belongs to pruned AlexNet.

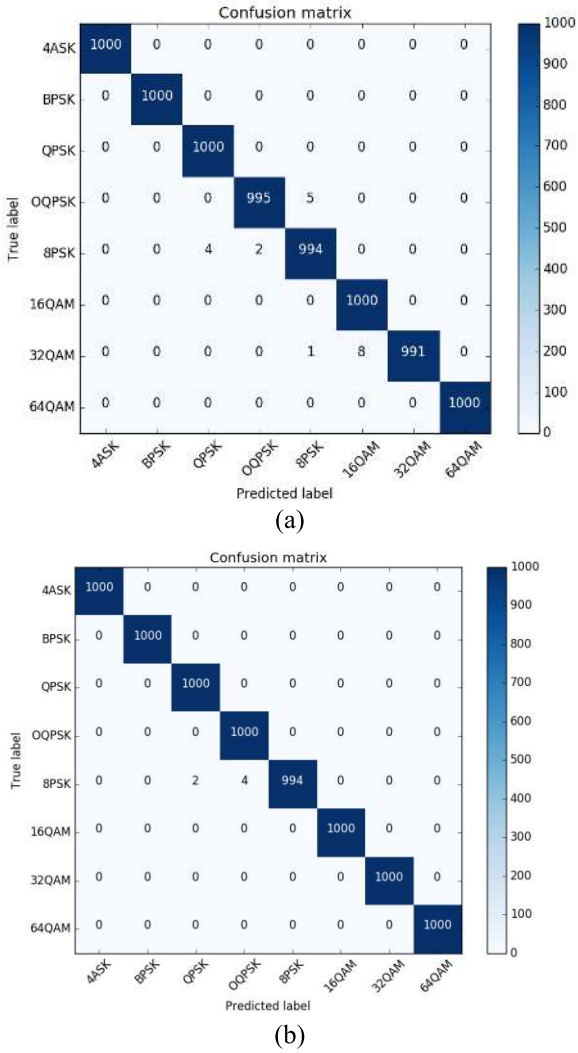

FIGURE 7. (a) Pruned AlexNet's confusion matrix at 0dB (b) Original AlexNet's confusion matrix at 0dB.

We give the result about FCR in Fig. 8:

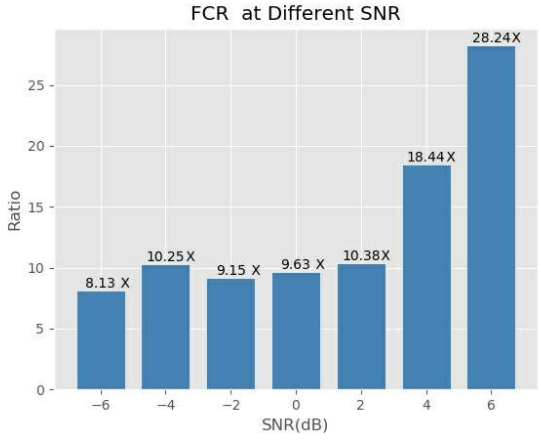

FIGURE 8. FCR result.

The average FCR is 13.76 $\times$, which shows our method has greatly reduced the model compute complex. At higher SNR, the image is easily to distinguish, so many feature extracted

by convolution filter are redundancy and there exists more space to compress the model.

D. CONVOLUTION LAYER PARAMETERS COMPRESS

Parameters amount will influence the storage and memory cost. This index should be taken into account for storage and memory limited devices. In this part, we will inspect how many parameters have been compressed. To compute parameters compression ratio, we use:

$$\mathrm{PCR} = \frac{P_{original}}{P_{pruned}} \tag{7}$$

where $P_{original}$ denotes the amount of convolution parameters belongs to original AlexNet, and P_{pruned} denotes the amount of parameters belongs to pruned AlexNet.

We give the result about PCR in Fig. 9:

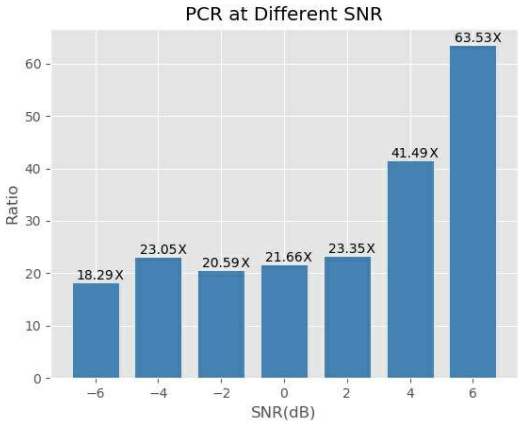

FIGURE 9. PCR result.

The average PCR is 30.28 ×, which shows our method has greatly reduced the model parameter amount. With the same tendency to FCR. At higher SNR, the image is easily to distinguish, so we just need few channels to finish classification and there exists more space to compress the model.

E. THE TRANSFORMATION OF CONVOLUTION LAYER AFTER PRUNE

In Fig. 5, we could see there exists a large amount of less activated convolution filter with (APoZ $>$ 0.9). To prove the method we adopt is competent in sliming AlexNet, we explore how convolution layer activated redundancy information change. We still deem those neurons which has APoZ $>$ 0.9 to be less activated and obtain following comparison result:

Seen from Fig. 10, we could easily see that less activated convolution filter (APoZ $>$ 0.9) makes up of most proportion in original AlexNet, which indicates original AlexNet great potential in network compression by APoZ. Secondly, we could also notice APoZ has pushed convolution layer to be denser which tell us APoZ success to remove redundancy from CNN.

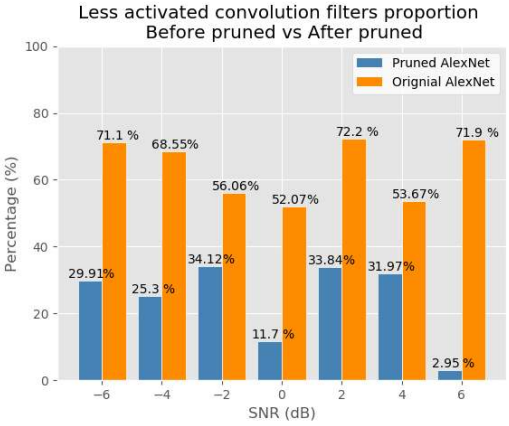

FIGURE 10. Less activated convolution filters proportion before pruned vs after pruned.

FIGURE 11. NVIDIA Jetson TX2 module.

F. DEVICE VALIDATION

In this part, we will start to deploy original AlexNet and pruned AlexNet into reality device. The device we used in this part is NVIDIA Jetson TX2 and we find out how our network slimming will effect signal classification task in compute limited capability device.

NVIDIA Jetson TX2 Module is an embedded AI computing device. It features a variety of standard hardware interfaces that make it easy to integrate it into a wide range of products and form factors. NVIDIA Jetson TX2 Module technical specification consists of NVIDIA Pascal™, 256 CUDA cores for GPU, HMP Dual Denver 2/2 MB L2 +, Quad ARM® A57/2 MB L2 for CPU and 8 GB 128 bit LPDDR4.

To measure the different performance of the pruned AlexNet and original AlexNet. We firstly read 1000 test data sample per SNR into RAM, then we estimate the forward propagation time by the time they need to predict all the images in RAM. To simulate the computability limited environment and eliminate random factor, we only set processor to be CPU and get 10 mean of forward propagation time, we get the comparison result as follow:

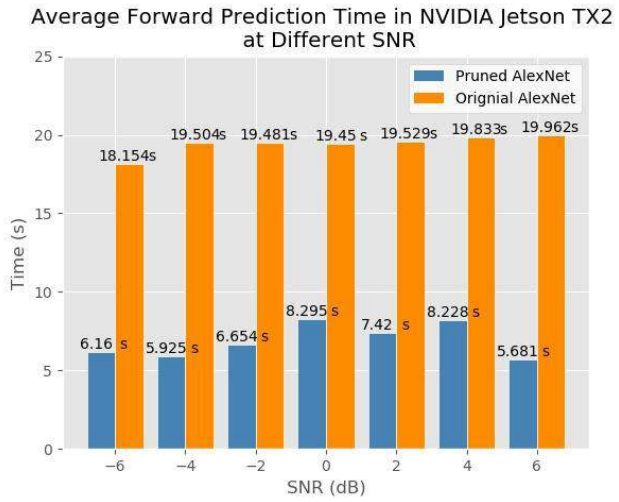

FIGURE 12. Prediction time comparison for 1000 samples.

It can been seen that pruned AlexNet at all SNR get faster about 3× in predicting samples in NVIDIA Jetson TX2 Module, since we cut down the FLOPs amount in convolution layer. We could also inspect that the predict time are equal at every SNR, even 6dB pruned AlexNet' FLOPs are much smaller to other SNR AlexNet' FLOPs. We believe it is caused by hardware communication delay.

V. CONCLUSION

Nowadays, more and more deep learning model have been applied into communication field. There is an also tendency to deploy powerful DNN model, such as AlexNet, into compute and storage limited device to confront the increasing complex communication environment. However, some powerful DNN model that is designed for computer vision may cause over-parameterized for edge equipment. In this paper, we choose APoZ criterion to prune the network. Result shows that light CNN convolution layer could use only 1.5% $\sim$ 5% parameter and 33% $\sim$ 35% time without losing accuracy more than 1.2%. However, we still lose some accuracy and we believe we could find a safe network prune technique to slim the network without losing accuracy and even get a better accuracy. In future, we will find a more effective criterion to conduct network prune.

REFERENCES

[1] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Proc. Int. Conf. Eng. Appl. Neural Netw.*, 2016, pp. 213–226.

[2] G. Ding, Q. Wu, L. Zhang, T. A. Tsiftsis, and Y. Yao, "An amateur drone surveillance system based on cognitive Internet of Things," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 29–35, Jan. 2018.

[3] 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. doi: 10.1109/TVT.2019.2900460.

[4] S. Peng *et al.*, "Modulation classification based on signal constellation diagrams and deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 718–727, Mar. 2019.

[5] J. Li, D. Bi, Y. Ying, K. Wei, and B. Zhang, "An improved algorithm for extracting subtle features of radiation source individual signals," *Electronics*, vol. 8, no. 2, p. 246, 2019.

[6] M. Jia, X. Liu, X. Gu, and Q. Guo, "Joint cooperative spectrum sensing and channel selection optimization for satellite communication systems based on cognitive radio," *Int. J. Satell. Commun. Netw.*, vol. 35, no. 2, pp. 139–150, 2017.

[7] M. Jia, Z. Yin, Q. Guo, G. Liu, and X. Gu, "Waveform design of zero head DFT spread spectral efficient frequency division multiplexing," *IEEE Access*, vol. 5, pp. 16944–16952, 2017.

[8] Y. Lin, X. Zhu, Z. Zheng, Z. Dou, and R. Zhou, "The individual identification method of wireless device based on dimensionality reduction and machine learning," *J. Supercomput.*, pp. 1–18, Dec. 2017.

[9] G. Ding, Q. Wu, Y.-D. Yao, J. Wang, and Y. Chen, "Kernel-based learning for statistical signal processing in cognitive radio networks: Theoretical foundations, example applications, and future directions," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 126–136, Jul. 2013.

[10] Z. Zhang, X. Guo, and Y. Lin, "Trust management method of D2D communication based on RF fingerprint identification," *IEEE Access*, vol. 6, pp. 66082–66087, 2018.

[11] T. Liu, Y. Guan, and Y. Lin, "Research on modulation recognition with ensemble learning," *EURASIP J. Wireless Commun. Netw.*, vol. 2017, no. 1, p. 179, 2017.

[12] H. Wang, L. Guo, Z. Dou, and Y. Lin, "A new method of cognitive signal recognition based on hybrid information entropy and D-S evidence theory," *Mobile Netw. Appl.*, vol. 23, no. 4, pp. 677–685, 2018.

[13] Y. Cao *et al.*, "Optimization or alignment: Secure primary transmission assisted by secondary networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 905–917, Apr. 2018.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 770–778.

[16] J. T. Zhou, H. Zhao, X. Peng, M. Fang, Z. Qin, and R. S. M. Goh, "Transfer hashing: From shallow to deep," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6191–6201, Dec. 2018.

[17] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.

[18] B. Tang, Y. Tu, Z. Zhang, and Y. Lin, "Digital signal modulation classification with data augmentation using generative adversarial nets in cognitive radio networks," *IEEE Access*, vol. 6, pp. 15713–15722, 2018.

[19] Y. Tu, Y. Lin, J. Wang, and J.-U. Kim, "Semi-supervised learning with generative adversarial networks on digital signal modulation classification," *Comput. Mater. Continua*, vol. 55, no. 2, pp. 243–254, 2018.

[20] M. Liu, J. Yang, and G. Gui, "DSF-NOMA: UAV-assisted emergency communication technology in a heterogeneous Internet of Things," *IEEE Internet Things J.*, to be published. doi: 10.1109/JIOT.2019.2903165.

[21] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell. (2018). "Rethinking the value of network pruning." [Online]. Available: https://arxiv.org/pdf/1810.05270.pdf

[22] J.-H. Luo, J. Wu, and W. Lin, "ThiNet: A filter level pruning method for deep neural network compression," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5058–5066.

[23] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1135–1143.

[24] H. Hu, R. Peng, Y.-W. Tai, and C.-K. Tang. (2016). "Network trimming: A data-driven neuron pruning approach towards efficient deep architectures." [Online]. Available: https://arxiv.org/abs/1607.03250

[25] C. Szegedy, W. Liu, Y. Jia, and P. Sermanet, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[26] L. Yann, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 605–698.