INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Multiple-Teacher Pruning Based Self-Distillation (MT-PSD) Approach to Model Compression for Audio-Visual Wake Word Spotting

Haotian Wang[1], *Jun Du*[1*], *Hengshun Zhou*[1], *Chin-Hui Lee*[2], *Yuling Ren*[3], *Jiangjiang Zhao*[3]

[1] National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, Anhui, China
[2] School of Electrical and Computer Engineering, Georgia Institute of Technology, American
[3] China Mobile Online Services Company Limited, China

htwang6@mail.ustc.edu.cn, jundu@ustc.edu.cn, chl@ece.gatech.edu

Abstract

We propose a novel model compression approach using multiple-teacher pruning based self-distillation for audio-visual wake word spotting, facilitating compact neural network implementations without sacrificing system performances. In each stage of the proposed framework, we prune a teacher model obtained in the previous stage to generate a student model, then fine-tune it with teacher-student learning and use it as a new teacher model for following stages. A normalized intra-class loss is designed to optimize this pruning based self-distillation (PSD) process. Both single-teacher PSD (ST-PSD) and multi-teacher PSD (MT-PSD) are adopted in the fine-tuning process each stage. When tested on audio-visual wake word spotting in MISP2021 Challenge, the two proposed techniques outperform state-of-the-art methods in both system performances and model efficiencies. Moreover, MT-PSD that leverages upon the complementarity of multiple teachers obtained in different stages also outperforms ST-PSD.

Index Terms: teacher-student learning, knowledge distillation, normalized intra-class correlation, structured pruning, audio-visual wake word spotting

1. Introduction

The goal of audio-visual wake word spotting (AVWWS) is to recognize a predefined wake word [1]. Generally, WWS modules are limited by restricted computational and memory resources. By integrating visual information, AVWWS systems can boost the performance of an audio-only system [2, 3, 4]. However, the increased number of parameters in AVWWS systems compared to the audio-only systems may hinder their deployment on mobile devices. Accordingly, designing an effective compression method for neural network based AVWWS models is crucial to ensure the practicality of audio-visual systems. Pruning is one method to compress the neural networks [5, 6], and it can be categorized into structured and unstructured pruning [7, 8, 9]. Filter pruning is one type of structured pruning that can be realized by Taylor Series expansion, geometric median (FPGM) and other methods [10, 11, 12]. Channel-level pruning [13, 14] using batch normalization (BN) layers can also yield promising results. Furthermore, the training and pruning approaches are also very important. In 2019, lottery ticket hypothesis (LTH) [15] was proposed by combining existing pruning and training modes. Researchers have compared various strategies based on LTH [16, 17, 18], and one effective scheme is to use learning rate rewinding [18]. Recently, CPLR [19] by integrating channel-level pruning and learning-rate rewinding strategies was proposed. The channel-level pruning method was originally a one-shot pattern but has since been improved to an iterative pattern, which is further guided by the learning rate rewinding strategy [18] in the case of CPLR. It leverages upon both channel-level and LTH pruning. When tested on the AVWWS task, CPLR yields consistent improvements in both system performance and model efficiency.

Knowledge distillation (KD) is another effective network compression method, which enhances the performance of smaller models by transferring knowledge from a larger model during training [20, 21, 22]. The traditional approach to KD is to match the probability prediction scores between the teacher and student models using Kullback-Leibler (KL) divergence [20, 23, 24]. Recently, several studies [25, 26, 27] have been conducted to tackle the problem of poor learning issues in the student network when the size of student and teacher models differs significantly. For instance, TAKD [26] reduces the gap between teacher and student networks by incorporating an intermediate teaching assistant of moderate model size. DIST [28] employs a correlation-based loss to explicitly capture the intrinsic inter-class relations from the teacher. In addition, leveraging upon distilled information from multiple teacher models can also improve the performance of the student model [29, 30, 31]. For example, researchers use knowledge distilled from multiple acoustic models to construct accurate and compact neural networks [30], and CA-MAD [31] further introduces sample-wise reliability for each teacher prediction.

Despite their advantages, the compression methods discussed above also come with a few limitations. For instance, pruning-based methods, like CPLR, tend to fall short of achieving high compression ratios. Moreover, the pruned network's performance can deteriorate rapidly when the network has limited parameters, which restricts its ability to learn from the ground-truth labels. On the other hand, KD-based methods, such as DIST, lack explicit guidance on how to generate the optimal student model for learning knowledge from the given teacher model, and choosing an unsuitable student model can lead to inefficient distillation. Additionally, most KD-based methods are a one-shot compression process, which limits their effectiveness in compression. In this paper, we proposed a data-driven network compression approach based on structured pruning and knowledge distillation called pruning based self-distillation (PSD), which leverages the complementarity between pruning-based methods and KD-based methods. The PSD approach we propose is a multi-stage method, wherein each stage involves generating an optimal student model from the previous teacher model using teacher-student pruning. Subsequently, the student model is fine-tuned using teacher-student fine-tuning to attain high-performance level and is then employed as the new teacher model for the following stages. In addition, a normalized intra-class loss is designed for teacher-

*corresponding author

10.21437/Interspeech.2023-1717

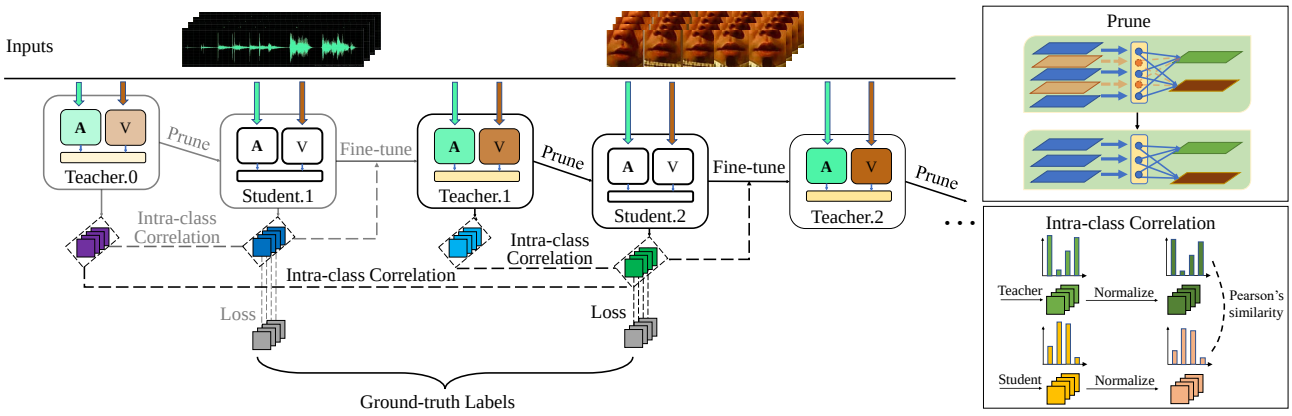


Figure 1: *Pruning based self-distillation for designing compact audio-visual network.*

student fine-tuning. We also adopt both single-teacher and multi-teacher learning strategies during the fine-tuning process, namely ST-PSD and MT-PSD. We tested the proposed approach on the MISP2021 AVWWS baseline system [32] and achieved a performance gain of 8.3% relative improvement using a lightweight model that only contained 5.9% of the original parameters.

2. Proposed Self-Distillation Frameworks

In this section, we introduce the proposed compression framework, which is illustrated in Figure 1. In the first stage, the initial model is pre-trained to the early-stop point, obtaining initial teacher model Teacher.0. Then, the model is pruned to obtain the next stage model Student.1. In the second stage, student model is updated with the guidance of the first-stage teacher model to become the second-stage teacher model Teacher.1. The student model and teacher model share the same network structure within the same stage, but they have different parameters. These aforementioned steps are repeated through several stages to obtain the most compact model. It is worth noting that only teacher model of last stage or all teacher models of previous stages can be used as guidance to fine-tune the generated student model results in varying guiding effects.

2.1. Initial teacher model

Firstly, we pre-trained the initial model with several data augmentation strategies mentioned in [32] to generate the first teacher model. The binary cross-entropy (BCE) loss function is used in our task. The sparse-training is also adopted to make the scale factor γ in BN layers more discriminating to achieve more accurate pruning of insignificant channels. The loss function used during pre-training is defined as follows:

$$L = \sum_{(x,y)} l(f(x,\theta,\gamma),y) + \lambda \sum_{\gamma \in \mathrm{BN}} h(\gamma) \tag{1}$$

$$h(\gamma) = |\gamma| \tag{2}$$

The sparse regular term $\lambda \sum h(\gamma)$ (L1-norm was selected in our algorithm) is added to the standard 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 represents the label of each input sample.

2.2. Teacher-student pruning

The optimal student model of a new stage is generated by pruning the teacher model of previous stage. We focus on eliminating less-significant channels in convolution and BN layers [13]. This pruning technique is a two-phase process, which can be outlined as follows, where k represents the pruning rate:

1. Mask generating
   - Copy all the factors in all BN layers of the teacher model to a list, then sort the list in ascending starting from the smallest factor, generating the new list: $\Gamma_{bn}[0:N]$.
   - Determine the pruning threshold: $thre = \Gamma_{bn}[k \cdot N]$.
   - Generate pruning mask for all BN layers based on the magnitude of the factors and threshold.
2. Pruning based on mask
   - Prune off the channels of BN layers based on the mask.
   - Prune off the channels of each convolution layer that are corresponding to the pruned channels in the BN layer followed.
   - Obtain the pruned model as the new student model.

Notably, some layers are interconnected with other layers in the residual block of ResNet [33], and we create the identical mask for these layers by weighted factors of the associated BN layers.

The student model generated via channel-level pruning efficiently eliminates superfluous connections from the previous teacher model, leading to greater utilization of weights. At the same time, it preserves those connections that convey comparatively significant information from the previous teacher model, ensuring that critical information is not lost during the following teacher-student fine-tuning process.

2.3. Teacher-student fine-tuning

In this section, we discuss two forms of teacher-student fine-tuning: Learning from teacher of single stage (ST-PSD) and learning from teachers of multiple stages (MT-PSD). The learning rate rewinding strategy is used to control the learning rate of the fine-tuning process.

2.3.1. Learning from teacher of single stage (ST-PSD)

Considering the structural divergence between the student model and teacher model, we allow the student model to learn from the nearest teacher model, as it has a more akin structure to the student model, thereby rendering the distillation process

more effective. In other words, similar to TAKD [26], we utilize the nearest teacher model as an assistant model to guide the fine-tuning process of the student model.

Formally, the output prediction vector of the student model of n-th stage and all previous teacher model can be represented as $\mathbf{Y}^{s_n}, \mathbf{Y}^{t_{n-1}}, \mathbf{Y}^{t_{n-2}}, ..., \mathbf{Y}^{t_0} \in \mathbb{R}^{B\times 1}$, where B denotes the batch size, s_n represents student model of n-th stage and t_i represents teacher model of i-th stage. The loss function of student model, with ground-truth labels $\mathbf{L} \in \mathbb{R}^{B\times 1}$, can be denoted as

$$L_{s_n} = \text{BCE loss}(\mathbf{Y}^{s_n}, \mathbf{L}) + \lambda \cdot \sum_{\gamma_n \in \text{BN}} h(\gamma_n) \tag{3}$$

where γ_n is the score factor of BN layers in n-th student model and $h(\gamma_n)$ is illustrated in Eq.2. The teacher-student loss function between $\mathbf{Y}^{s_n}$ and $\mathbf{Y}^{t_i}$ is expressed as

$$L_{t_i} = 1 - \rho_p(f_{\text{norm}}(\mathbf{Y}^{s_n}), f_{\text{norm}}(\mathbf{Y}^{t_i})) \tag{4}$$

Same with DIST [28], Pearson's similarity is adopted [34],

$$\begin{aligned}\rho_p(\mathbf{u}, \mathbf{v}) &= \frac{\text{Cov}(\mathbf{u}, \mathbf{v})}{\text{Std}(\mathbf{u})\text{Std}(\mathbf{v})} \\ &= \frac{\sum_{i=1}^{B}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{B}(u_i - \bar{u})^2 \sum_{i=1}^{B}(v_i - \bar{v})^2}}\end{aligned} \tag{5}$$

f_{norm} is the normalized function to balance the impact of each sample in the batch, we employ linear normalized function, as shown in Eq. 6

$$f_{\text{norm}}(\mathbf{Y}) = \frac{\mathbf{Y}}{\sum_{k=1}^{B} Y_k} \tag{6}$$

To summarize, the single-teacher loss can be expressed as below, where α denotes the weight factor:

$$L_{\text{ST}}^{n} = L_{s_n} + \alpha \cdot L_{t_{n-1}} \tag{7}$$

2.3.2. Learning from teachers of multiple stages (MT-PSD)

Using the ST-PSD approach, various stages of models are obtained. Demonstrated in Section 4.2 , We have observed that these models seem to be complementary in terms of their predictions.

Therefore, it is a logical notion to have the student model of the n-th stage learn from all teacher models of previous stages to achieve a complementary distillation effect, the KD loss of MT-PSD can be expressed in Eq. 8, as follows:

$$L_{\text{MT}}^{n} = L_{s_n} + \alpha \cdot \sum_{i=0}^{n-1} \beta^{n-1-i} L_{t_i} \tag{8}$$

where $\beta \in (0, 1)$ is the forgetting factor over stages. There are two main reasons for adopting the forgetting factor: Firstly, the student model and teacher model that are closer in stage share a stronger structural correlation, which enables the student model to learn better from the closer teacher model. Secondly, the closer teacher model usually performs better, and assigning a higher weight can boost the performance of the student model.

Learning from teachers of multiple stages can be viewed as a type of system fusion approach that does not require additional parameters. The teacher models from different stages offer complementary knowledge that can be learned by the student model to enhance its performance. The forgetting factor assigned to each stage enables the student model to learn more significant information while discarding less important information, resulting in improved overall performance.

3. Application to AVWWS

The proposed compressing approach is evaluated on the AVWWS task based on the MISP2021 challenge [32]. We use the official baseline system, whose details will be elaborated in the following subsections.

3.1. Single WWS subsystems

The baseline WWS system consist of audio subsystem and video subsystem. The audio-only system consists of two layers of 2D-convolution layers as frontend, one long short-term memory (LSTM) layer and three convolution layers as backend. We added one BN layer after per convolution layer to execute channel-level pruning. The video-only system consists of a ResNet-18 [33] with 3D-convolution layers as frontend, an LSTM layer and three convolution layers as backend. The ResNet-18 has been pre-trained on the lip-reading task [32].

3.2. Audio-visual fusion

Consistent with [32], we adopted the decision-level fusion combining the posterior probabilities from separate audio and visual WWS subsystems. Whose principle is as follows:

$$P_{\text{AV}} = k_a \times P_{\text{A}}(y_{\text{A}}|f_{\text{A}}) + k_v \times P_{\text{V}}(y_{\text{V}}|f_{\text{V}}) \tag{9}$$

where $P_{\text{A}}(y_{\text{A}}|f_{\text{A}})$ and $P_{\text{V}}(y_{\text{V}}|f_{\text{V}})$ are the posterior probabilities of wake word presence generated by input features, respectively. k_a and k_v 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.

4. Experiments and Results

To test the efficiency of the proposed compression approach, several related experiments have been conducted.

4.1. Dataset and metric

We conduct the experiments on the MISP2021 AVWWS dataset [32] in this research. 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, whose principle is as follows:

$$Score = FRR + FAR \tag{10}$$

The lower Score, the better the system performance.

4.2. Complementarity of multiple teacher models

To demonstrate complementary in various stages of teacher models in terms of prediction,

we apply the proposed ST-PSD approach to audio-only and video-only systems and perform fusion over stages, whose principle is as follows:

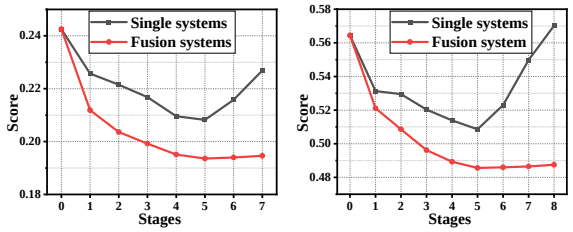
$$P_n^{\text{mul}} = \frac{1}{n} \sum_{i=0}^{n} P_i(y|f) \tag{11}$$

where n is the pruning rounds, $P_i(y|f)$ is the posterior probabilities of wake word presence generated by input features of model of $i{-}th$ stage. The output of systems is compared with preset threshold, the results on far-field are shown in Figure 2.

Based on the results above, it is evident that the fusion systems surpass the signal systems in terms of performance. Moreover, the incorporation of teacher models from more stages leads to significant improvement in performance, this indicates the complementary nature of multiple teacher models.

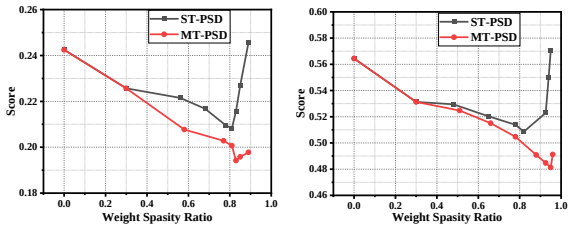
Table 1: *Performance comparison of different compression methods on AVWWS system.*

Methods	Audio-only system		Video-only system		Fusion system	
	Parameters	Score	Parameters	Score	Parameters	Score
Baseline [32]	2.68M(100%)	0.2610	13.03M(100%)	0.5840	15.75M(100%)	0.2510
Channel-level pruning [13]	1.32M(49.2%)	0.2805	4.20M(32.2%)	0.5692	5.52M(32.2%)	0.2653
LTH-IF [35]	0.56M(20.9%)	0.2656	6.91M(53.0%)	0.5631	7.47M(47.4%)	0.2500
CPLR [19]	0.56M(20.9%)	0.2611	1.85M(14.2%)	0.5855	2.41M(15.3%)	0.2432
DIST [28]	0.82M(30.6%)	0.2078	2.18M(16.7%)	0.5102	3.00M(19.0%)	0.1906
ST-PSD (ours)	0.52M(19.4%)	0.2017	1.06M(8.1%)	0.5086	1.58M(10.0%)	0.1877
MT-PSD (ours)	**0.28M(10.4%)**	**0.1978**	**0.66M(5.1%)**	**0.4813**	**0.94M(5.9%)**	**0.1679**



(a) Audio-only system. (b) Video-only system.

Figure 2: *Performance comparison of single and fusion systems of ST-PSD on separate subsystems of AVWWS.*



(a) Audio-only system. (b) Video-only system.

Figure 3: *Performance comparison of ST-PSD and MT-PSD on separate audio-only and video-only systems.*

4.3. ST-PSD vs MT-PSD

We conducted a performance comparison of the compact models generated by ST-PSD and MT-PSD individually, the results are shown in Figure 3. Weight Sparsity Ratio is the rate of the pruned parameters with the original parameters.

The results indicate that MT-PSD is generally more effective than ST-PSD. The compact systems generated by MT-PSD obtain lower Score than those produced by ST-PSD. Additionally, MT-PSD can achieve a higher weight sparsity ratio without any compromise on performance.

Additionally, it is noteworthy that the performance of ST-PSD and MT-PSD on both systems steadily improves in the initial few stages, and then declines when exceeding a certain weight sparsity ratio. Figure 4 illustrates the weight distribution in all BN layers for each round.

The results suggest that in the initial few rounds, minor parameters are pruned and the weights of the compact network display a gaussian-like pattern. As the network structure becomes more compressed, the weights tend to be averaged, making it difficult to distinguish between insignificant channels. Further compression may then prune off some essential channels, leading to a decline in system performance.

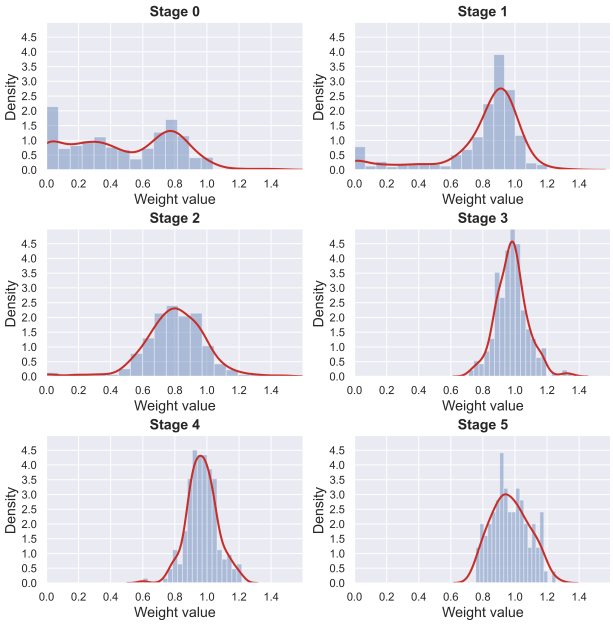


Figure 4: *Distribution of weights in all BN layers during the multiple stages of MT-PSD.*

4.4. Comparison with other techniques

We further compared the performance of the proposed method with other compression methods on the same AVWWS system. The results are presented in Table 1.

Experimental results indicate that the total AVWWS network parameters are compressed to 5.9% by MT-PSD, resulting in a decrease of 0.82 in the $Score$. MT-PSD achieves optimal fusion performance with minimal parameter usage, outperforms other pruning based and KD based compression methods.

5. Summary

We propose a multi-stage compression approach combining pruning and knowledge distillation into each stage. Using a specially designed normalized intra-class loss, our approach employs channel-level pruning to generate compact student models from previous teacher models and utilizes knowledge distilled from previous teacher models to fine-tune the student model of the current stage, then use it as a teacher model in the next stage. Evaluated on the MISP2021 AVWWS challenge data set our proposed ST-PSD and MT-PSD frameworks achieve good results in both system performances and model efficiencies.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62171427, and China Mobile Inc.

7. 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] 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.

[3] 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.

[4] R. Ding, C. Pang, and H. Liu, "Audio-visual keyword spotting based on multidimensional convolutional neural network," in *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 2018, pp. 4138–4142.

[5] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, "A comprehensive survey on model compression and acceleration," *Artificial Intelligence Review*, vol. 53, pp. 5113–5155, 2020.

[6] 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.

[7] 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.

[8] 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.

[9] S. Anwar, K. Hwang, and W. Sung, "Structured pruning of deep convolutional neural networks," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 13, no. 3, pp. 1–18, 2017.

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

[11] 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.

[12] 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.

[13] 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.

[14] C. Zhao, B. Ni, J. Zhang, Q. Zhao, W. Zhang, and Q. Tian, "Variational convolutional neural network pruning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2780–2789.

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

[16] C.-I. J. Lai, Y. Zhang, A. H. Liu, S. Chang, Y.-L. Liao, Y.-S. Chuang, K. Qian, S. Khurana, D. Cox, and J. Glass, "Parp: Prune, adjust and re-prune for self-supervised speech recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 256–21 272, 2021.

[17] N. Liu, G. Yuan, Z. Che, X. Shen, X. Ma, Q. Jin, J. Ren, J. Tang, S. Liu, and Y. Wang, "Lottery ticket preserves weight correlation: Is it desirable or not?" in *ICML*, 2021, pp. 7011–7020.

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

[19] H. Wang, J. Du, H. Zhou, H. Lu, and Y. Cao, "A novel approach to structured pruning of neural network for designing compact audio-visual wake word spotting system," in *APSIPA ASC*, 2022, pp. 820–826.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[21] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *ICCV*, 2019, pp. 1921–1930.

[22] Y. Lv, L. Wang, M. Ge, S. Li, C. Ding, L. Pan, Y. Wang, J. Dang, and K. Honda, "Compressing transformer-based asr model by task-driven loss and attention-based multi-level feature distillation," in *ICASSP*, 2022, pp. 7992–7996.

[23] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3967–3976.

[24] T. Kim, J. Oh, N. Kim, S. Cho, and S.-Y. Yun, "Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation," *arXiv preprint arXiv:2105.08919*, 2021.

[25] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *ICCV*, 2019, pp. 4793–4801.

[26] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5191–5198.

[27] W. Son, J. Na, J. Choi, and W. Hwang, "Densely guided knowledge distillation using multiple teacher assistants," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9395–9404.

[28] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," in *NeurIPS*, 2022.

[29] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1285–1294.

[30] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers." in *Interspeech*, 2017, pp. 3697–3701.

[31] H. Zhang, D. Chen, and C. Wang, "Confidence-aware multi-teacher knowledge distillation," in *ICASSP*, 2022, pp. 4498–4502.

[32] 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.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[34] K. Pearson, "Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia," *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, no. 187, pp. 253–318, 1896.

[35] 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.