A Novel Approach to Structured Pruning of Neural Network for Designing Compact Audio-Visual Wake Word Spotting System

Haotian Wang†, Jun Du†*, Hengshun Zhou†, Heng Lu‡ and Yuhang Cao‡
† National Engineering Research Center of Speech and Language Information Processing
University of Science and Technology of China, Hefei, Anhui, China
E-mails: az1522702192@mail.ustc.edu.cn, jundu@ustc.edu.cn, zhhs@mail.ustc.edu.cn
‡ Ximalaya Inc., ShangHai, China
E-mails: bear.lu@ximalaya.com, adam.cao@ximalaya.com

***Abstract*—In this paper, we propose a novel approach to structured pruning of neural network. Firstly, we extend the original channel-level pruning from one-shot manner to iterative manner. Then we further employ the learning rate rewinding strategy in the lottery ticket hypothesis (LTH) to guide the channel-level pruning, yielding a new algorithm named channel-level pruning with learning rate rewinding (CPLR). Finally, we apply CPLR to prune the audio and video networks for designing compact audio-visual wake word spotting (AVWWS) system. Tested on MISP-2021 AVWWS database, the results show that the proposed CPLR approach performs better than either the channel-level pruning approach or LTH approach in term of both system performance and model efficiency. More interestingly, we observe that while the network parameters are greatly reduced by CPLR, the network generalization capability can be even better.**

Index Terms: structured pruning, channel-level pruning, learning rate rewinding, lottery ticket hypothesis, audio-visual wake word spotting

I. Introduction

Wake word spotting (WWS) can be regarded as a specific case of keyword spotting (KWS), which plays a very important role in human-computer interaction. The goal of the WWS task is to recognize a predefined wake-up word [1]. Typical WWS systems are based on audio modality [2, 3], one of the main challenges is that the performance of these systems usually declines a lot under noisy conditions [4–6]. In addition, due to the interference of signal transmission, WWS is still a challenging task under far-field conditions [7]. In the past few years, research efforts have been made in detection of audio wake-up words under noisy and far-field conditions, such as introducing speech enhancement [6, 7] and the novel training methods [8–10].

In [11–14], the authors demonstrate that visual information from video can improve the system performance based on clear and noisy audio signals. However, the introduction of visual information also greatly increases the total amount of parameters. For example in the audio-visual wake word spotting (AVWWS) system [15], the model size of proposed video network is much larger than that of the audio-only network, which will limit its application.

Accordingly, an effective pruning method for neural network is essential for the feasibility of audio-visual systems. Network pruning includes structured pruning and unstructured pruning [16, 17]. Structured pruning can not only improve the computational efficiency, but also effectively reduces the number of network parameters, which has attracted more and more attention [18, 19]. Filter pruning is one type of structure pruning which can be realized by Taylor expansion, geometric median (FPGM) and other methods [20–22]. Channel pruning using the batch normalization (BN) layers can obtain better results [23]. Moreover, the training and pruning strategy is also very important. The classic strategy is the "weights initialization - pruning - fine-tuning" procedure [24]. Then the authors proposed a dynamic pruning approach, which uses sparse training in convenient pruning operation [23]. In 2019, the lottery ticket hypothesis (LTH) [25] was proposed by integrating existing pruning and training modes. Some researchers have compared various strategies based on LTH and one effective scheme is based on learning rate rewinding [26].

So far, there are not many studies on the neural network pruning for AVWWS task [15]. One remarkable work is our previously proposed LTH-IF [27], namely LTH in an iterative fine-tuning manner. In this study, we develop a new structure pruning approach called Channel-level Pruning with Learning rate Rewinding (CPLR) that integrates the channel-level pruning method and the LTH strategy. Firstly, we improve the original channel-level pruning method from the one-shot pattern to an iterative pattern. Then we adopt the learning rate rewinding strategy to guide the channel-level pruning. CPLR can fully utilize the advantages of both channel-level and the LTH pruning. Finally, we apply the proposed pruning algorithm to the audio and video neural network architectures for WWS. Evaluated on MISP-2021 AVWWS database, the CPLR approach can yield consistent improvements in terms of both system performance and model efficiency over the LTH-IF approach.

*corresponding author

The remainder of this paper is organized as follows. Section II gives a review of related works on network slimming with their advantages and disadvantages. Section III elaborates our proposed CPLR approach and presents the innovation points. In Section IV we apply CPLR to the AVWWS system. In Section V experimental results and analysis are discussed. Finally, we conclude in Section VI.

II. Related work

In this section, we describe the two types of neural network pruning approaches related to our work, namely channel-level pruning and LTH-based pruning.

A. Channel-level neural network pruning

Channel-level pruning is a kind of structured pruning, which was first proposed in [23]. The authors chose the scale factor of BN layer γ as the indicators for two reasons. The one is structured pruning can reduce both the parameter number and the float point operations (FLOPS). The other is channel-level pruning scheme has certain flexibility that can be applied to most convolution neural networks (CNNs). The channel with larger γ is considered more important for the network and should be avoided from being pruned. In the training process, sparse training was firstly adopted for pre-training the network. Then an algorithm for pruned channels depending on the absolute value (ABS) of the scale factor γ of BN layer was developed. Finally, a smaller learning rate was set to fine-tune the pruned network. This method can achieve a 20% to 70% compression ratio on different networks.

However, the channel-level pruning method also has some disadvantages: (1) The channel-level pruning method only uses one round of training and one round of fine-tuning, the preset pruning rate must be very high to get a satisfactory compression ratio, making it a challenging task to fine-tune the network. (2) An excessive pruning rate may destroy the network structure, so the network can not achieve the performance of the original network with a high compression ratio. (3) The sparsity of channels is highly related to the results of the sparse-training process, it is difficult to find the appropriate default hyper-parameters.

B. LTH-based neural network pruning

LTH is a hypothesis that a randomly-initialized, dense neural network contains a sub-network (winning lottery ticket) which can match the test accuracy of the original network after training in isolation for at most the same number of iterations [25]. The original LTH strategy is based on weight pruning belonging to unstructured pruning. First, the network is trained to the early-stop point (the training epoch that reaches the lowest loss on the development set). Then the same parameters at the beginning of the training process are used to reset the network. Afterward, the network is retrained using a smaller learning rate. Finally, the above steps are repeated to reach the required compression ratio. In the recent work on LTH, the authors compare three training and adjustment strategies for LTH, namely weight rewinding strategy (lottery strategy), pruning and fine-tuning strategy, and learning rate rewinding strategy. The conclusion is the learning rate fine-tuning strategy is the best in most instances [26].

The LTH based method works well on some small-size networks [25], which can even achieve the compression ratio of more than 95%. But it still has following disadvantages: (1) The LTH algorithm works very well on small-size network, but it is difficult to find the wining lottery ticket in large-scale network. Some training strategy may be used to solve the problem above like using warmup strategy in the training process [25], but these methods also increase the training cost. (2) The original LTH strategy is based on weight pruning, and this pruning method can not reduce the parameters efficiently on normal hardware because although the pruned weights are set to zero, they still exist on the hardware as floats, taking up storage space [16, 17].

III. Channel-level pruning with learning rate rewinding

Algorithm 1 CPLR algorithm

1 : Pre-train the initial network to the early-stop point $f(x;\theta_0;\gamma_0)$ by using sparse-training.
2 : Set the pruning rate $k = 30\%$ per round.
3 : Create the mask m based on BN scale factor γ and the pruning rate k.
4 : Prune 30% channels globally using the mask, obtain the reinit network $f(x;\theta_0 \odot m;\gamma_0 \odot m)$.
5 : Use learning rate rewinding strategy to retrain the network to obtain network $f(x;\theta_1;\gamma_1)$.
6 : Repeat **2** to **5** until the desired weight sparsity ratio is reached.
7 : Fine-tune the pruned network using the same learning rate as the last retraining process, return fine-tuned network $f(x;\theta_N;\gamma_N)$.

In this section, we elaborate on the proposed CPLR algorithm and compare it with channel-level pruning and LTH strategy. The detailed process is illustrated in Fig. 1. Firstly, the network without pruning is trained with sparse training to the early stop point. Secondly, the pruning mask is created based on the scale factor γ in the BN layer, which is used to prune less important channels. To obtain a more compact construction, learning-rate rewinding strategy is used to retrain the pruned network and sparse training is still adapted. Repeat the above steps to achieve the desired compression ratio. Finally, the pruned network is fine-tuned using the same learning rate as the last training round. The whole algorithm is shown in Algorithm 1. The symbol $f(x;\theta_n;\gamma_n)$ represents the network function given input data x, parameter set θ_n and the BN scale factor set γ_n at round n.

A. Iterative pruning strategy

In order to prune the network more accurately, we introduce iterative pruning into the original channel-level pruning method

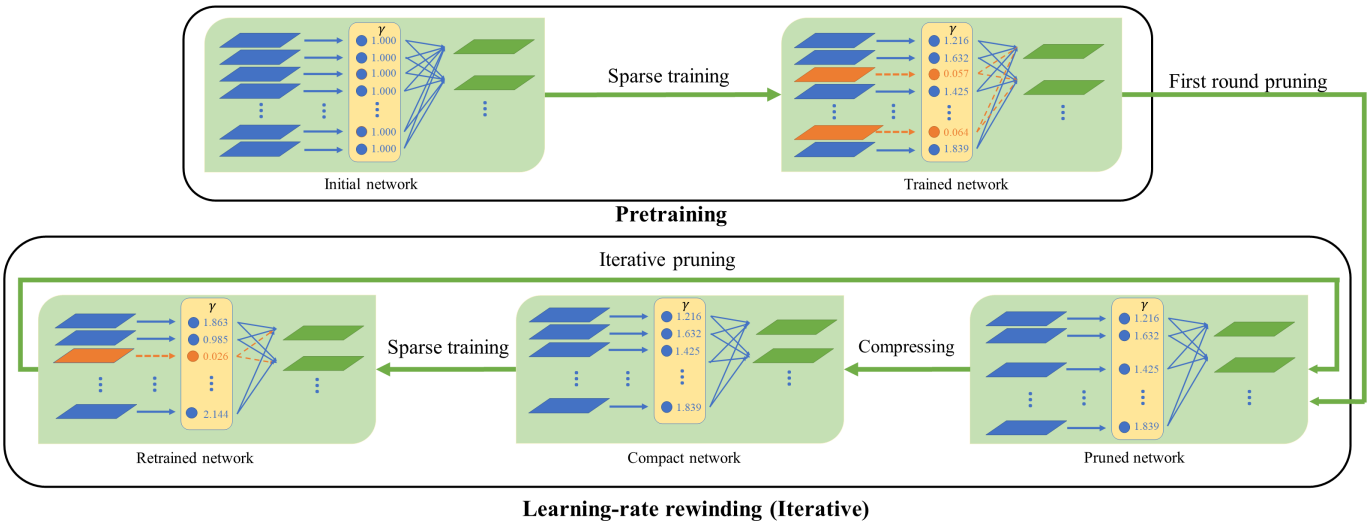

Fig. 1. Channel-level pruning with learning rate rewinding (CPLR).

to replace one-shot pruning. Assuming that percentage α of all the channels are pruned in each round, after n rounds, the compression ratio (CR) and weight sparsity ratio (WSR) are shown in the following equation.

$$CR(n) = (1-\alpha)^n \tag{1}$$

$$WSR(n) = 1 - CR = 1 - (1-\alpha)^n \tag{2}$$

where $CR(n)$ is the rate of remaining parameters to the original parameters, and $WSR(n)$ is the rate of the pruned parameters with the original parameters. Then, we introduce sparse-training strategy to every training rounds, whose principle is as follows:

$$L = \sum_{(x,y)} l(f(x,\theta,\gamma),y) + \lambda \sum_{\gamma \in BN} h(\gamma) \tag{3}$$

$$h(\gamma) = |\gamma| \tag{4}$$

The sparse regular term $\lambda \sum h(\gamma)$ (L1-norm was selected in our algorithm) is added to the normal training loss function. $f(x,\theta,\gamma)$ is the network function given input data x, parameter set θ and the BN scale factor set γ. y repesents the label of each input sample. $\sum_{(x,y)} l(f(x,\theta,\gamma),y)$ is function of the binary cross entropy (BCE) loss used as our loss function.

Sparse training has been proved to be effective for dynamic pruning networks [23]. In the training process, the scale factor γ will become more and more discriminative (the closer γ is to 0, the less important is the channel).

Using iterative pruning instead of one-shot pruning has the following superiorities. First, a small pruning rate in each round can avoid a great deal of destruction on the original network, so the degradation process of the network generalization performance will be slower using iterative pruning. Second, the fine-tuning process is easier because we only prune off a small proportion of total parameters, and also iterative pruning can get better performance than one-shot pruning. In addition, the adjustment of the compression ratio is more flexible using an iterative way because we can change the number of training rounds to get a different compression ratio. Finally, the sparse-training is used during each round of training other than only one round in the original channel-level pruning method, which is more sufficient to obtain the higher weight sparse ratio.

B. Leaning-rate rewinding strategy

In our algorithm, we use the LTH leaning-rate rewinding strategy to guide the training and pruning process, the process is as follows:

$$W_0 \xrightarrow{\alpha_1} W_1^{\alpha_1} \xrightarrow{m_1} W_1 \xrightarrow{\alpha_2} \ldots \xrightarrow{\alpha_n} W_n^{\alpha_n} \xrightarrow{m_n} W_n \xrightarrow{\alpha_{n+1}} \ldots \tag{5}$$

$$\alpha_{n+1} = \alpha_n \times k \tag{6}$$

$$k = \begin{cases} 1 & (CR(n) \geq 0.2) \\ [2,10] & (CR(n) < 0.2) \end{cases} \tag{7}$$

where α_n represents the learning rate used in the n-th round. $W_n^{\alpha_n}$ represents the parameters trained in the n-th round using the learning rate α_n, m_n represents the generated pruning mask after n rounds and W_n represents the parameters of the pruned network. k is the amplification factor of the learning-rate controlled by $CR(n)$.

When the preset pruning-rate is too small, the weights of the network change slightly, and the similar structure may be utilized by using the same learning rate to retrain the network, makes the network converge faster to the early-stop point with the similar performance as the original network [26]. In addition, when the compression ratio is less than 20%, the learning rate increases by 2-times to 10-times, as shown in (7).

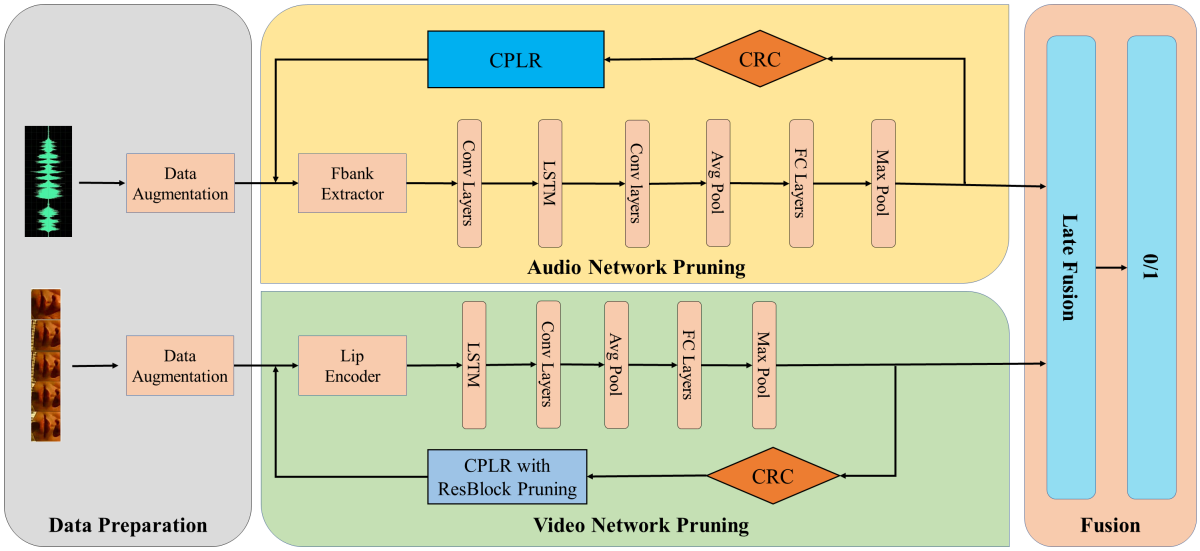

Fig. 2. The architecture of audio-visual wake word spotting system using the proposed CPLR, "CRC" represents "Compression Ratio Calculator"

IV. Application on AVWWS network

The proposed CPLR approach is evaluated on audio-visual wake word spotting task based on the MISP2021 challenge dataset [15]. Fig. 2 shows the our architecture, which mainly consists of three parts: audio-only system, video-only system and fusion part. The details will be elaborated in the following subsections.

A. Audio-only WWS system

Audio system officially provided by the challenge [15] was selected, which consists of two layers of 2D-convolution layers, one long short term memory (LSTM) layer and three convolution layers. We added one batch normalization layer after per convolution layer.

B. Video-only WWS system

Same as the audio system, the corresponding video system [15] officially provided was selected, which consists of a ResNet-18, a LSTM layer and three convolution layers. The ResNet-18 has been pre-trained on the lip-reading task with the details can be referred to [15]. There are two kinds of the convolution layers in the residual block (ResBlock) of ResNet-18, one kind of convolution layers are connected with other convolution layers, called "cross-layer connections", while the other do not connect with any other convolution layers [28]. Some changes have been designed to prune the ResNet-18 when implementing channel-level pruning. Specifically, for the cross-layer connections, which cannot be pruned because they must be the same as the convolution layers connected with them [28]. Accordingly, we only prune the channels not in cross-layer connections.

C. Audio-visual fusion

Consistent with [15], we adopted the decision-level fusion combining the posterior probabilities from separate audio and visual WWS subsystem, as shown in the following formula:

$$P_{\mathrm{AV}} = \alpha \times P_{\mathrm{A}}(y_{\mathrm{A}}|f_{\mathrm{A}}) + \beta \times P_{\mathrm{V}}(y_{\mathrm{V}}|f_{\mathrm{V}}) \tag{8}$$

where $P_{\mathrm{A}}(y_{\mathrm{A}}|f_{\mathrm{A}})$ and $P_{\mathrm{V}}(y_{\mathrm{V}}|f_{\mathrm{V}})$ are the posterior probabilities of wake word presence (y_{A} or y_{V}) generated by input the audio features f_{A} and the video features f_{V}, respectively. α and β are the weights of audio-only and video-only systems. The output of systems is compared with the preset threshold (th_{A}, th_{V}, th_{AV}) after the sigmoid operation.

V. Experiments and Results

A. Dataset and metric

We conduct the experiments on the MISP2021 AVWWS dataset [15], which contains about more than 120 hours of audio-visual data and has been divided with non-overlapping speakers across the training, development and evaluation sets. The wake word is "Xiao T Xiao T". The combination of false reject rate (FRR) and false alarm rate (FAR) is adopted as the evaluation metric, which is defined as follows:

$$Score = FRR + FAR = \frac{N_{FR}}{N_{wake}} + \frac{N_{FA}}{N_{non-wake}} \tag{9}$$

where N_{wake} and $N_{non-wake}$ denote the number of samples with the wake word and without the wake word in the evaluation set, respectively. N_{FR} denotes the number of samples that include the wake word but where the WWS system erroneously does not detect it. N_{FA} is the number of samples that do not contain the wake word but where the WWS system erroneously detect it. The lower Score, the better the system performance.

B. Results for audio-only system

Considering the complexity and challenge of far-field environment, we evaluated the proposed approach on far-field audio. Some data augmentation methods mentioned in [15] were

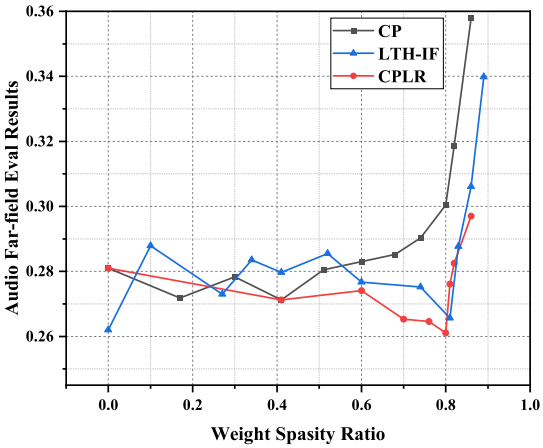

Fig. 3. Performance curves of different pruning methods on the evaluation set for audio-only system.

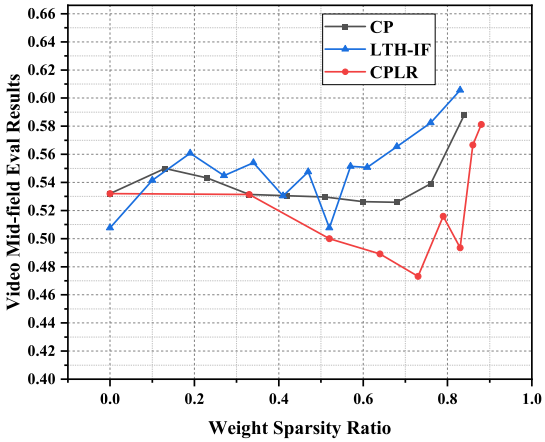

Fig. 4. Performance curves of different pruning methods on the mid-field evaluation set for video-only system.

adopted. The original one-shot channel-level pruning, LTH-IF and the proposed CPLR approach were used to prune the audio-only system separately. The results are shown in Fig. 3. The x-axis represents the WSR and the y-axis represents the $Score$ on the evaluation set.

According to Fig. 3, the proposed CPLR can not only achieve better performance when pruning but also obtain a higher WSR. With the decreasing of the parameters, the scores of the networks compressed by the above three methods all show a tendency of falling first and then rising. When the WSR is more than 0.8, the performance of all these three networks degrades rapidly. Compared to the other two methods, CPLR achieves the best score (0.2611) with a highest WSR of 0.8.

C. Results for video-only system

For the video-only system, we use the mid-field data and the far-field data. Same data augmentation methods used in [15] were also adopted.

1) Video-only WWS results on mid-field: The original one-shot channel-level pruning, LTH-IF approach and the proposed CPLR approach were used to prune the video-only system separately. The results are shown in Fig. 4.

According to Fig. 4, the proposed CPLR achieves the best performance compared to the other two methods. When the WSR is about 73%, the score of 0.4732 is obtained. Finally we can reach the maximum WSR of 0.83 at round 6 without performance degradation. Compared with the original channel-level pruning method, the minimum score decreases by 0.0527 and the WSR increases by 0.15. In comparison to the LTH-IF method, the proposed CPLR achieves better performance with a higher WSR.

The introduction of sparse-training per round makes the weights of the network more sparse, making it easier to identify the important channels, so the proposed CPLR can achieve better performance than the LTH-IF and one-shot

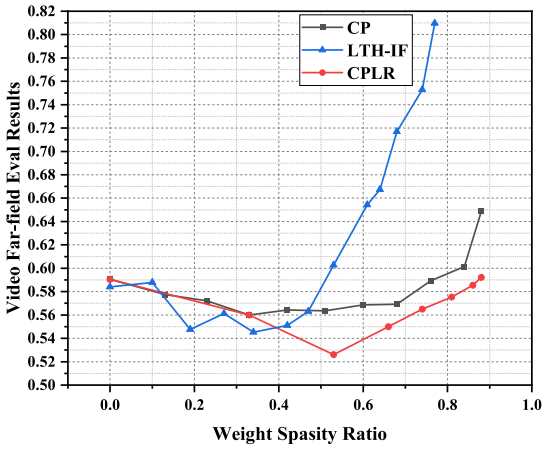

Fig. 5. Performance curves of different pruning methods on the far-field evaluation set for video-only system.

channel-level pruning, especially when WSR is high. Besides, the iterative manner performs better than the one-shot manner and the effectiveness of the learning-rate rewinding strategy has been proved by this experiment.

2) Video-only WWS results on far-field: The one-shot channel-level pruning, LTH-IF and the proposed CPLR were compared in Fig. 5.

According to Fig. 5, the lowest score (0.5260) is obtained by the proposed CPLR with a WSR of about 0.51. Compared with the original channel-level pruning method, the minimum score decreases by 0.034 and the final WSR increases by 0.1. In addition, compared to the LTH-IF algorithm, the proposed CPLR performs worse in the first few rounds, but performs better when WSR is more than 0.40. The minimum score decreases by 0.022 and the WSR increases by 0.39 from

TABLE I
FUSION PERFORMANCE OF DIFFERENT PRUNING METHODS.

Fusion	Audio		Video		Audio-Video	
Pattern	Para	Score	Para	Score	Para	Score
Original network						
Far+Middle	2.68M	0.2620	13.07M	0.5130	15.75M	0.2190
Far+Far	2.68M	0.2620	13.07M	0.5840	15.75M	0.2510
CPLR algorithm						
Far+Middle	0.56M	0.2611	2.21M	0.4935	2.77M	0.2092
Far+Far	0.56M	0.2611	1.85M	0.5855	2.41M	0.2432
CP algorithm						
Far+Middle	1.32M	0.2805	4.20M	0.5259	5.52M	0.2206
Far+Far	1.32M	0.2805	4.20M	0.5692	5.52M	0.2653
LTH-IF algorithm						
Far+Middle	0.56M	0.2656	6.25M	0.5076	6.81M	0.2174
Far+Far	0.56M	0.2656	6.91M	0.5631	7.47M	0.2500

LTH-IF to CPLR. The proposed CPLR not only obtains better performance but also achieves higher WSR compared with the other two methods.

With the decreasing of the parameters in the network, the scores of the pruned networks all tend to fall first and then rise. The performance of LTH-IF deteriorates rapidly when WSR is more than 0.4 while the performances of CP and CPLR deteriorate more steadily. The reason may be that the performance of the LTH-IF approach largely depends on the effectiveness of the pre-train process, and if the pre-train process does not achieve good performance, some important connections in the network may be pruned inaccurately leading to worse performance. If it drops into a vicious circle, the performance will deteriorate rapidly. While CPLR approach considers all connections in a layer, the deteriorating process of the system performance will be slower.

D. Results for audio-visual fusion systems

Besides the above results, we also conduct audio-visual fusion, and the results under multiple hybrid configurations are shown in TABLE I. "Pattern" indicates the data type combination of audio and video inputs. For example, "Far+Middle" means the combination of far-field audio and mid-field video. "Para" denotes the size of model parameters in Bytes.

CPLR algorithm achieves the best fusion performance with the highest WSR of 0.83, significantly outperforming the other two algorithms. Benefiting from the proposed CPLR compression algorithm, a score gain of 0.0528 compared with the original audio-only WWS system is obtained, and the size of parameter set only increases by 0.09M Bytes in the setting of using mid-field video. The proposed CPLR algorithm makes it possible to introduce visual information to the audio WWS system, improves the performance of the WWS system significantly without increasing a large amount of parameters.

E. Brief summary

Based on the results of the compression of audio-only, video-only and audio-visual fusion system using the above-mentioned three methods CP, LTH-IF and CPLR, we can give a brief summary of the advantages and disadvantages of each method. As a one-shot pruning method, CP only needs one training and pruning round. CPLR usually needs 5 to 8 training and pruning rounds to reach the maximum WSR and LTH-IF approach needs over 10 rounds, takes up the most training cost of above three methods. When WSR is low, LTH-IF and CPLR method can obtain better performance. When the WSR is high, CPLR can achieve the best performance, which is the most notable advantage of CPLR.

VI. CONCLUSIONS

In this article, we propose a structured pruning approach CPLR, which combines channel-level pruning and the learning rate rewinding strategy. Specifically, we extend the original channel-level pruning from one-shot manner to iterative manner. Then, we further introduce the learning rate rewinding strategy to the iterative channel-level pruning. Verified on the evaluation set of the MISP2021 AVWWS challenge dataset in both single modalities and fusion systems, the proposed approach yields consistent improvements.

VII. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427 and the iFLYTEK Co.,Ltd.

REFERENCES

[1] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," in *ICASSP*, 2019, pp. 6341–6345.

[2] I. López-Espejo, Z.-H. Tan, and J. Jensen, "Improved external speaker-robust keyword spotting for hearing assistive devices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1233–1247, 2020.

[3] Y. Wang, H. Lv, D. Povey, L. Xie, and S. Khudanpur, "Wake word detection with streaming transformers," in *ICASSP*, 2021, pp. 5864–5868.

[4] X. Ji, M. Yu, J. Chen, J. Zheng, D. Su, and D. Yu, "Integration of multi-look beamformers for multi-channel keyword spotting," in *ICASSP*, 2020, pp. 7464–7468.

[5] I. López-Espejo, Z.-H. Tan, and J. Jensen, "A novel loss function and training strategy for noise-robust keyword spotting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2254–2266, 2021.

[6] Y. A. Huang, T. Z. Shabestary, and A. Gruenstein, "Hotword cleaner: Dual-microphone adaptive noise cancellation with deferred filter coefficients for robust keyword spotting," in *ICASSP*, 2019, pp. 6346–6350.

[7] Y. Gao, Y. Mishchenko, A. Shah, S. Matsoukas, and S. Vitaladevuni, "Towards data-efficient modeling for wake word spotting," in *ICASSP*, 2020, pp. 7479–7483.

[8] Y. Gao, N. D. Stein, C. Kao, Y. Cai, M. Sun, and T. Z. et al., "On front-end gain invariant modeling for wake word spotting," in *Interspeech*, 2020, pp. 991–995.

[9] H. Park, P. Zhu, I. Lopez-Moreno, and N. Subrahmanya, "Noisy student-teacher training for robust keyword spotting," in *Interspeech*, 2021, pp. 331–335.

[10] A. Hard, K. Partridge, C. Nguyen, N. Subrahmanya, A. Shah, P. Zhu, I. Lopez-Moreno, and R. Mathews, "Training keyword spotting models on non-iid data with federated learning," in *Interspeech 2020*, 2020, pp. 4343–4347.

[11] D. Stewart, R. Seymour, A. Pass, and J. Ming, "Robust audio-visual speech recognition under noisy audio-video conditions," *IEEE transactions on cybernetics*, vol. 44, no. 2, pp. 175–184, 2013.

[12] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *ICASSP*, 2015, pp. 2130–2134.

[13] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[14] L. Momeni, T. Afouras, T. Stafylakis, S. Albanie, and A. Zisserman, "Seeing wake words: Audio-visual keyword spotting," in *31st British Machine Vision Conference 2020, BMVC 2020,*, 2020.

[15] H. Zhou, J. Du, G. Zou, Z. Nian, C. Lee, S. M. Siniscalchi, S. Watanabe, O. Scharenborg, J. Chen, S. Xiong, and J. Gao, "Audio-visual wake word spotting in misp2021 challenge: Dataset release and deep analysis," in *Interspeech*, 2022.

[16] J. Cheng, P.-s. Wang, G. Li, Q.-h. Hu, and H.-q. Lu, "Recent advances in efficient computation of deep convolutional neural networks," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 64–77, 2018.

[17] L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485–532, 2020.

[18] H. Mao, S. Han, J. Pool, W. Li, X. Liu, Y. Wang, and W. J. Dally, "Exploring the regularity of sparse structure in convolutional neural networks," *CoRR*, vol. abs/1705.08922, 2017.

[19] Z. Huang and N. Wang, "Data-driven sparse structure selection for deep neural networks," in *ECCV*, 2018, pp. 304–320.

[20] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," in *ICLR*. OpenReview.net, 2017.

[21] Y. He, P. Liu, Z. Wang, Z. Hu, and Y. Yang, "Filter pruning via geometric median for deep convolutional neural networks acceleration," in *CVPR*, 2019, pp. 4335–4344.

[22] C. Gamanayake, L. Jayasinghe, B. K. K. Ng, and C. Yuen, "Cluster pruning: An efficient filter pruning method for edge ai vision applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 4, pp. 802–816, 2020.

[23] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *ICCV*, 2017, pp. 2755–2763.

[24] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in neural information processing systems (NIPS)*, vol. 28, 2015.

[25] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *ICLR*, 2019.

[26] A. Renda, J. Frankle, and M. Carbin, "Comparing rewinding and fine-tuning in neural network pruning," in *ICLR*, 2020.

[27] H. Zhou, J. Du, C.-H. Huck Yang, S. Xiong, and C.-H. Lee, "A study of designing compact audio-visual wake word spotting system based on iterative fine-tuning in neural network pruning," in *ICASSP*, 2022, pp. 7572–7576.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.