Interspeech 2024
1-5 September 2024, Kos, Greece

Outlier Reduction with Gated Attention for Improved Post-training Quantization in Large Sequence-to-sequence Speech Foundation Models

Dominik Wagner[1], *Ilja Baumann*[1], *Korbinian Riedhammer*[1], *Tobias Bocklet*[1,2]

[1]Technische Hochschule Nürnberg Georg Simon Ohm
[2]Intel Labs

dominik.wagner@th-nuernberg.de

Abstract

This paper explores the improvement of post-training quantization (PTQ) after knowledge distillation in the Whisper speech foundation model family. We address the challenge of outliers in weights and activation tensors, known to impede quantization quality in transformer-based language and vision models. Extending this observation to Whisper, we demonstrate that these outliers are also present when transformer-based models are trained to perform automatic speech recognition, necessitating mitigation strategies for PTQ. We show that outliers can be reduced by a recently proposed gating mechanism in the attention blocks of the student model, enabling effective 8-bit quantization, and lower word error rates compared to student models without the gating mechanism in place.

Index Terms: post-training quantization, Whisper, gated attention, outliers

1. Introduction

Foundation models are deep neural networks trained on extensive and diverse datasets, capable of addressing a wide range of downstream tasks with minimal or no adaptation required. Speech foundation models (SFMs) like wav2vec 2.0 [1], HuBERT [2], and Whisper [3] have demonstrated remarkable performance across various speech-related tasks. However, improved performance also led to a notable increase in both the number of parameters and computational complexity. The increased demand for computational resources not only results in higher energy consumption but also limits the accessibility of SFM-based applications on resource-constrained devices. Consequently, there is a growing focus on enhancing the efficiency of SFMs [4, 5, 6, 7, 8].

Common approaches to reduce the size of automatic speech recognition (ASR) systems include knowledge distillation (KD) [8, 9, 10], model quantization [11, 12, 13, 14, 7], weight pruning [15, 16, 17], or combinations thereof [18, 19]. Two widely used quantization techniques are quantization-aware training (QAT) and post-training quantization (PTQ), with the later being the focus of this study. QAT involves simulating quantization operations during model training. PTQ methods are typically easier to apply, as they either require no direct interaction with the model or involve passing only a small calibration dataset through the model to determine the optimal quantization parameters.

Prior research has shown that weight quantization has minimal effect on the accuracy of transformer-based systems [20, 21], whereas activation tensors exhibit significantly wider value ranges [22, 23], making them more difficult to quantize.

Several efforts have been made to reduce the size of the Whisper SFM, a model also utilized in this study. These approaches either focus exclusively on KD [4, 24], QAT [6], or apply PTQ to calibrate personalized quantization schemes for particular speakers [7].

Analyzing the intricacies of SFMs is crucial for optimizing their performance, especially in resource-constrained environments. By understanding why models perform less effectively after quantization, we may gain insights into the determinants of their limitations.

Investigations into the behavior of transformer-based language models indicate that these models learn outliers within their weights and activation tensors [25, 26, 21, 27, 20, 28, 23]. Outliers typically appear within a limited, fixed set of hidden dimensions but materialize across various layers irrespective of the input sequence. Moreover, these outliers influence the quality of the model's predictions, and attempts to mitigate their impact, e.g., by dropping the corresponding values, can result in a significant degradation of the model's performance [27]. The presence of outliers within hidden dimensions also poses challenges for model quantization due to the trade-off between rounding errors and clipping errors [21, 20, 28]. Previous works argue that these anomalies stem from specific behaviors exhibited by attention heads attempting to either learn a null operation or a partial update of the residuals [29, 20, 28]. To obtain the zero values necessary for a non-update scenario in the attention matrix, the softmax input undergoes continual amplification during training, resulting in outliers in other network components.

So far, investigations into outliers have been focused on transformer-based language and vision models. In this work, we show that the observations made in the text and image domain also translate to the speech domain and that quantizing both weight and activation tensors to 8-bit benefits from outlier mitigation. In particular, we employ the gating mechanism for attention blocks introduced in [28], and demonstrate that Whisper-based models distilled with this gating mechanism in place learn smaller outliers and exhibit less performance degradation after quantization.

Knowledge distillation provides a robust framework for leveraging the knowledge encapsulated within the pretrained Whisper model, while also providing freedom in the choice of the student architecture. Given the unavailability of the original training data and the significant computational resources required for Whisper pretraining, we opt for student-teacher training using a diverse dataset of $\sim$16k hours of English speech collected from various publicly available sources comprising a large number of speakers and speaking styles to transfer knowledge to the student. Our focus is on analyzing outliers and evaluating the effectiveness of gated attention in mitigating them, along with improving ASR performance after PTQ.

10.21437/Interspeech.2024-2105

Our main contributions are:

- Identification of outlier behavior in Whisper, mirroring findings in transformer-based language and vision models
- Application of a gated attention mechanism to effectively address outliers in hidden dimensions of SFMs
- Creation of several smaller distilled versions of Whisper using student-teacher training to reduce the size of both encoder and decoder
- Demonstration of a practical framework for post-training quantization of SFMs to INT8, enhancing ASR efficiency and deployability

2. Method

2.1. Knowledge distillation

Knowledge distillation (KD) [30] is a learning framework that entails training a more compact student model to emulate the performance of a larger teacher model. KD involves training by aligning the student's predictions with those of the teacher. Following [31], we utilize a linear combination of cross-entropy loss across a set of N target labels $\boldsymbol{y}_{1:N} = \{y_1 \ldots, y_N\}$ and the Kullback-Leibler (KL) divergence [32] to optimize the student:

$$\mathcal{L} = \alpha_{\text{CE}} \left(-\sum_{i=1}^{N} p(y_i \mid \boldsymbol{y}_{<i}, \boldsymbol{H}_{1:M}) \right) + \alpha_{\text{KL}} \left(KL\left(\mathbf{S}, \mathbf{T}\right) \right), \tag{1}$$

where $\mathbf{S}$ and $\mathbf{T}$ denote the output probability distribution of the student and the teacher model, respectively. The constant weighting factors are set to $\alpha_{\text{CE}} = 1$ and $\alpha_{\text{KL}} = 0.8$ based on [4].

2.2. Post-training quantization

We consider post-training quantization (PTQ), where a trained full precision (FP32) model is converted into an 8-bit (INT8) fixed-point model directly without any additional training. Quantization is the process of mapping the values of a tensor $\mathbf{x} \in \mathbb{R}$ to corresponding values on an integer grid $\mathbf{x}_q \in \mathbb{Z}$. Following [23, 28], we emulate the quantization process according to [33]:

$$\mathbf{x}_q = s \cdot \left(\min\left(\max\left(\left\lfloor \frac{\mathbf{x}}{s} \right\rceil + z, 0 \right), 2^b - 1 \right) - z \right), \tag{2}$$

where $\mathbf{x}$ represents either model weights or activation tensors, $s \in \mathbb{R}_+$ is a scaling factor specifying the quantization step size, $b \in \mathbb{N}$ is the target bitwidth, $z \in \mathbb{Z}$ is the zero-point, and $\lfloor \cdot \rceil$ indicates rounding to the nearest integer.

The quantization procedure in our experiments encompasses both weight and activation quantization to INT8. Since activations are dependent on the input data, a critical aspect of the PTQ process for activation tensors involves identifying appropriate minimum and maximum values for each quantizer (i.e., the scaling factor s applied to the full-precision values). Several methods are available for establishing the boundaries of the interval. We employ a static approach, which estimates the quantization range based on 16 batches from the validation set utilizing an exponential moving average of the minimum and maximum values across those batches [34]. For weight quantization, we use the full range of the weight tensors.

All weights in every layer of the student's encoder and decoder blocks are quantized. Despite the presence of gating mechanisms in each encoder layer, we found that the last encoder layer retains a relatively large dynamic range, resulting in imprecise INT8 representation and consequently higher word error rates. Therefore, for activation quantization, all layers of the student model, except for the final output projection in the last encoder layer along with the final layer norm, are quantized.

2.3. Gated attention

A recently proposed conditional gating method that helps controlling the update process of hidden representations has been shown to be effective for reducing outliers in transformer-based language and vision models [28]. This approach allows the model to selectively retain or nullify updates to the representation of specific tokens, independent of the attention probabilities and values.

In [28], a gating function $\mathcal{G}$ is activated through a sigmoid nonlinearity σ and subsequently multiplied with the attention outputs using the Hadamard product:

$$\mathcal{A}(\mathbf{x}) := \text{softmax}\left(\frac{\boldsymbol{Q}(\mathbf{x})\boldsymbol{K}(\mathbf{x})^T}{\sqrt{d}} \right) \boldsymbol{V}(\mathbf{x}) \tag{3}$$

$$\text{gated_att}(\mathbf{x}) := \sigma(\mathcal{G}(\mathbf{x})) \odot \mathcal{A}(\mathbf{x}) \tag{4}$$

$\mathcal{A}(\mathbf{x})$ is the self-attention mechanism defined in [35] with the trainable linear projections $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$, as well as an input $\mathbf{x}$. Whisper employs multi-headed self-attention, in which the feature representations are divided into n parts of dimensionality d. The attention mechanism is applied to each n and concatenated to the final output. $\mathcal{G}$ is a neural network with a single linear layer, trained along with the rest of the model. We substitute Equation 3 with Equation 4 across all encoder and decoder layers of the student model and use a single $\mathcal{G}$ that is shared across different attention heads but not across different token positions. The authors of [28] use one gating function per attention head that is shared across different token position in their main experiments. They also proposed an additional method for outlier mitigation called clipped softmax. We also tried these methods in preliminary investigations but found that the gated attention mechanism was more effective than clipped softmax and gating on a per-head basis did perform worse than sharing the gate across attention heads.

2.4. Whisper

Whisper [3] is a family of transformer-based sequence-to-sequence [35] SFMs trained to perform multiple tasks such as multilingual ASR, language identification, and speech translation. The models are trained on ∼680k hours of proprietary data retrieved from the world wide web and are available in five sizes ranging from 39M parameters to 1.55B parameters. All Whisper SFMs utilize an encoder-decoder structure but differ in parameters such as the number of transformer blocks, the number of attention heads, and hidden layer dimensions.

The Whisper encoder $\mathcal{E}$ maps a sequence of L log-Mel spectrogram features $\boldsymbol{F}$ obtained from the raw audio waveform $\boldsymbol{A}$: $\boldsymbol{F}(\boldsymbol{A})_{1:L} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_L\}$ to a sequence of M hidden representations $\boldsymbol{H}_{1:M}$:

$$\mathcal{E} : \boldsymbol{F}(\boldsymbol{A})_{1:L} \mapsto \boldsymbol{H}_{1:M}.$$

The decoder predicts the probabilities for the next token y_i, based on the preceding tokens tokens $\boldsymbol{y}_{<i}$ and the hidden representations $\boldsymbol{H}_{1:M}$: $p(y_i \mid \boldsymbol{y}_{<i}, \boldsymbol{H}_{1:M})$. The model is trained on pairs of log-Mel spectrogram features and target transcriptions, using the cross-entropy objective.

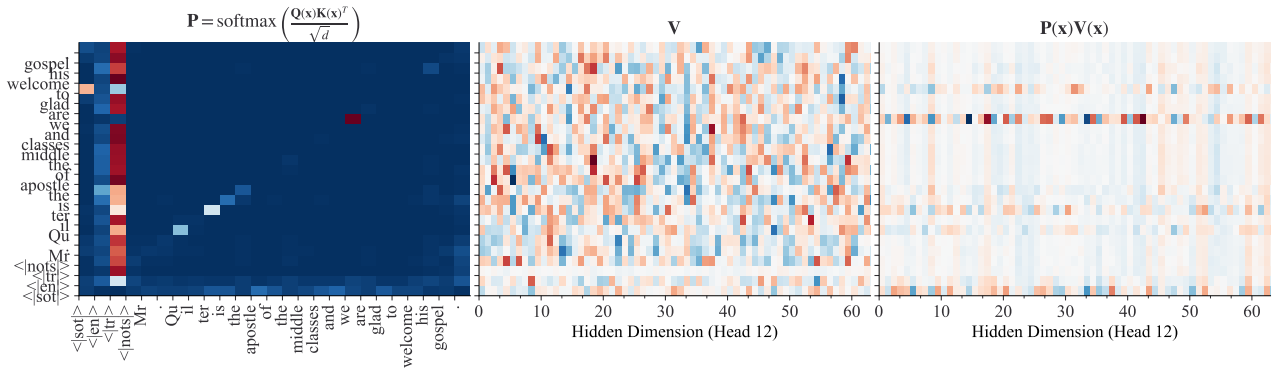


Figure 1: *Behavior of the self-attention mechanism in the pretrained Whisper model (`whisper-large-v2`) computed for the first example of the LibriSpeech test-clean set. The left matrix shows the attention probabilities* $\mathbf{P}$*, where* $d = 64$ *is the dimensionality of the attention head. The middle matrix are the values* $\mathbf{V}$ *in the twelfth attention head. The right matrix is the product of the two.*

We distill the 1.55B parameter version of Whisper (`whisper-large-v2`) into three variants using either 8, 16 or 24 encoder layers instead of the original 32 layers. Based on findings that drastically reducing or even eliminating the Whisper decoder does not significantly compromise performance, we reduce its size to 2 layers [5, 4]. The total number of trainable parameters for the resulting student model variants are 283M, 440M, and 598M parameters, respectively.

3. Experiments and Results

3.1. Data

We used a combination of four publicly available English datasets as our training corpus: People's Speech [36], Common-Voice (version 16) [37], LibriSpeech [38], and Voxpopuli [39]. The combined training data comprises approximately $\sim$16k hours of speech sourced from diverse domains such as audiobooks, political speeches, interviews, and narrated Wikipedia articles. In addition to evaluating ASR performance on in-distribution (ID) data comprised of the test portions of the training corpora, we employed three out-of-distribution (OOD) test sets to further asses the robustness of the gated attention approach towards PTQ: TED-LIUM [40], Fleurs [41], and Gigaspeech [42]. These datasets encompass diverse speech characteristics and additional domains such as podcasts and talks.

3.2. Evaluation metrics

To assess ASR performance, we examined word error rates (WERs) across both ID and OOD test sets. We evaluated ASR performance for differently sized student models with and without INT8 quantization on weights and activations using either coventional attention or the gating mechanism. Following [28], we also analyzed outliers measured by kurtosis and infinity norm $\|\cdot\|_\infty$, exploring the impact of gated attention across different model sizes. Kurtosis was averaged across the outputs of all attention layers. Based on [20], we count values that exceed six standard deviations from the mean of the activation tensor as outliers.

3.3. Modeling and architecture details

Leveraging the shared dimensionality of hidden layers between teacher and student networks, we initialized each student using the pre-trained weights of the teacher [43, 44, 4]. In particular, we selectively copied the weights of 8, 16 or 24 layers from the teacher encoder. For example, when the number of layers in the encoder was reduced by a factor of 4 (i.e., 8 layers instead of 32), we copied every 4$^{\text{th}}$ weight matrix from the teacher starting with the first one. In the student decoder, we copied the weights from the initial and final layers of the teacher.

Each model was trained for 2×10^5 steps with an effective batch size of 64, which amounts to two training epochs. We used the AdamW optimizer [45] ($\lambda = 10^{-4}$, $\epsilon = 10^{-8}$, $\beta_1 = 0.99$, $\beta_2 = 0.999$) with an initial learning rate of 10^{-4}, a linear schedule and a warm-up phase of 1000 steps. The models used in the evaluation were selected based on the lowest WER achieved on the validation set comprised of all validation portions from the four datasets used for training. We utilized greedy decoding for its increased inference speed.

3.4. Whisper exhibits outlier behavior similar to LMs

As an initial analysis, we followed [28] and visualized the behavior of the multi-headed self-attention mechanism in the Whisper decoder. The large version of Whisper employs 20 attention heads with 64 hidden dimensions each. Figure 1 illustrates the attention mechanism in the 31$^{\text{st}}$ layer in the decoder part of the pretrained 1.55B parameter version of Whisper. We observe that the attention head predominantly allocates its probability mass to the transcription token `<|tr|>`, while the same token has small values associated with it in $\mathbf{V}$ (cf. third row from the bottom in the middle matrix). As a result, the product between the two is small, representing only a minimal or no update of the hidden representation. Only a small portion of the probability mass is distributed to other tokens (in this case mostly the token `we`), resulting in a local update of the hidden representation for those tokens. Similar patterns can be found across all decoder layers and attention heads. These results are in line with the findings on transformer-based language and vision models [46, 28], supplementing them with SFMs.

3.5. Gated attention improves PTQ in student models

Figure 2 illustrates which hidden dimensions contribute the most outliers for one student model trained without gated attention (left) and another one trained with gated attention (right). Outliers were counted separately for each dimension in the output projection of the self-attention block in the last decoder layer (i.e., the 2$^{\text{nd}}$ decoder layer). We see that the distribution of outliers becomes more uniform and each dimension contributes less to the overall outliers with the gating mechanism in place. For example, with conventional attention, the hidden dimensions #819 and #1054 contribute $\sim$15% of all outliers, whereas with gated attention the two highest shares (dimension #287 and

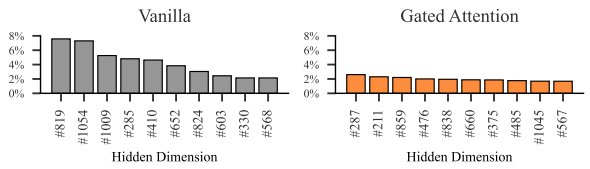


Figure 2: *Top 10 shares of activation outliers per hidden dimension at the output projection of the self-attention block of the last decoder layer in trained student models. Left are the relative outliers for a student trained without gated attention and right are the relative outliers with gated attention. Both models were trained using 24 layers for the encoder. The hidden dimensions are zero-indexed. Each output projection layer has a dimensionality of 1280.*

Table 1: *Average kurtosis and maximum infinity norm on the full ID test set for different model sizes before quantization.*

Decoder Layers	Gated Attention	Average Kurtosis	Maximum Inf. Norm
24	✗	105.4	66.8
	✓	42.8	44.0
16	✗	48.5	35.3
	✓	19.5	29.8
8	✗	28.6	29.9
	✓	20.7	23.7

#211) amount only to a total of ∼5% relative to all outliers.

Table 1 shows the average kurtosis and $\|\cdot\|_\infty$ across test sets with and without gated attention. Consistently lower outlier metrics were observed for models trained with gated attention compared to those trained without gated attention, indicating improved robustness and stability in the former case.

Table 2 presents WERs on each ID and OOD test set for student models with 24, 16, and 8 encoder layers before and after INT8 quantization of weights and activations. Each model was trained once with the gated attention mechanism in place and once without. The 24-layer encoder system demonstrated the best overall performance across the test sets and WERs increased with fewer encoder layers used in the student model. Before quantization, using the gated attention mechanism during student-teacher training resulted in similar WERs compared to training without gated attention across all test sets. However, a notable difference emerges after quantization. With gated attention, Table 2 shows only slight increases in WERs relative to non-quantized models, whereas quantized models trained without gated attention experience significant increases in WERs across all test sets. This highlights the effectiveness of gated attention in maintaining ASR performance after quantization. Additionally, in some cases, the gating mechanism also improves the full precision performance. However, this is not consistent across all test sets and encoder layer configurations.

Our findings align with outlier analyses conducted for language and vision models [27, 20, 26, 28], showing that the importance of outlier mitigation is crucial to control performance degradation in PTQ.

4. Conclusions

In line with observations in transformer-based language and vision models, we found that outliers generated by the attention mechanism, attempting to perform no-update operations, also manifest in the Whisper model architecture. These outliers pose challenges when applying post-training quantization, particularly concerning activation quantization. We focused on reducing the model size through student-teacher training, followed by applying INT8 quantization to both weight and activation tensors. To enhance robustness against outliers, we used a gated attention mechanism that learns to perform the null operation when necessary. Our experiments demonstrated that the student model equipped with the gating mechanism achieved similar WERs as the model without gated attention at floating-point precision. However, when quantization was applied, models with gated attention exhibited significantly more stable WERs. This underscores the efficacy of gated attention in complementing student-teacher training, particularly in scenarios where the ultimate goal involves weight and activation quantization.

Table 2: *Word Error Rates on different in-distribution (ID) and out-of-distribution (OOD) test sets.*

Dataset	INT8 Quant.	Gated Attention	WER 24 layer	WER 16 layer	WER 8 layer
Voxpopuli (ID)	✗	✗	13.7	15.6	19.2
		✓	12.6	13.8	18.1
	✓	✗	21.7	22.2	24.1
		✓	13.9	15.6	19.6
LibriSpeech test-clean (ID)	✗	✗	7.0	6.7	10.7
		✓	6.3	6.9	10.0
	✓	✗	9.7	12.7	13.7
		✓	7.7	8.6	12.3
LibriSpeech test-other (ID)	✗	✗	13.1	14.6	21.1
		✓	12.7	14.0	20.6
	✓	✗	18.4	23.2	24.0
		✓	14.4	15.8	22.8
CommonVoice 16 (ID)	✗	✗	23.6	27.6	36.5
		✓	22.4	26.2	36.3
	✓	✗	34.4	33.2	39.3
		✓	25.0	28.4	38.6
People's Speech (ID)	✗	✗	33.7	33.9	46.1
		✓	32.7	35.0	41.6
	✓	✗	38.7	45.6	47.6
		✓	35.5	37.6	44.7
TED-LIUM (OOD)	✗	✗	13.6	15.0	19.6
		✓	14.3	14.9	18.0
	✓	✗	19.7	22.4	27.1
		✓	15.5	17.2	19.8
Gigaspeech (OOD)	✗	✗	18.4	19.9	26.2
		✓	17.9	20.0	25.7
	✓	✗	23.1	25.7	28.5
		✓	19.2	21.5	27.8
Fleurs (OOD)	✗	✗	17.1	17.7	25.2
		✓	15.8	18.6	25.5
	✓	✗	25.2	21.3	28.9
		✓	17.4	20.5	28.1

5. Acknowledgments

We gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b196ac14. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. This work was supported by the Bavarian State Ministry of Science and the Arts under grant H.2-F1116.NÜ/61/2.

6. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, p. 3451–3460, 2021.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," 2022, arXiv:2212.04356.

[4] S. Gandhi, P. von Platen, and A. M. Rush, "Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling," 2023, arXiv:2311.00430.

[5] Y. Peng, Y. Sudo, M. Shakeel, and S. Watanabe, "OWSM-CTC: An open encoder-only speech foundation model for speech recognition, translation, and language identification," 2024, arXiv:2402.12654.

[6] H. Shao, W. Wang, B. Liu, X. Gong, H. Wang, and Y. Qian, "Whisper-KDQ: A lightweight whisper via guided knowledge distillation and quantization for efficient ASR," 2023, arXiv:2305.10788.

[7] E. Fish, U. Michieli, and M. Ozay, "A Model for Every User and Budget: Label-Free and Personalized Mixed-Precision Quantization," in *Interspeech*, 2023, pp. 3232–3236.

[8] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *ICASSP*, 2022, pp. 7087–7091.

[9] X. Gong, Z. Zhou, and Y. Qian, "Knowledge Transfer and Distillation from Autoregressive to Non-Autoregessive Speech Recognition," in *Interspeech*, 2022, pp. 2618–2622.

[10] T. Ashihara, T. Moriya, K. Matsuura, and T. Tanaka, "Deep versus Wide: An Analysis of Student Architectures for Task-Agnostic Knowledge Distillation of Self-Supervised Speech Models," in *Interspeech*, 2022, pp. 411–415.

[11] C.-F. Yeh, W.-N. Hsu, P. Tomasello, and A. Mohamed, "Efficient speech representation learning with low-bit quantization," 2022, arXiv:2301.00652.

[12] S. Ding, P. Meadowlark, Y. He, L. Lew, S. Agrawal, and O. Rybakov, "4-bit Conformer with Native Quantization Aware Training for Speech Recognition," in *Interspeech*, 2022, pp. 1711–1715.

[13] K. Zhen, H. D. Nguyen, R. Chinta, N. Susanj, A. Mouchtaris, T. Afzal, and A. Rastrow, "Sub-8-Bit Quantization Aware Training for 8-Bit Neural Network Accelerator with On-Device Speech Recognition," in *Interspeech*, 2022, pp. 3033–3037.

[14] A. Fasoli *et al.*, "4-Bit Quantization of LSTM-Based Speech Recognition Models," in *Interspeech*, 2021, pp. 2586–2590.

[15] T. Moriya, H. Sato, T. Tanaka, T. Ashihara, R. Masumura, and Y. Shinohara, "Distilling attention weights for ctc-based asr systems," in *ICASSP*, 2020, pp. 6894–6898.

[16] V. S. Lodagala, S. Ghosh, and S. Umesh, "PADA: Pruning assisted domain adaptation for self-supervised speech representations," in *SLT*, 2023, pp. 136–143.

[17] C.-I. J. Lai *et al.*, "PARP: Prune, adjust and re-prune for self-supervised speech recognition," in *NeurIPS*, vol. 34, 2021, pp. 21 256–21 272.

[18] J. Kim, S. Chang, and N. Kwak, "PQK: Model Compression via Pruning, Quantization, and Knowledge Distillation," in *Interspeech*, 2021, pp. 4568–4572.

[19] S. Ding *et al.*, "USM-Lite: Quantization and sparsity aware fine-tuning for speech recognition with universal speech models," 2024, arXiv:2312.08553.

[20] Y. Bondarenko, M. Nagel, and T. Blankevoort, "Understanding and overcoming the challenges of efficient transformer quantization," in *EMNLP*, Nov. 2021, pp. 7947–7969.

[21] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "GPT3.int8(): 8-bit matrix multiplication for transformers at scale," in *NeurIPS*, 2022.

[22] Z. Yao, R. Yazdani Aminabadi, M. Zhang, X. Wu, C. Li, and Y. He, "Zeroquant: Efficient and affordable post-training quantization for large-scale transformers," in *NeurIPS*, vol. 35, 2022, pp. 27 168–27 183.

[23] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, "SmoothQuant: Accurate and efficient post-training quantization for large language models," in *ICML*, 2023.

[24] T. P. Ferraz, M. Z. Boito, C. Brun, and V. Nikoulina, "Multilingual DistilWhisper: Efficient distillation of multi-task speech models via language-specific experts," 2024, arXiv:2311.01070.

[25] Z. Luo, A. Kulmizev, and X. Mao, "Positional artefacts propagate through masked language model embeddings," in *ACL*, 2021, pp. 5312–5327.

[26] X. Wei, Y. Zhang, X. Zhang, R. Gong, S. Zhang, Q. Zhang, F. Yu, and X. Liu, "Outlier suppression: Pushing the limit of low-bit transformer language models," in *NeurIPS*, vol. 35, 2022, pp. 17 402–17 414.

[27] O. Kovaleva, S. Kulshreshtha, A. Rogers, and A. Rumshisky, "BERT busters: Outlier dimensions that disrupt transformers," in *ACL-IJCNLP*, 2021, pp. 3392–3405.

[28] Y. Bondarenko, M. Nagel, and T. Blankevoort, "Quantizable transformers: Removing outliers by helping attention heads do nothing," in *NeurIPS*, 2023.

[29] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, "Revealing the dark secrets of BERT," in *EMNLP-IJCNLP*, 2019, pp. 4365–4374.

[30] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[31] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *EMNLP*, 2016, pp. 1317–1327.

[32] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951.

[33] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *CVPR*, 2018.

[34] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," 2018, arXiv:1806.08342.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *NeurIPS*, vol. 30, 2017.

[36] D. Galvez, G. Diamos, J. M. C. Torres, J. F. Cerón, K. Achorn, A. Gopi, D. Kanter, M. Lam, M. Mazumder, and V. J. Reddi, "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," in *NeurIPS Datasets and Benchmarks Track*, 2021.

[37] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," in *LREC*, 2020, pp. 4218–4222.

[38] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[39] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, 2021, pp. 993–1003.

[40] A. Rousseau, P. Deléglise, and Y. Estève, "TED-LIUM: an automatic speech recognition dedicated corpus," in *LREC*, 2012, pp. 125–129.

[41] A. Conneau *et al.*, "FLEURS: Few-shot learning evaluation of universal representations of speech," in *SLT*, 2023, pp. 798–805.

[42] G. Chen *et al.*, "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio," in *Interspeech*, 2021, pp. 3670–3674.

[43] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020, arXiv:1910.01108.

[44] S. Shleifer and A. M. Rush, "Pre-trained summarization distillation," 2020, arXiv:2010.13002.

[45] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[46] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? an analysis of BERT's attention," in *ACL Workshop BlackboxNLP*, 2019, pp. 276–286.