경량 실시간 음성인식 신경망의 양자화 영향 분석

안성환, 김세민, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실

{shahn, smkim21}@hi.snu.ac.kr, nkim@snu.ac.kr

Impact of Quantization on Lightweight Real-time Automatic Speech Recognition Networks

Sung Hwan Ahn, Semin Kim, Nam Soo Kim
Human Interface Laboratory,
Department of Electrical and Computer Engineering and INMC,
Seoul National University

요 약

본 논문은 경량화된 실시간 음성인식 모델에 여러 가지 양자화 방법을 적용한 후 성능 결과를 분석한다. 여러 정밀도에 대해 양자화 하지 않은 모델과 동일하거나 비슷한 수준의 성능을 내는 양자화 방법을 보고한다.

Ⅰ. 서 론

음성인식(automatic speech recognition, ASR)이란 입력으로 들어온 오디오를 상응하는 텍스트로 전사하는 시스템이다. 최근 딥러닝의 발전과 함께 심층신경망 기반 ASR 시스템의 성능이 매우 좋아졌으며, 자원이 제한적인 온디바이스 (on-device) 환경에 사용 가능할 정도로 경량화된 실시간 ASR 시스템도 등장했다 [1]. 이 때 ASR 성능은 최대한 유지하면서 모델 파라미터 크기 및 연산 복잡도를 줄이기 위한 연구가 활발히 이루어지고 있다.

모델 경량화를 위한 핵심 기술 중 하나인 양자화 (quantization)는 데이터의 표현 정밀도를 32-bit 또는 16-bit 부동소수점 (floating point, FP)보다 낮춰 모델 크기 및 연산 비용을 낮춘다. 특히 합성곱 계층 (convolutional layer, Conv)은 전체 시스템에서 차지하는 parameter 크기 및 연산량의 비율이 매우 높아 양자화의 주 타겟이다. 일반적으로 [그림 1]과 같이 Conv 의 입력과 weight 을 정수 (integer, Int)로 양자화하여 정수 합성곱 연산을 수행하고, 그 결과를 다시 FP 로 역양자화 (dequantization)하는 방식이 사용된다.

심층신경망 모델의 파라미터 개수가 많을수록 양자화 이후 성능이 잘 유지되지만, 작은 모델일수록 양자화 이후 성능이 많이 떨어진다는 것이 실험적으로 알려져 있다. 하지만 ASR 모델에 양자화를 적용하는 연구는 아직 활발히 진행되지 않는 상황이다. 대부분 온디바이스 환경에 사용되기 힘든 큰 모델을 대상으로 하였고, 양자화 이후 학습을 진행하지 않는 post-training quantization (PTQ) 기법에 한정되었다 [2].

이에 본 연구는 온디바이스용으로 설계된 매우 작은 크기의 실시간 ASR 모델에 양자화를 적용하였다. 여러 양자화 기법을 활용하여 여러 정밀도의 양자화를 적용하고, 이에 따른 성능 변화를 분석했다.

Ⅱ. 본론

[그림 1]은 본 연구에서 사용한 ASR 모델의 양자화 개요도이다. ASR 모델은 [1]에서 제안된 1 차원 Conv, Batch Normalization, activation 으로만 이루어진 실시간 경량 모델을 사용하되, 모델 크기를 온디바이스용으로 더욱 경량화하기 위해 아래와 같이 세부 설정을 변경했다. Mel spectrogram 입력에 대해 320ms chunk 단위로 실시간 음성인식을 수행한다. Encoder 는 11 개의 Conv Block 으로 구성되어 있으며, 각 block 의 채널 크기는 256, 은닉 채널 크기는 1024 이다. Decoder 및 joiner 의 채널 크기는 256 이다. 총 파라미터 개수는 7.1M 개이다.

온디바이스를 타겟으로 하는 만큼 symmetric, per-tensor 의 간단한 양자화 granularity 를 사용하였다. 정밀도는 두 가지를 적용하였다. 모든 Conv 에 대해, 입력과 weight 을 모두 Int8 로 양자화 (A8W8) 하거나 입력은 Int8, weight 은 Int4 로 양자화 (A8W4) 하였다.

입력 x, 양자화 비트 수 b, scale s에 대해 양자화 결과 $\hat{x}$는 아래와 같이 계산된다.

$$\hat{x} = Clamp\left(Round\left(\frac{x}{s}\right), -2^{b-1}+1, 2^{b-1}-1\right) \times s \quad (1)$$

PTQ 를 위해서는 수식 (1)에서 s를 잘 정해야 양자화 이후 성능하락을 최소화할 수 있다. 여러 방법을 시도해 본 후 가장 성능이 좋은 방법을 사용하였으며, 구체적으로는 아래와 같다.

- 입력: 양자화 전/후의 MSE loss 를 minimize 하도록 s를 계산.
- Weight: $max(abs(weight))$를 s로 사용.

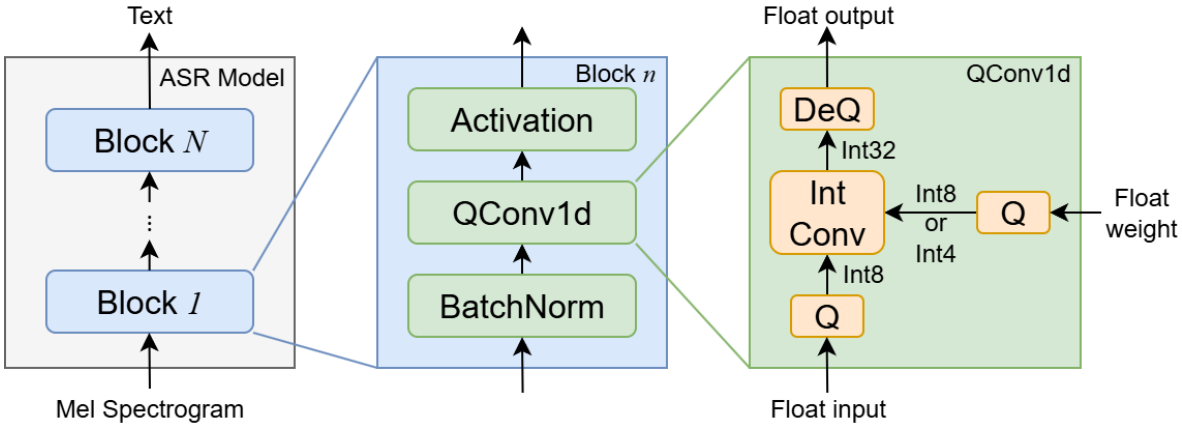

그림 1. 음성인식 모델 양자화 개요도. Q 는 quantization, DeQ 는 de-quantization, Conv 는 convolutional layer 를 의미한다.

Quantization-aware training (QAT)의 경우, fine-tuning 과 from scratch 를 진행했다. Finetuning 은 FP 학습을 완료하고, 해당 모델을 PTQ 한 다음, 양자화된 모델을 작은 learning rate 로 추가 학습하는 것을 의미한다. From scratch 는 FP 학습 없이 처음부터 QAT 를 진행하는 것을 의미하며, scale s 를 직접 학습하는 방법 [3] [4]을 채택하였다.

모든 모델은 LibriSpeech[5] 의 train-clean 과 train-other 데이터셋으로 학습하고 LibriSpeech test-clean 과 test-other 데이터셋으로 성능을 검증했다. 검증시에는 weight averaging 기법을 적용하였다.

정밀도	양자화 방법	test-clean	test-other
Full-precision	-	6.99	16.99
A8W8	PTQ	7.27	17.61
	QAT (finetuning)	6.81	16.89
A8W4	PTQ	99.27	99.62
	QAT (finetuning)	11.28	24.59
	QAT (from scratch)	8.29	19.21

[표 1] LibriSpeech test 데이터셋에서의 WER 성능

각 양자화 기법에 대해 word error rate (WER)을 측정한 결과는 [표 1]과 같다. A8W8 를 살펴보면, PTQ 이후 성능 하락이 WER 차이 1% 이내로 유지되며, 이후 QAT finetuning 까지 진행하면 원래의 성능이 복구될 뿐 아니라 FP 보다 성능이 아주 소폭 좋아진다. 이는 학습을 더 오래 한 효과로 해석할 수 있다. A8W4 의 경우, PTQ 만으로는 성능이 전혀 나오지 않는다. QAT finetuning 을 진행하면 성능이 어느 정도 나오지만, FP 에 비해서는 많이 떨어진다. QAT from scratch 는 FP 의 성능에 많이 근접한 것을 확인할 수 있다. 파라미터 크기가 7.1M 에 불과한 점과, 일반적으로 파라미터가 많아야 양자화 이후 성능이 유지된다는 점을 고려하면 이는 놀라운 성능으로 볼 수 있다.

Ⅲ. 결론

본 논문은 온디바이스용 실시간 경량 ASR 모델에 대해 여러 정밀도의 여러 양자화 방법을 적용해 보았다. 입력과 weight 을 int8 로 양자화 하면 성능이 유지되고, 입력은 int8, weight 은 int4 로 양자화 하면 QAT from scratch 를 통해 성능 손실을 최소화할 수 있다.

ACKNOWLEDGMENT

이 논문은 2026 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참 고 문 헌

[1] 안성환, 우범준, 이동준, 김남수, “임베디드 시스템을 위한 심층신경망 기반 스트리밍 음성 인식 모델,” 한국통신학회 학술대회논문집, 2024, pp. 1450-1451.

[2] Chen Feng, Yicheng Lin, Shaojie Zhuo, and Chenzheng Su, “Edge-ASR: Towards Low-Bit Quantization of Automatic Speech Recognition Models,” *arXiv preprint arXiv:2507.07877*, 2025.

[3] Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha, “Learned Step Size Quantization,” in *Proc. ICLR* 2020.

[4] Wenqi Shao et al., “OmniQuant: Omnidirectionally Calibrated Quantization for Large Language Models,” in *Proc. ICLR* 2024.

[5] Vassil Panayotov et al., “Librispeech: An ASR corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206-5210.