INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Adaptive Neural Network Quantization for Lightweight Speaker Verification

Haoyu Wang, Bei Liu, Yifei Wu, Yanmin Qian[†]

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{fayuge, beiliu, yifei.wu, yanminqian}@sjtu.edu.cn

Abstract

Recently, speaker verification systems benefit from deep neural networks and the size of speaker embedding encoder increases with these sophisticated architectures. Nevertheless, mobile devices have inadequate memory for oversized embedding extractors, thus demanding compact networks. In this paper, we explore neural network quantization for model compression. Specifically, we first propose a novel uniform quantization method based on K-Means clustering. Then, to further improve the small model performance, mixed precision quantization is introduced. Besides, we implement a multi-stage fine-tuning (MSFT) recipe to boost the accuracy of mixed-precision model. In experiments, the performance degradation of 4 bit quantized ResNet34 is **negligible**. Our quantized models outperform former model compression methods in terms of size and accuracy. In addition, mixed-precision quantization with MSFT strategy further improves the model performance.

Index Terms: speaker verification, model compression, neural network quantization, mixed precision quantization

1. Introduction

With the implementation of deep neural networks in speaker verification systems, evident progress in model performance has made [1, 2, 3, 4]. Among these deep architectures, ResNet [5] and ECAPA-TDNN [4] are two of the most popular and efficient speaker embedding extractors. However, while larger models generally lead to better performance, their excessive memory usage restricts their application in mobile devices. Thus, there is a growing interest in finding a balance between model size and system performance.

To address this issue, model compression approaches are proposed to reduce the model size while ensuring high performance. Various methods have been developed, including reducing the number of parameters through knowledge distillation [6] and model pruning [7, 8]. Besides, some innovative lightweight model designs are raised. They build more minor architecture with spatial separable convolution kernels and less repetitive blocks [9, 10]. Others trade the accuracy of parameters for the compactness of models, namely model quantization [11, 12, 13]. Compared with other methods, the superiority of model quantization is keeping the integrity of model structure, model compression is achieved by reducing the parameter precision. Previous work states the redundancy of parameter precision commonly exists in neural networks [14, 15]. In view of this observation and previous work [16] in our field, we found model quantization a promising way to decrease the model size.

Although the quantization compression method has the advantage of maintaining the integrity of the model, there are some problems in the previous quantization methods that limit their performance[17]. The previous quantization method[16] pursues the high compression ratio of the model, but the performance of the model is seriously degraded; Other methods[18, 19] adopt fixed quantization values and training methods for all parameters in the model, resulting in excessive compression loss. To solve this problem, we propose an adaptive centroids selection strategy, which derives different quantization values for each part of the model. Our method realizes lossless compression under higher compression ratio.

Contributions of this paper are as follows: First, we develop a novel quantization method and obtain efficient uniform quantized models. Second, we implement a mixed-precision quantization algorithm to achieve better results than uniform quantization. Third, we further improve the performance of mixed-precision quantization with the multi-stage fine-tuning (MSFT) recipe. Our proposed quantization algorithm has yielded satisfactory results in experiments conducted on the Voxceleb dataset. We achieve lossless quantization at 4 bit uniform precision. The accuracy of our mixed-precision quantization model with MSFT has surpassed that of uniform quantization at a comparable model size. Additionally, our proposed method has outperformed other model compression methods in terms of performance and model size.

The content of this paper is organized as follows: we introduce related works on model quantization in section 2, the description of our methods is presented in section 3, and the experimental setup and implementation details are given in section 4. The experimental results are shown in section 5. Finally, section 6 is the conclusion.

2. Related work

In this section, we refer to previous work in model compression in speaker verification domain and briefly introduce mixed-precision quantization.

2.1. Model compression in speaker verification

Previously proposed researchers adopt various methods to reduce the speaker verification system size. [20, 21] implement knowledge distillation to speaker embedding extractor, with this technology, small models gain comparable performance against original models. [16] applies a binary quantization to the extractor. The model is compressed by 32 times and all parameters are quantized to two values; however, the model performance drops dramatically. Other approaches [10, 22] redesign encoder architectures with fewer parameters to compress model size.

2.2. Mixed-precision quantization

Mixed-precision quantization allocates the precision of parameters according to the sensitivity of each layer to quantiza-

[†]corresponding author

10.21437/Interspeech.2023-927

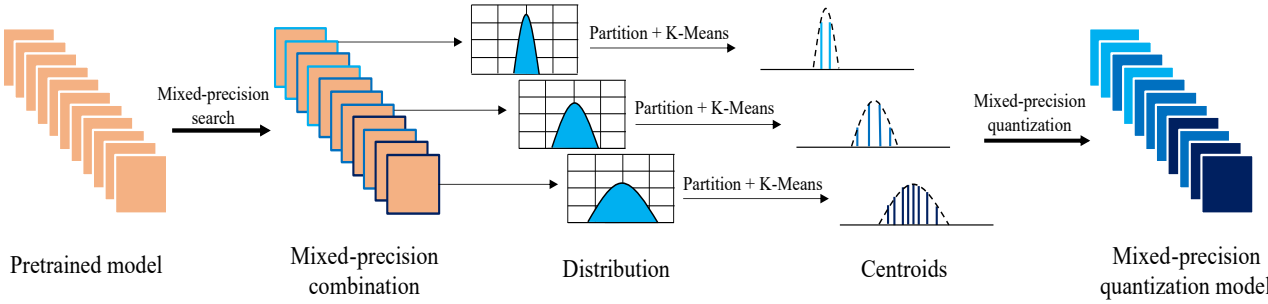


Figure 1: *The pipeline of mixed-precision quantization. Quantization centroids are determined independently for each model layer to adapt the layer-wise distribution of parameters. The algorithm of mixed-precision research is introduced in section 3.3; quantization centroids are derived by the method in section 3.2.*

tion. Uniform precision quantization applies the same bit-width quantization to all parameters, which ignores the sensibility variance of different layers to quantization. To measure this sensibility, [23] proposed a criterion related to the Hessian matrix. According to the order of sensibility and overall model size restriction, mixed-precision search is implemented to produce a reasonable quantization combination for the model. In this paper, we realize the mixed-precision quantization for performance improvement.

3. Proposed methods

In this section, we introduce in detail our uniform and mixed-precision quantization techniques and fine-tuning methods.

3.1. Basic definition of quantization

Quantization algorithms aim at preserving model performance while compressing the bit-width of parameters. Model size decreases by reducing parameter-level memory occupation. Unlike traditional CNNs, each learnable parameter in quantization model is stored at lower bit precision (3 or 4 bit, even 1 bit for binary quantization) instead of 32 bit. All parameters are approximated in few fixed values.

Full-precision quantization centroids are stored for every layer. To reduce the memory usage of parameters, the original parameters are converted into integers. The integer set N is defined as follow:

$$\boldsymbol{N} \in \{0, 1, 2, \ldots, 2^n - 1\} \tag{1}$$

where n represents the quantization precision in bit. And we build a bijection from N to the quantization centroids set C:

$$\boldsymbol{C}^{(l)} = \boldsymbol{\Psi}(\boldsymbol{N}^{(l)}) = \left\{q_1^{(l)}, q_2^{(l)} \ldots, q_{2^n}^{(l)}\right\} \tag{2}$$

where $C^{(l)}$ denotes the quantization value set in the l-th layer, Ψ is the bijection. In this way, learnable parameters can be stored in a fewer bits and converted to full-precision values in the inference stage.

Then the quantization operation $f(\cdot)$ and the final quantized parameters Q are defined as below:

$$\boldsymbol{Q}^{(l)} = \alpha^{(l)} \boldsymbol{C}^{(l)} \tag{3}$$

$$\boldsymbol{f}\left(\boldsymbol{W}^{(l)}\right) = \arg\min_{Q^{(l)}} \left|\boldsymbol{W}^{(l)} - \boldsymbol{Q}^{(l)}\right| \tag{4}$$

where α represents the learnable scaling factor, $W^{(l)}$ is the weights in the l-th layer. We achieve quantization f of parameters in the model by approximating the original weights in the nearest quantization value.

3.2. K-Means based quantization aware training

We introduce the derivation of quantization centroids of our model in this section. The weight distribution of each layer in the model differs; However, previous quantization approaches applied the same centroids for all layers.

To alleviate this mismatch between quantization centroids and parameter distribution, we proposed K-Means based Quantization Aware Training (KMQAT). First of all, kmqat divides the weight of each layer into n intervals, which ensures that each part of the parameter distribution will be assigned a corresponding centroid; Secondly, KMQAT performs K-Means clustering in each interval to ensure the minimum quantization loss in this interval. For each layer, parameters are partitioned into n intervals with same width according to their value:

$$\widehat{W}^l = W_1^l \parallel W_2^l \parallel \cdots \parallel W_{2^n}^l \tag{5}$$

where n is the bit precision of current layer, $\widehat{W^l}$ denotes the clipped weight of l-th layer, W_i^l is the i-th interval of weight.

In partition, we focus on 90% of the parameters that lie near the peak of distribution to avoid the negative effect of outliers. The ablation study conducted on this percentage is shown in Table 1. Then we implement K-Means clustering algorithm with only one center in each weight partition. The centroid set of the l-th layer is given by:

$$\boldsymbol{C}_{KMQAT}^{(l)} = \left\{\Phi(W_1^l), \Phi(W_2^l), \cdots, \Phi(W_{2^n}^l)\right\} \tag{6}$$

where Φ denotes the cluster operation. So far, we have achieved adaptive quantization by adapting the centroid to the weight distribution.

3.3. Mixed-precision quantization

From the method in section 3.2, we can achieve uniform quantization. However, due to the characteristics of deep neural network, different layers have different sensitivity to quantization, however, uniform quantization ignores this sensitivity difference. Thereby, we implement mixed-precision quantization to overcome this issue.

Mixed-precision quantization allows different bit-width in one quantized model. It improves model performance by reallocating the quantization precision across layers. As illustrated in Figure 1, we utilize uniform quantized models as candidates for mixed-precision search. Mixed-precision search distributes quantization precision to layers of the ResNet model. Then the layers are quantized at different precision. Finally, we fine-tune the mixed-precision quantized model.

We adopt different precision for each layer regarding its sensitivity to quantization. Inspired by [24], the sensitivity of

layers is estimated by the trace of the Hessian matrix. The total sensitivity is defined as:

$$\mathbf{\Omega^{Hes}} = \sum_{i=1}^{L} \mathbf{\Omega_i^{Hes}} = \sum_{i=1}^{L} \mathrm{Tr}\left(\boldsymbol{H}^{(i)}\right) \parallel \boldsymbol{W}^{(i)} - \boldsymbol{Q}^{(i)} \parallel_2^2 \quad (7)$$

where Ω^{Hes} represents the total sensitivity of the model, $H^{(i)}$ denotes the Hessian matrix of the i-th layer and L is the number of layers. There are three principals that the search algorithm obeys: First, the layer with high sensitivity to quantization should have higher precision than the ones with lower sensitivity. Second,the total model size is limited. Third, combination with lowest Ω^{Hes} is selected for mixed-precision quantization.

3.4. Multi-stage fine-tuning

Especially, we develop the multi-stage fine-tuning (MSFT) to further improve the performance of mixed-precision quantization. Due to various weight precision in the model, if the parameters of different layers are trained together during the training session, [23] indicate that the model will possibly fall into the sub-optimal solution. MSFT progressively quantizes the layers in accordance with their bit-width instead of quantizing the whole model entirely. [23] demonstrates that quantization begins from lower precision layers achieves better results. In our experiments, quantization begins with lower precision layers.

Algorithm 1 Our proposed multi-stage fine-tuning (MSFT) recipe for mixed-precision quantization

Input: f: quantization operation; W: pretrained full-precision weights; P: bit precision of each layer derived by mixed-precision search; Q: bit precision contained in P
Output: Mixed-precision quantized model

Initial $W = \{W_1, W_2, ..., W_L\}$;$P = \{p_1, p_2, ..., p_L\}$; $Q = \{q_1, q_2, ..., q_n\}$ where $q_1 < q_2 < ... < q_n$; $j = 1$
repeat
  for i in $\{1, 2, ..., L\}$ **do**
    if $p_i \leq q_j$ **then**
      $W_i \leftarrow f(W_i)$;
    else
      $W_i \leftarrow W_i$;
    end if
  end for
  Train W until convergence;
  $j \leftarrow j + 1$;
until The model is fully quantized

4. Experimental setups

4.1. Datasets

Our experiments are conducted in VoxCeleb1 [25] and Voxceleb2 [26] dataset. The pretrained model and the quantized models are trained with the development set of Voxceleb2. The test sets are Voxceleb1-O, Voxceleb1-E and Voxceleb1-H. We apply data augmentation and speed perturbation in experiments in order to obtain the robustness of the system. RIRs [27] and MUSAN [28] noises are added to training data. Speed perturbation changes the original speed of training utterances to 0.9 and 1.1 times, thus adding twice as many speakers.

4.2. Implementation Details

As a mainstream embedding extractor, pretrained ResNet34 is quantized in our experiment. Utterances length is set to 200 frames in training session. We conduct 80-dimensional Fbank input features. Additive Angular Margin (AAM) loss is adopt [29] as loss calculator, angular margin m is set to 0.2. We set the initial learning rate to 0.0001 with 3 warm-up epochs and the final learning rate of 0.00001. The equal error rate (EER) is considered as the performance reference index. The uniform quantization models are trained for 40 epochs, and 60 epochs are trained in total for mixed-precision quantization ones.

Table 1: *The ablation study on the percentage of parameters nearby 0 in partition.*

Model	Percentage of param	Vox1-O EER(%)	Vox1-E EER(%)	Vox1-H EER(%)
4 bit-KMQAT ResNet34	100%	**0.925**	1.047	1.914
	90%	0.957	**1.024**	**1.898**
	80%	0.963	1.039	1.910

5. Results and analysis

5.1. Analysis of quantization results

5.1.1. Uniform quantization results

Uniform quantization is adopted at 1, 2, 3 and 4 bit precision to the original ResNet34. We quantize all weights in convolution and linear layers, which represent 99.42% parameters in the model. Experimental results are shown in Table 2. At 4 bit uniform quantization, the performance drops relatively by 7.7% on Voxceleb1-O, the relative degradation of performance is only 1.6% and 2.6% on Voxceleb1-E and Voxceleb1-H, respectively. KMQAT realizes a lossless quantization on Vox1-E and Vox1-H at a compression ratio of 7.72x.

5.1.2. Mixed-precision quantization results

Mixed-precision quantization is proposed to reassign quantization precision among layers to improve performance. From the results of mixed-precision search, the shallower convolution layers are more sensitive to quantization, and the last convolution and fully-connected layers have lower sensitivity. Experimental results in Table 2 show that the performance improvement achieved through mixed-precision is limited. Due to the variety of bit-widths in the model, quantizing all parameters at same time impedes the normal optimization of scaling factor. Therefore, we design multi-stage fine-tuning to promote mixed-precision quantization training.

5.1.3. MSFT results

In our experiments, mixed-precision quantization models with MSFT further improves the performance at comparable size of uniform quantization. The 2.57MB mixed precision model occupies less memory than the uniform one, and achieves 5.8%, 2.6% and 3.2% relative amelioration on Vox1-O,Vox1-E and Vox1-H. Through experiments, we demonstrate that mixed-precision quantization and MSFT are effective methods to boost performance of quantization models.

5.2. Comparison of KMQAT and other compressed models

We implement previous quantization methods on our baseline for the convenience of comparison. The results are shown in Table 3. In the experiment, quantized ResNet34 surpasses other quantization model [18, 19] and compressed full-precision models with similar architecture [22, 30] in terms of EER.

At 1, 2 and 3 bit precision, experiments indicate KMQAT has obvious advantages over other model compression and quantization methods. In the extreme quantization domain, our proposed method has better results than existing binary[16] and ternary[31] quantization methods. Moreover, KMQAT overtakes some recently proposed lightweight full-precision networks [10, 32] at comparable size.

Table 2: *Performance of the full-precision and quantized ResNet34 on the Voxceleb1 test sets. "MP" means mixed-precision quantization. "MSFT" corresponds to the multi-stage fine-tuning strategy. "$\{2, 3, 4\}$" refers to a mixed-precision quantization model with a combination of 2, 3 and 4 bit quantized layers.*

Architecture	Bit-width (bit)	Model size	Compression ratio	Voxceleb1-O EER(%)	Voxceleb1-E EER(%)	Voxceleb1-H EER(%)
ResNet34	32	26.66MB	-	0.888	1.008	1.850
+KMQAT	4	3.45MB	7.72x	0.957	1.024	1.898
+KMQAT	3	2.63MB	10.11x	1.074	1.128	2.057
+KMQAT	2	1.80MB	14.81x	1.398	1.428	2.566
+KMQAT	1	0.97MB	27.48x	2.133	2.076	3.592
++MP	$\{2, 3, 4\}$	2.57MB	10.37x	1.026	1.136	2.054
+++MSFT		2.57MB	10.37x	1.015	1.099	1.992
++MP	$\{1, 2, 3, 4\}$	1.78MB	14.94x	1.367	1.456	2.592
+++MSFT		1.78MB	14.94x	1.388	1.381	2.452

Figure 2: *Centroids of different quantization methods with identical parameter distribution in 3 bit quantization.*

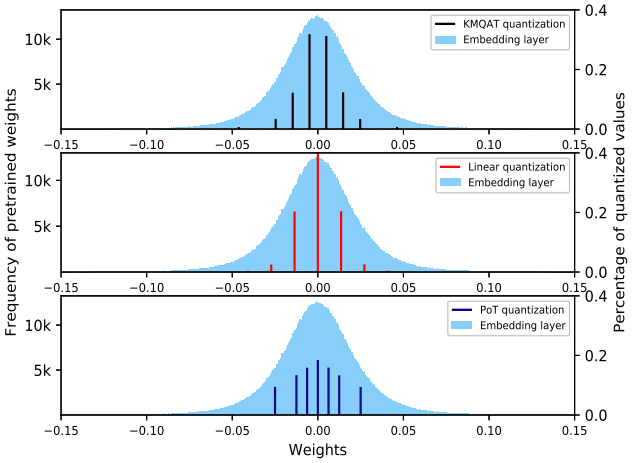
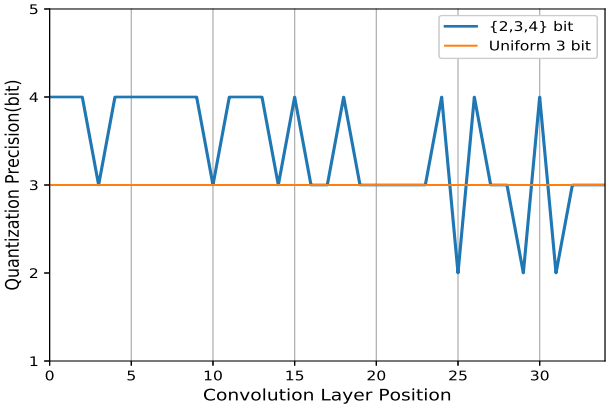


Table 3: *The experiment results of compressed/quantized ResNet34 and other full-precision compact architectures.*

Model	Size (MB)	Bit-width (bit)	Vox1-O EER(%)
KMQAT-ResNet34(Ours)	**3.45**	**4**	**0.957**
PoT-ResNet34[19](our impl.)	**3.45**	**4**	0.973
ADMM-ResNet34[18](our impl.)	5.13	6	1.73
Thin-ResNet-34[30]	5.6	32	2.36
Fast-ResNet-34[22]	5.6	32	2.37
KMQAT-ResNet34(Ours)	**1.80**	**2**	**1.398**
TWN-ResNet34[31](our impl.)	**1.80**	**2**	1.473
KMQAT-ResNet34(Ours)	0.97	**1**	**2.133**
ResNet34(binary)[16]	**0.66**	**1**	5.355
CS-CTCSConv1d[32]	0.96	32	2.62
ECAPA-TDNNLite[10]	1.2	32	3.07

We interpret this performance gap as the diversity of centroids' selection strategies. The imbalance between quantization value density and weight distribution degrades the accuracy of model. As shown in Figure 2, linear quantization (e.g., ADMM[18]) allocates few centroids for the region nearby 0 where most parameters exist. PoT quantization[19] deploys too many centroids in the area near 0. As a comparison, KMQAT takes into account the parameters of all positions in the distribution. The distribution of KMQAT centroids is relatively close to the parameter distribution of the pretrained model, thus bringing better performance.

Figure 3: *The quantization precision number of each layer of ResNet34 in mixed precision quantization.*



5.3. Analysis of mixed-precision quantization

We discuss the performance improvement of mixed-precision quantization in this section. As shown in Figure 3, shallower convolution layers require higher precision and deeper layers are less sensible to quantization operation. First several convolution layers are critical in the data processing, the lower precision of initial layers may cause irreversible performance degradation. Through reasonable allocation of weight precision and appropriate training recipe, our mixed-precision quantization model gains additional improvement.

6. Conclusions

This paper introduces a novel quantization method KMQAT and mixed-precision quantization for speaker verification system. We realized lossless 4 bit quantization of ResNet34. Our approach outperforms previous model compression and model quantization methods in terms of model size and accuracy. Experiments on Voxceleb prove that mixed-precision quantization with multi-stage fine-tuning further improves the performance of quantization model. In addition, we analyze the advantages of KMQAT centroids distribution and mixed-precision quantization algorithm.

7. Acknowledgement

This work was supported in part by China NSFC projects under Grants 62122050 and 62071288, and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102.

8. References

[1] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Proc. Interspeech 2017*, 2017, pp. 999–1003.

[3] B. Liu, Z. Chen, S. Wang, H. Wang, B. Han, and Y. Qian, "DF-ResNet: Boosting Speaker Verification Performance with Depth-First Design," in *Proc. Interspeech 2022*, 2022, pp. 296–300.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 3830–3834.

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

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

[7] J. Luo, J. Wu, and W. Lin, "Thinet: A filter level pruning method for deep neural network compression," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5058–5066.

[8] P. Molchanov, A. Mallya, S. Tyree, I. Frosio, and J. Kautz, "Importance estimation for neural network pruning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 264–11 272.

[9] J. Balian, R. Tavarone, M. Poumeyrol, and A. Coucke, "Small footprint text-independent speaker verification for embedded systems," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6179–6183.

[10] Q. Li, L. Yang, X. Wang, X. Qin, J. Wang, and M. Li, "Towards lightweight applications: Asymmetric enroll-verify structure for speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7067–7071.

[11] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 525–542.

[12] J. Xu, X. Chen, S. Hu, J. Yu, X. Liu, and H. Meng, "Low-bit quantization of recurrent neural network language models using alternating direction methods of multipliers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7939–7943.

[13] C. Leng, Z. Dou, H. Li, S. Zhu, and R. Jin, "Extremely low bit neural network: Squeeze the last bit out with admm," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[14] Y. Izui and A. Pentland, "Analysis of neural networks with redundancy," *Neural Computation*, vol. 2, no. 2, pp. 226–238, 1990.

[15] Y. Cheng, F. Yu, R. Feris, S. Kumar, A. Choudhary, and S. Chang, "An exploration of parameter redundancy in deep networks with circulant projections," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2857–2865.

[16] T. Zhu, X. Qin, and M. Li, "Binary Neural Network for Speaker Verification," in *Proc. Interspeech 2021*, 2021, pp. 86–90.

[17] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," in *Low-Power Computer Vision*. Chapman and Hall/CRC, 2021, pp. 291–326.

[18] J. Xu, J. Yu, S. Hu, X. Liu, and H. Meng, "Mixed precision low-bit quantization of neural network language models for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3679–3693, 2021.

[19] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen, "Incremental network quantization: Towards lossless cnns with low-precision weights," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[20] S. Wang, Y. Yang, T. Wang, Y. Qian, and K. Yu, "Knowledge distillation for small foot-print deep speaker embedding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6021–6025.

[21] B. Liu, H. Wang, Z. Chen, S. Wang, and Y. Qian, "Self-knowledge distillation via feature enhancement for speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7542–7546.

[22] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In Defence of Metric Learning for Speaker Recognition," in *Proc. Interspeech 2020*, 2020, pp. 2977–2981.

[23] Z. Dong, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "Hawq: Hessian aware quantization of neural networks with mixed-precision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[24] Z. Dong, Z. Yao, D. Arfeen, A. Gholami, M. Mahoney, and K. Keutzer, "Hawq-v2: Hessian aware trace-weighted quantization of neural networks," *Advances in neural information processing systems*, vol. 33, pp. 18 518–18 529, 2020.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2616–2620.

[26] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: deep speaker recognition," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 1086–1090.

[27] T. Ko, V. Peddinti, D. Povey, M. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[28] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[29] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[30] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 74–81.

[31] F. Li, B. Zhang, and B. Liu, "Ternary weight networks," *arXiv preprint arXiv:1605.04711*, 2016.

[32] L. Cai, Y. Yang, X. Chen, W. Tu, and H. Chen, "CS-CTCSCONV1D: Small footprint speaker verification with channel split time-channel-time separable 1-dimensional convolution," in *Proc. Interspeech 2022*, 2022, pp. 326–330.