Interspeech 2024
1-5 September 2024, Kos, Greece

Sign Value Constraint Decomposition for Efficient 1-Bit Quantization of Speech Translation Tasks

*Nan Chen, Yonghe Wang, Feilong Bao**

College of Computer Science, Inner Mongolia University, China
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology
chennannlp@gmail.com, cswyh@imu.edu.cn, csfeilong@imu.edu.cn

Abstract

Speech-to-text translation is vital in converting speech input to text output in different languages. While combining speech and machine translation pre-trained models enhances translation quality, it also escalates the number of parameters, resulting in substantial hardware costs for model training and deployment. We propose a 1-bit quantized model based on Sign Value Constraint Decomposition (SVCD) for linear layers to address this challenge. SVCD approximates the weight matrix of the linear layer as a sign matrix and two trainable vectors, preserving higher information capacity at a minor space cost. Additionally, we utilize knowledge distillation to transfer the capability of the original fine-tuned model to the quantized model. The experimental results demonstrate the critical importance of the decoder's attention module in the performance of the quantized speech translation model. Our code is available at https://github.com/myaxxxxx/onebit-st.

Index Terms: speech-to-text translation, quantization

1. Introduction

Speech-to-text translation tasks [1, 2, 3] aim to convert speech input in one language into text output in another. In recent years, speech translation models [4, 5] that combine speech [6, 7, 3] and machine translation [8, 9, 10, 11] pre-trained models have been widely used. However, the combined use of speech and machine translation pre-trained models significantly increases the parameters of the speech translation model. This leads to two issues: 1) it increases the difficulty of training the speech translation model that is combined with pre-trained models [12]. Using a speech translation model that combines the Hubert speech pre-trained model with the M2M100 machine translation model, only the A100 and H100 GPU can train the model in FP16 mixed precision. 2) the excessive size of the models causes difficulties in storing, inferring, and deploying the models.

Recently, quantization methods have attracted widespread attention due to their effective reduction of computation and memory usage. There are mainly two types of quantization methods: post-training quantization (PTQ) [13, 14, 15], and Quantization Aware Training (QAT) [16, 17, 18, 19]. Currently, most quantization methods are PTQ, which is simple and easy to deploy without requiring retraining or changing the training process. However, low-bit PTQ often leads to significant performance degradation. The latest PTQ method compresses parameter weights to at least 3-bit values, but when the quantization bandwidth is reduced to 1-bit or 2-bit values, the performance of the model drops sharply. We aim to break through this bottleneck and achieve quantization with even lower bits.

Compared to PTQ, QAT achieves better results because QAT-based models allow continuous fine-tuning during training to improve performance further. In recent QAT work, Bitnet [18] reduces the bit bandwidth to 1 bit. However, the challenge faced by QAT is that optimizing the model is relatively difficult. Onebit [19] finds that when the quantization bandwidth is particularly low, the performance drops sharply, and the training process becomes unstable.

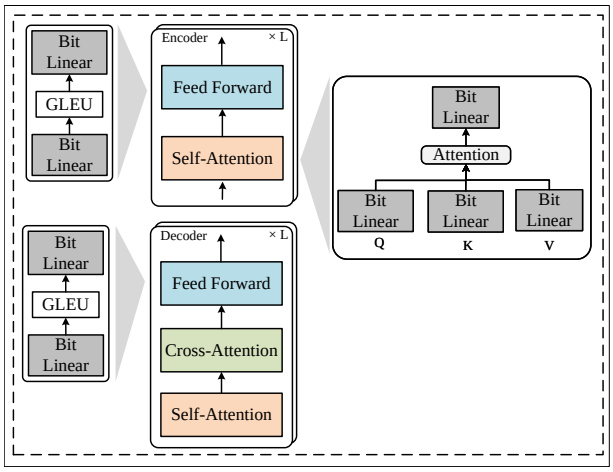

Figure 1: *Overview of our method.*

How do we balance between extremely low-bit quantization and quantization model performance? In this paper, we propose a one-bit linear layer based on Sign Value Constraint Decomposition (SVCD) to achieve this balance and maintain training stability. In SVCD, we approximate the weight matrix of the linear layer as a sign matrix and two trainable vectors. The sign matrix maintains the high rank of the original weight matrix at a more minor space cost, thus preserving higher information capacity. The two trainable vectors provide additional trainable parameters, storing more weight information of the linear layer at a lower cost and making the model easier to train. Additionally, we use knowledge distillation [20, 21] to transfer the capability of original fine-tuned model to the quantized model. The experimental results indicate that in quantizing speech translation, the attention module of decoder is crucial. When quantizing the attention module of the decoder, the model's performance sharply declines. Additionally, the model's performance decreases minimally by approximately 1% as the quantizing the encoder of the machine translation model in the speech translation model. The model's performance decreases by approximately 5.5% when quantizing the speech pre-training model.

The main contributions of this paper are summarized as follows:

10.21437/Interspeech.2024-2022

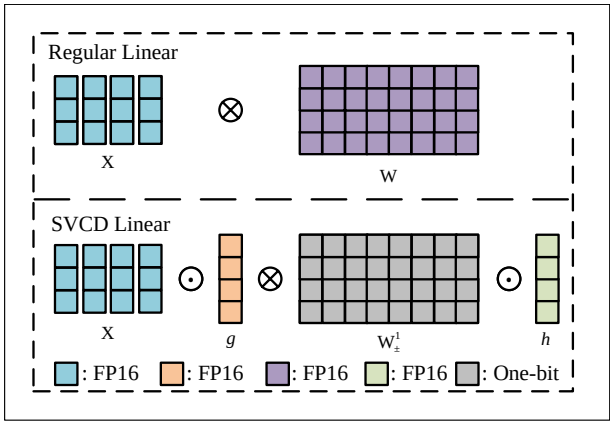

Figure 2: *The overview of our SVCD. The upper part is the original Linear. The lower part is our SVCD.* $\odot$ *represents the Hadamard product operation. the weight matrix consists of ±1 instead.* g *and* h *represents the two trainable vectors.*

- We propose a 1-bit quantized model that effectively reduces the storage space required for training and loading the model. To our knowledge, we are the first to explore extreme-low-bit quantization in the context of pre-trained model-based speech translation tasks.
- Building on quantized models, we introduce Sign Value Constraint Decomposition (SVCD) for quantizing linear layers. We balance low-bit quantization and model performance by adding two trainable parameter vectors.

2. Approach

2.1. One-bit Linear Layer

The core idea of model quantization is to compress the weight matrix W in linear layers into a low-bit form, reducing the model's storage space and computational cost. Traditional quantization methods often use the round-to-nearest (RTN) method, which rounds the weight w to the nearest value in the quantization grid. However, this method may lead to precision loss. To address this issue, we introduce two additional learnable parameter vectors to adjust the bias and scaling factors in the quantization process. We first binarize the weights to either +1 or 1 with the signum function:

$$\operatorname{Sign}\left(\mathbf{W_{ij}}\right)=\begin{cases}+1, & \text{if } \mathbf{W_{ij}}>0\\ -1, & \text{if } \mathbf{W_{ij}}\leq 0\end{cases}. \tag{1}$$

where $\mathbf{W_{ij}}$ represents the weight value at the ith row and jth column and Sign$(\cdot)$ functions return the sign matrix. The quantized linear layer we propose is as follows:

$$\mathbf{W}_{\pm 1}=\operatorname{Sign}(\mathbf{W}). \tag{2}$$

where $\mathbf{W}$ denotes the quantized weight matrix with the shape $m \times n$ and $W_{\pm 1}$ denotes the 1-bit quantized matrix. Next, we introduce trainable vectors g and h into the forward process:

$$\mathbf{Y}=\left[(\mathbf{X}\odot \mathbf{g})\mathbf{W}_{\pm 1}^{\mathrm{T}}\right]\odot \mathbf{h}. \tag{3}$$

where $\odot$ represents the Hadamard product operation and $\mathbf{Y}$ represents the outputs of our quantized linear. One-bit [19] discover that models might experience floating-point overflow during the QAT process. As depth increases, the activation can become progressively larger. Therefore we use Post-LayerNorm instead of Pre-LayerNorm:

$$\mathbf{Z}=\operatorname{LayerNorm}(\mathbf{Y}). \tag{4}$$

where $\mathbf{Z}$ is the final input of the quantized linear layer. As $\mathbf{g}$ and $\mathbf{h}$ are randomly initialized, to constrain their initialization values, we follow the initialization method of bias in linear layers and constrain the initialization of tensors $\mathbf{g}$ and $\mathbf{h}$ to a uniform distribution. Unlike BitNet and One-Bit, our method not only adds two trainable parameter vectors but also constrains the initialization of these two vectors. This constrained initialization method can help stabilize the training process of the model, improve the convergence speed, and enhance the performance of the model.

In our proposed SVCD, the original parameter matrix $\mathbf{W}$ is decomposed into Sign and two learnable vectors. For $\mathbf{g}$ and $\mathbf{h}$, we can approximate them as the rank-1 approximation of the $\mathbf{g}$ matrix and $\mathbf{h}$ matrix of the singular value decomposition (SVD) of matrix $\mathbf{W}$. Therefore, $\mathbf{W}$ can be represented as:

$$\mathbf{W}\approx \mathbf{W}_{\text{sign}}\odot\left(\mathbf{a}\mathbf{b}^{\mathrm{T}}\right). \tag{5}$$

In the forward propagation process of the quantization model, given the weight matrix W and the output X, the linear layer can be reformulated as follows:

$$\mathbf{X}\mathbf{W}^{\mathrm{T}}\approx\left[\left(\mathbf{X}\odot\mathbf{b}^{\mathrm{T}}\right)\mathbf{W}_{\text{sign}}^{\mathrm{T}}\right]\odot\mathbf{a}^{\mathrm{T}}. \tag{6}$$

2.2. knowledge Distillation

During training, in addition to the cross-entropy loss, we employ knowledge distillation to transfer knowledge from a 32-bit unquantized teacher model to a 1-bit quantized student model. The hidden states of the teacher model, computed with full precision based on mean squared error, are used to guide the training of the student model.

$$\mathcal{L}_{\mathrm{MSE}}=-\frac{1}{n_s}\sum_{i=1}^{n_s}\left(P^{\mathcal{T}}\left(\boldsymbol{x}_i\right)-P^{\mathcal{S}}\left(\boldsymbol{x}_i\right)\right)^2. \tag{7}$$

where n_s denotes the number of training samples in the current batch. $\mathcal{T}$ and $\mathcal{S}$ are the teacher model and student model, respectively. Hence the final objective function can be formulated as:

$$\mathcal{L}_{\mathrm{KD}}=\mathcal{L}_{\mathrm{CE}}+\alpha\mathcal{L}_{\mathrm{MSE}}. \tag{8}$$

where α is the hyper-parameter that balances the importance of the cross-entropy loss and the hidden features MSE loss.

3. Experiments

3.1. Datasets and Hyperparameters

Datasets: We conduct experiments on the MuST-C [22] dataset, which comprises 350 hours of TED talks. We select eight translation directions, including English (En) to German (De), French (Fr), Spanish (Es), Romanian (Ro), Russian (Ru), Italian (It), Portuguese (Pt), and Dutch (Nl). Each translation direction is accompanied by corresponding transcripts and translation text. We use the `tst-COMMON` set for testing.

Model settings: We use a speech translation architecture based on multi-task learning, which includes three parts: speech encoder, machine translation encoder, and decoder. We employ Hubert [6] as the speech pre-trained model and Deltalm [8] as

Table 1: *BLEU scores on MuST-C* `tst-COMMON` *set. We set the FP16 finetuning as baselines. All methods were trained in FP16 format. "-" indicates that Bit-net could not be trained due to gradient overflow during training and all the reported results are from three runs.*

Method	BLEU							
	De	Fr	Es	Ro	Ru	It	Pt	Nl
FP16	25.3	35.7	30.5	23.8	17.2	26.0	31.3	29.5
GPTQ [13] (2-bit)	16.25	23.46	21.11	14.45	9.42	17.25	21.31	20.09
LLM-QAT [16] (2-bit)	17.13	24.22	22.09	15.32	10.65	18.18	22.24	20.99
OmniQuant [17] (2-bit)	17.44	24.87	22.54	15.75	11.21	18.63	22.76	21.44
Bit-net [18] (1-bit)	-	-	-	-	-	-	-	-
One-bit [19] (1-bit)	14.43	22.34	18.43	12.31	6.42	15.52	20.19	17.59
Ours								
SVCD (Speech Encoder)	20.54	28.86	24.88	17.45	11.94	21.23	25.21	23.21
SVCD (MT Encoder)	**24.45**	**34.64**	**29.75**	**22.91**	**16.95**	**25.45**	**30.94**	**29.04**
SVCD (Speech & MT Encoder)	19.63	27.86	24.43	17.11	11.33	20.99	24.77	23.11

Table 2: *Compression ratio of speech translation models.*

Pre-trained Models		Full	One-bit	Ratio(%)
Speech Model	MT Model			
Base	Base	6GB	2.1GB	65%
Large	Large	63GB	19.4GB	69%

Table 3: *The performance of loss ablation study. CE represents the cross entropy loss. MSE represents the mean square error loss.*

No.	Module		BLEU	
	CE	MSE	En-De	En-fr
1	✘	✓	23.53	33.89
2	✓	✘	24.01	34.23
3	✓	✓	24.45	34.64

the machine translation model. We use Adam [23] for optimization due to its robustness. All models are trained on 4 Tesla A100 GPUs with fp16 and all models are evaluated by BLEU score [24]. For more details of hyper-parameter settings, please refer to https://anonymous.4open.science/r/onebit-st-7033.

3.2. Baselines and Evaluation Metric

We selected five baseline models for comparison. To the best of our knowledge, previous work has not explored 1-bit quantization in speech translation tasks. Given that speech translation tasks based on pre-trained models introduce machine translation pre-trained models and speech pre-trained models, we first selected GPTQ [13], LLM-QAT [16], and OmniQuant [17], which are related to quantization of pre-trained models, for comparison, and relaxed the quantization bits of these models to W2A16 (2-bit weight and 16-bit activation). Additionally, we compared them with popular OneBit [19, 18] baseline models, including one-bit [19] and bit-net [18] models, with quantization bits set to W1A16 (1-bit weight and 16-bit activation). For evaluation, All models are evaluated by BLEU [24] score.

4. Results and Analysis

4.1. Main Results

Table 1 showcases BLEU scores across various speech translation models, highlighting the impact of different quantization approaches. The FP16 model serves as the baseline with scores such as 25.3 in German (De), 35.7 in French (Fr), and 30.5 in Spanish (Es).

In contrast, the 2-bit quantized methods, including GPTQ, LLM-QAT, and OmniQuant, show that the performance drops significantly compared to the baseline. The SVCD approach is applied in three variations: Speech Encoder, MT Encoder, and both. The SVCD (MT Encoder) shows a decrease in performance after quantization, with reductions under 1% compared to the FP16 baseline, achieving scores such as 34.64 in Fr and 30.94 in Portuguese (Pt). The SVCD (Speech Encoder) experiences a more noticeable performance decrease, around 5%, yet this reduction comes with the advantage of significantly lower memory requirements.

Bit-net could not be evaluated due to training difficulties, as indicated by the dashes in the table. While the most memory-efficient, the One-bit quantization results in the lowest BLEU scores among the quantization methods, reinforcing the challenge of maintaining translation fidelity with aggressive quantization techniques. Overall, the results indicate that while all quantization methods result in some performance loss, SVCD (MT Encoder) provides the most promising approach for maintaining near-baseline accuracy with reduced memory usage.

4.2. The Efficiency of SVCD

Utilizing extreme-low-bit quantization for model weights can markedly diminish their memory requirements. As depicted in Table 2, a notable increase in compression ratio is observed with larger model sizes, which is especially beneficial for substantial models, enabling their accommodation within a single GPU despite a slight decline in performance. Moreover, reducing quantization to ±1 also enhances the efficiency of matrix multiplication operations on CPUs, as it allows the conversion of floating-point multiplications between matrix elements into quicker bitwise operations on these processors. Consequently, this significant decrease in memory consumption allows these low-bit speech translation models to fulfill the operational demands on desktops and mobile phones.

Table 4: *BLEU scores of different quantied modules.* ✓ *denotes quantifying this module.* ✘ *indicates training with FP16 mixed precision. Encoder represents the machine translation encoder in the multi task learning (MTL) architecture*

No.	Module		BLEU	
	Encoder	Decoder	En-De	En-fr
1	✘	✓	2.46	0.36
2	✓	✓	3.24	2.37
3	✓	✘	25.45	34.79

4.3. The Effectiveness of Knowledge Distillation

To investigate the impact of knowledge distillation on our proposed SVCD model, we conduct ablation experiments to validate the performance of models with and without knowledge distillation. Specifically, we further explore the influence of the teacher network's encoder outputs on the model. The experimental results are shown in 3.

Table 3 illustrates the results of a loss ablation study, as measured by BLEU scores for English-to-German (En-De) and English-to-French (En-Fr) language pairs. The study evaluates the effects of cross-entropy loss (CE) versus means square error loss (MSE) on the model's performance.

The results show that using MSE loss (No. 1) results in BLEU scores of 23.53 for En-De and 33.89 for En-Fr, indicating a substantial impact on performance. When switching to CE loss (No. 2), there is a slight decrease in the En-Fr score to 34.23 but a minor increase for En-De to 24.01. Notably, combining both CE and MSE losses (No. 3) achieves the highest BLEU scores of 24.45 for En-De and 34.64 for En-Fr, suggesting that integrating both types of loss can lead to the most effective translation model performance. These results highlight that while MSE can be effective, the traditional use of CE loss, especially when combined with MSE, is more advantageous for speech translation tasks.

4.4. The Effectiveness of the Decoder

Table 4 illustrates the impact of quantization on different modules of a machine translation architecture. The table demonstrates that quantizing the decoder module undermines model performance in translation tasks, as seen with the BLEU scores of 2.46 for English-to-German (En-De) and 0.36 for English-to-French (En-Fr), where only the decoder is quantized. In contrast, the BLEU score drop is minimal, where only the encoder is quantized. This indicates that the encoder's quantization has a comparatively negligible impact on performance, with scores remaining high at 25.45 for En-De and 34.64 for En-Fr. These results suggest that while the encoder can be quantized with little effect on translation quality, the decoder's quantization is detrimental and should be cautiously approached.

4.5. The Ablation of Decoder Modules

As shown in Table 4, the decoder component of the speech translation model is crucial. When the decoder module is quantized, the model's performance significantly decreases. To further investigate which decoder sub-modules greatly impact the model's quantization. Table 5 delineates the BLEU scores for different quantized modules within the decoder of a speech translation system. That the quantization of the cross-attention (Cross-Att) module leads to a catastrophic decline in model performance,

Table 5: *BLEU scores of different decoder quantied modules. FFN represents the feed-forward layer. Self-Att denotes the self-attention module in the decoder layer. Cross-Att represents the cross-attention module in the decoder layer.*

No.	Module			BLEU	
	Self-Att	Cross-Att	FFN	En-De	En-fr
1	✓	✘	✘	23.90	33.95
2	✘	✓	✘	2.37	0.09
3	✘	✘	✓	24.33	34.27
4	✓	✓	✘	0.31	1.54
5	✘	✓	✓	0.13	0.06

with BLEU scores plummeting to 2.37 for English-to-German (En-De) and 0.09 for English-to-French (En-Fr) when it is the only quantized module (No. 2). Conversely, quantizing either the self-attention (Self-Att) module or the feed-forward layer (FFN) independently (No. 1 and No. 3), has a far less severe impact on performance, yielding scores that are comparably close to non-quantized baselines with 23.90 and 24.33 for En-De, and 33.95 and 34.27 for En-Fr, respectively. These observations underscore the sensitivity of the cross-attention module to quantization and its crucial role in maintaining translation quality. In contrast, the other modules appear to be more robust to quantization. Quantization of Self-Att and FFN without Cross-Att (No. 4) also results in detrimental scores, further highlighting the Cross-Att module's unique importance.

4.6. Comparsion with Bitnet

Forward stability in quantized matrix multiplication, particularly vulnerable to overflow from minor input perturbations, arises from the greater magnitude of quantized elements, notably ±1, compared to typical FP16 matrix parameters. Restoring output activation variance to FP16 levels involves multiplication with value vectors of comparable magnitude. Conversely, backward stability is challenged by the non-differentiable nature of the Sign(·) function, which can lead to infinite gradients with matrix element changes. Employing numerically smaller value vectors for multiplication prevents layer-by-layer gradient accumulation and explosion during back-propagation. Furthermore, using the hyperbolic tangent function's derivative as a proxy for the Sign(·) function's derivative circumvents gradient explosion at weight zero points.

5. Conclusion

In this paper, we introduce a pioneering approach to extreme low-bit quantization in speech translation models, leveraging a one-bit linear layer based on Sign Value Constraint Decomposition (SVCD) and knowledge distillation. This method effectively balances the need for reduced model size with performance maintenance, particularly addressing the challenges posed by the decoder's quantization of the attention module. We hope to explore a promising direction for developing efficient speech translation models.

This work is supported by the Natural Science Foundation of China (62366037, 62066033); Outstanding Youth Project of Inner Mongolia Natural Science Foundation (2022JQ05); Young science and technology talents cultivation project of Inner Mongolia University (21221505).

6. References

[1] Q. Fang and Y. Feng, "Understanding and bridging the modality gap for speech translation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 15 864–15 881. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-long.884

[2] Q. Fang, R. Ye, L. Li, Y. Feng, and M. Wang, "STEMM: Self-learning with speech-text manifold mixup for speech translation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7050–7062. [Online]. Available: https://aclanthology.org/2022.acl-long.486

[3] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[4] Y. Tang, J. M. Pino, X. Li, C. Wang, and D. Genzel, "Improving speech translation by understanding and learning from the auxiliary text translation task," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021, pp. 4252–4261. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.328

[5] Y. Tang, J. M. Pino, C. Wang, X. Ma, and D. Genzel, "A general multi-task learning framework to leverage text data for speech to text tasks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 6209–6213. [Online]. Available: https://doi.org/10.1109/ICASSP39728.2021.9415058

[6] W. Hsu, B. Bolte, Y. 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 Lang. Process.*, vol. 29, pp. 3451–3460, 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[8] S. Ma, L. Dong, S. Huang, D. Zhang, A. Muzio, S. Singhal, H. H. Awadalla, X. Song, and F. Wei, "Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders," *CoRR*, 2021.

[9] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, M. Auli, and A. Joulin, "Beyond english-centric multilingual machine translation," *J. Mach. Learn. Res.*, 2021.

[10] Y. Tang, C. Tran, X. Li, P. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," *CoRR*, 2020.

[11] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguistics*, 2020.

[12] G. I. Gállego, I. Tsiamas, C. Escolano, J. A. R. Fonollosa, and M. R. Costa-jussà, "End-to-end speech translation with pre-trained models and adapters: UPC at IWSLT 2021," in *Proceedings of the 18th International Conference on Spoken Language Translation, IWSLT 2021, Bangkok, Thailand (online), August 5-6, 2021*, M. Federico, A. Waibel, M. R. Costa-jussà, J. Niehues, S. Stüker, and E. Salesky, Eds. Association for Computational Linguistics, 2021, pp. 110–119. [Online]. Available: https://doi.org/10.18653/v1/2021.iwslt-1.11

[13] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "GPTQ: accurate post-training quantization for generative pre-trained transformers," *CoRR*, vol. abs/2210.17323, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2210.17323

[14] S. Kim, C. Hooper, A. Gholami, Z. Dong, X. Li, S. Shen, M. W. Mahoney, and K. Keutzer, "Squeezellm: Dense-and-sparse quantization," *CoRR*, vol. abs/2306.07629, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.07629

[15] 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 *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 38 087–38 099. [Online]. Available: https://proceedings.mlr.press/v202/xiao23c.html

[16] Z. Liu, B. Oguz, C. Zhao, E. Chang, P. Stock, Y. Mehdad, Y. Shi, R. Krishnamoorthi, and V. Chandra, "LLM-QAT: data-free quantization aware training for large language models," *CoRR*, vol. abs/2305.17888, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.17888

[17] W. Shao, M. Chen, Z. Zhang, P. Xu, L. Zhao, Z. Li, K. Zhang, P. Gao, Y. Qiao, and P. Luo, "Omniquant: Omnidirectionally calibrated quantization for large language models," *CoRR*, vol. abs/2308.13137, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2308.13137

[18] H. Wang, S. Ma, L. Dong, S. Huang, H. Wang, L. Ma, F. Yang, R. Wang, Y. Wu, and F. Wei, "Bitnet: Scaling 1-bit transformers for large language models," *CoRR*, vol. abs/2310.11453, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.11453

[19] Y. Xu, X. Han, Z. Yang, S. Wang, Q. Zhu, Z. Liu, W. Liu, and W. Che, "Onebit: Towards extremely low-bit large language models," *CoRR*, vol. abs/2402.11295, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.11295

[20] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[21] ——, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[22] M. A. D. Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: a multilingual speech translation corpus," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 2012–2017. [Online]. Available: https://doi.org/10.18653/v1/n19-1202

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. of ACL*, 2002.