USM-Lite

Quantization And Sparsity Aware Fine-Tuning For Speech Recognition With Universal Speech Models

Authors: Shaojin Ding, David Qiu, David Rim, Yanzhang He, Oleg Rybakov, Bo Li, Rohit Prabhavalkar, Weiran Wang, Tara N Sainath, Zhonglin Han, Jian Li, Amir Yazdanbakhsh, Shivani Agrawal

Presenter: Shaojin Ding

Highlights

We propose a model compression approach for Universal Speech Model fine-tuning

- With a **low-bit quantization** and **N:M structured sparsity** aware paradigm on the model weights
- Compress a 2-billion-parameter USM to **9.4% of the original model size** with modest WER regressions

Model	Quantization	Sparsity	WER (%)	Model Size Ratio*
2B CTC USM (baseline)	float32	dense	4.1	N/A
2B CTC USM (best candidate)	int4	2:4 sparsity	4.4	9.4%

* Model Size Ratio is computed as the ratio of the estimated model size relative to the baseline.

Agenda

- Motivations
- Proposed approach
    - Native Quantization-Aware Training
    - Magnitude based Pruning with N:M Sparsity
    - Joint optimization with Quantization and Sparsity
- Experimental setup
- Results
- Conclusions, Limitations, and Future work

Motivations

Automatic Speech Recognition (ASR)

- End-to-end ASR has been widely integrated into modern user-interactive AI services and devices
- **Improving latency and serving cost without losing recognition quality** to benefit live ASR apps with both server-side and on-device model
- Even more important in this *large* model era

Motivations

Universal/Foundational Speech Model (USM)

- Self-supervised learned (SSL) speech representations dramatically improves ASR quality
- Universal Speech Model scales SSL models up
  - Massive model sizes (billions of parameter)
  - Capture multi-domain and multi-lingual distributions
  - Serve for increasing number of speech processing tasks
- Challenges
  - USMs are expensive to be deployed, due to the need of large amount of memory and computational resources

Motivations

Existing ASR compression studies

- With a single compression technique, we usually see significant quality drop at high compression ratio (e.g., quantization, sparsity, knowledge distillation, etc.)
- Experiment with smaller backbones (millions of parameters)

Proposal

- Compressing ASR models from different perspectives at the same time
  - **Quantization**: reduces the model complexity from the **parameter precision**
  - **Sparsity**: reduces the model complexity from the **matrix topology**
- We propose a USM fine-tuning approach for ASR on model weights with joint
  - **Low-bit quantization**
  - ***N:M* structured sparsity**
- Both techniques are **hardware friendly** and are supported by modern GPUs and TPUs

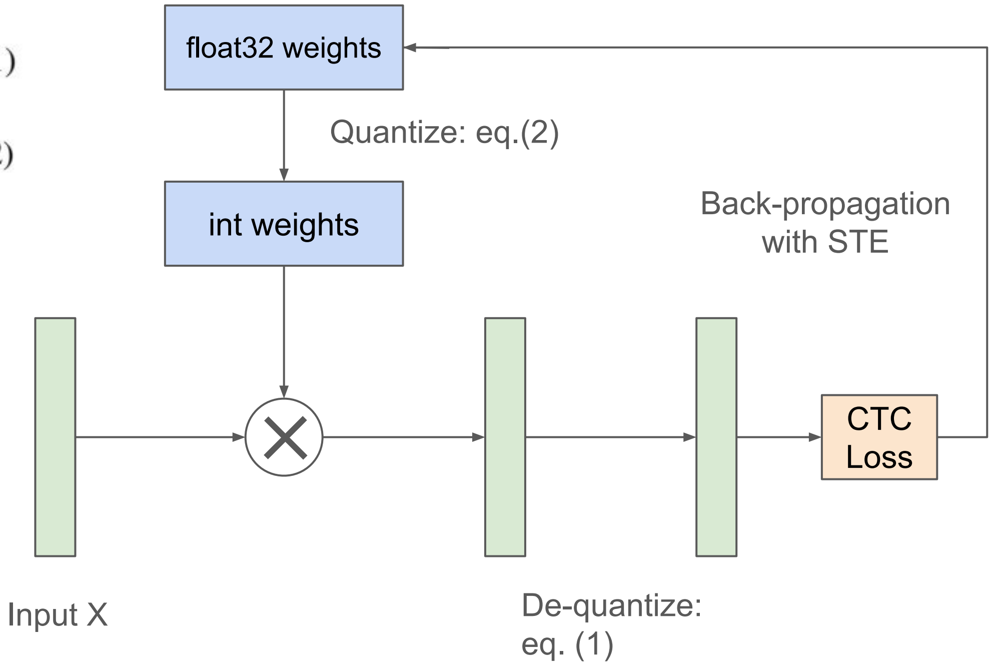
Native Quantization-Aware Training (QAT)

Example on simple matrix multiplication

$$\mathbf{Y}_j = \mathbf{s}_j \cdot \left[\mathbf{X} \otimes \text{Quantize}(\mathbf{W}_j)\right], 1 \leq j \leq J, \tag{1}$$

$$\text{Quantize}(\mathbf{W}_j) = \text{round}\left(\frac{\mathbf{W}_j}{\mathbf{s}_j}\right), \tag{2}$$

- Run eq. (1) and (2) during FP
- Cast the quantized weight from eq.(2) to the **native integer type**
- Straight Through Estimator (STE) to bypass the rounding function during BP



Google

Magnitude based Pruning with *N:M* Sparsity

- Sparsity pattern
  - For each group of M consecutive weights, there are **at most N non-zero values**
- Pruning schedule
  - One-shot
    - Only update the mask once at the beginning of the fine-tuning
  - Few-shot
    - Updates the mask for T_p times at the beginning of the fine-tuning

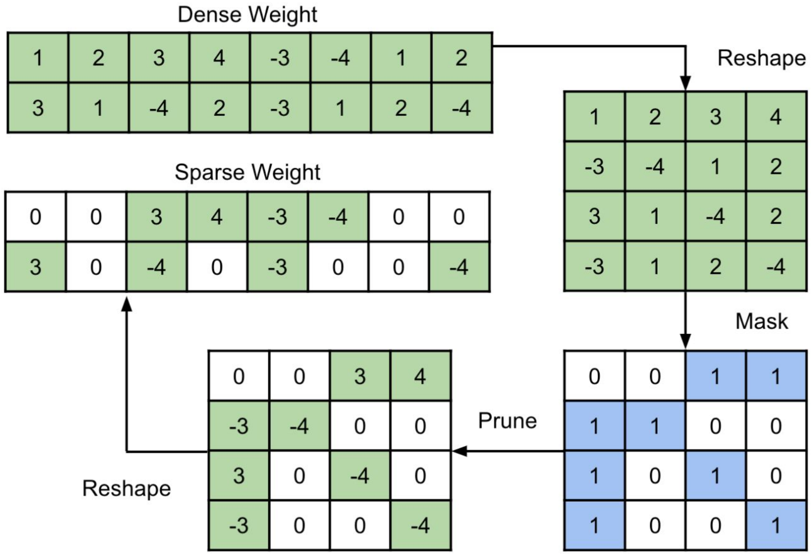


Fig. 1. Illustration of magnitude based pruning with $N{:}M$ sparsity on a weight matrix. This example has $N = 2$ and $M = 4$.

Joint optimization with Quantization and Sparsity

- Prune-and-quantize fashion
  - The pruned weights are set to zero
  - Directly maps to the zero-point of symmetric quantization
  - Has no effect on calculating the quantization scale - zero-point weights do not contribute to scale calculation

Experimental Setups

- Pre-trained with BEST-RQ [13] on over 12 million hours of multilingual speech data from YouTube *

Model	# Params (B)	# Layers	Dimension	Att. Heads	Conv. Kernel Size
Conformer CTC	2.0	32	1536	16	5

- Fine-tuning datasets
  - 1.2-million-hour U.S. English audio-text pairs from voice search, anonymized *
  - A small portion of the dataset is hand-transcribed
  - The rest is pseudo-transcribed with a 600-million-parameter teacher model

[13] Self-supervised learning with random-projection quantizer for speech recognition

* Our data handling abides by Google AI Principles: https://ai.google/responsibility/principles/

Ablation Studies on Quantization

int8 quantization

- PTQ and QAT can retain float32 quality

int4 quantization

- Need QAT to retain float32 quality

int2 quantization

- Quality regressions across the board
- Need sub-channel quantization [25] to reach a reasonable quality

Table 1. Results of ablation studies on quantization. *Model Size Ratio* is computed as the ratio of the estimated model size relative to *B0*. PTQ refers to post-training quantization.

Exp	Model	Voice Search WER	Model Size Ratio
B0	float32 dense 2B CTC USM	4.1	-
E0	int8 PTQ	4.2	25.0%
E1	int8 QAT	4.2	25.0%
E2	int4 PTQ	86.7	12.5%
E3	int4 QAT	4.3	12.5%
E4	int2 QAT	99.9	6.3%
E5	int2 QAT + 16 sub-channel	45.2	7.3%
E6	int2 QAT + 32 sub-channel	32.0	8.3%
E7	int2 QAT + 64 sub-channel	12.3	10.4%

[25] 2-bit conformer quantization for automatic speech recognition

Ablation Studies on Sparsity

2:4 sparsity

- One-shot and 1k-shot prunings both have minimal WER regressions

1:4 sparsity

- Quality regressions across the board
- 1k-shot significantly outperforms one-shot pruning

Table 2. Results of ablation studies on N:M sparsity. *Model Size Ratio* is computed as the ratio of the estimated model size relative to *B0*.

Exp	Model	Voice Search WER	Model Size Ratio
B0	float32 dense 2B CTC USM	4.1	-
E8	2:4 sparsity one-shot	4.4	53.1%
E9	2:4 sparsity 1k-shot	4.3	53.1%
E10	1:4 sparsity one-shot	11.7	28.1%
E11	1:4 sparsity 1k-shot	10.6	28.1%

Combining Quantization with Sparsity

Smaller backbones

- Increasing regressions when reducing model sizes

Combining quantization with sparsity

- 9.4% of the original model size with 7.3% relative WER regressions
- Superior quality compared to applying either technique solely
- Parity with 1B USM but much smaller

Table 3. Results of the proposed paradigm of combining quantization and $N{:}M$ sparsity. Results on baseline USM with different model sizes are also presented here for comparisons. *Model Size Ratio* is computed as the ratio of the estimated model size relative to *B0*.

Exp	Model	Voice Search WER	Model Size Ratio
B0	float32 dense 2B CTC USM	4.1	-
B1	float32 dense 1B CTC USM	4.5	50.2%
B2	float32 dense 600M CTC USM	4.7	33.5%
B3	float32 dense 300M CTC USM	5.0	18.9%
E7	int2 QAT + 64 sub-channel	12.3	10.4%
E11	1:4 sparsity 1k-shot	10.6	28.1%
E12	int4 QAT + 2:4 sparsity one-shot	4.4	9.4%
E13	int4 QAT + 2:4 sparsity 1k-shot	4.5	9.4%

Conclusions, Limitations, and Future work

Conclusions

- Ablation studies corroborate the effectiveness of quantization and sparsity during USM fine-tuning
- Compressing the model jointly from the **parameter precision** and the **matrix topology** perspectives are more effective than an individual technique

Limitations and Future work

- STE is not enabled for **pruning operator**, which can possibly improve the performance of models with N : M sparsity
- Investigate more aggressive combinations such as **int2 + 2:4 sparsity** in future work
- Validate the proposed approach on other speech processing tasks

Thanks! Q&A