Transformer 기반 음성 인식 모델의 경량화를 위한 Quantization 및 LoRA 적용 연구

한승민, 정의현, 정진우, 차민수, 윤수연*
국민대학교

handsomemin@kookmin.ac.kr, rabbitdy0101@kookmin.ac.kr, imaboybut@kookmin.ac.kr,
minns00@kookmin.ac.kr, *1104py@kookmin.ac.kr

Applying Quantization and LoRA for Efficient Transformer-Based Speech Recognition Models

Seung Min Han, Ui Hyun Jung, Jin Woo Chung, Min Su Cha, Soo Yeon Yoon*
Kookmin Univ.

요 약

공학 · 수학 · 화학 등 수식 발화가 잦은 도메인에서는 일반적인 음성 인식 모델만으로는 숫자, 기호, 단위 등이 포함된 발화를 정확히 처리하기 어렵다. 이러한 특성을 반영하려면 해당 task 에 특화된 파인 튜닝이 필요하지만 고성능 모델은 높은 VRAM 자원을 요구하고, 경량 모델은 수식 인식 성능이 부족한 한계가 있었다. 본 연구에서는 Whisper-Large 모델에 8-bit Quantization 과 LoRA 를 적용하여 수식 인식 성능을 유지하면서도 자원 효율을 높일 수 있는 구조를 실험적으로 검토하였다. 그 결과 학습에 필요한 VRAM 소모량은 기존 Whisper-small 대비 약 68% 감소하였으며 학습에 사용되지 않은 공학 수학 대학 강의 데이터를 통해 성능을 평가한 결과 기존 대비 CER 은 15.59%, WER 은 37.82% 개선되었다.

I. 서론

본 연구는 음성 인식 모델에 대하여 수식이나 공식이 포함된 발화를 보다 정확하게 추출하기 위한 목적에서 비롯되었다. 초기 실험에서는 Whisper-small 모델을 기반으로 파인 튜닝하였고 학습 데이터와 동일한 출처 의 데이터에 대해서는 CER 약 9% 수준으로 준수한 성능을 보였다. 그러나 외부 강의 음성 입력 시 수식 발화에 관하여 결과가 좋지 못 하였고, 이는 whisper-small 모델의 용량에서 오는 근본적인 한계라고 판단하였다. 이에 따라 모델의 규모가 더 큰 Whisper-large 로의 파인 튜닝을 시도하였으나 보유한 자원에서는 VRAM 사용량 초과로 인하여 학습 진행이 불가능하였다.

이에 본 연구에서는 VRAM 소모량을 줄이면서도 도메인 특화 발화를 반영할 수 있도록 양자화와 LoRA 기반의 경량화 구조를 적용하는 실험을 진행하였다.

II. 관련 연구

2.1 음성 인식 모델

2.1.1 Whisper

Whisper 는 OpenAI 에서 공개한 End-to-End 음성 인식 모델로, Common Voice, LibriSpeech 등 주요 벤치마크에서 상위권 성능을 기록하고 있다. Transformer 기반의 only-decoder 구조를 통해 입력 음성 전체를 컨텍스트로 활용하며 텍스트를 생성한다. 특히 mel-spectrogram 을 직접 입력받아 예측을 수행하는 구조는 복잡한 수식 표현이나 수치가 조합된 발화에서도 높은 표현력을 보이는 구조라서 본 연구의 도메인 특화 음성 인식 목적에 적합하다고 판단하였다.[1]

2.2 모델 경량화

2.2.1. Quantization

Quantization 은 모델의 연산 정밀도를 낮춰 연산량과 메모리 사용량을 줄이는 방법이다. 일반적으로 32 비트 또는 16 비트 부동소수점 연산을 8 비트 정수 수준으로 변환하며 성능 저하 없이도 GPU 메모리 점유율을 줄일 수 있다고 알려져 있다.[2]

2.2.2 LoRA(Low-Rank Adaption)

LoRA 는 기존 모델의 가중치를 고정한 채 일부 선형 계층에 저차원 행렬을 삽입하여 미세 조정하는 방식이다.[3] 전체 파라미터를 학습하지 않고도 효율적인 파인 튜닝이 가능하며 학습 자원 절감에 효과적이다.

III. 연구 방법

3.1 데이터셋 구성

데이터셋은 AI-Hub 에서 제공하는 국내 '대학 강의 음성 데이터' 를 기반으로 구성하였으며 전체 약 36 만 개의 샘플 중 영어 단어나 숫자 발화가 모두 포함된 약 4 만 7 천 개의 데이터를 선별하여 학습에 활용하였다. 수식이나 공식의 발화는 대부분 영어 단어나 숫자 표현으로 구성되기 때문에 이러한 조건을 포함하는 데이터만을 추출하였다.

3.2 실험 환경

모델 학습은 단일 GPU 기반 Linux 환경에서 수행하였으며 사용한 주요 라이브러리는 [표 2]에 정리하였다.

표 2. 라이브러리 환경

Category	Value
Python Version	3.10.16
GPU	NVIDIA RTX 3090 24GB
PyTorch	2.6.0
Transformers	4.51.3
PEFT	0.15.2.dev0

3.3 모델 구성 및 학습 방식

Whisper-Large-v2 모델은 8-bit 정밀도로 양자화된 형태로 로드한 뒤 q_proj 와 v_proj 계층에만 LoRA 를 적용하여 파인 튜닝을 수행하였다. 모델 양자화는 Hugging Face 에서 제공하는 BitsAndBytesConfig 를 활용하였고 LoRA 는 PEFT 프레임워크를 통해 구성하였다. 전체 파라미터를 학습하지 않고 일부 계층만을 학습 대상으로 설정한 구조로 양자화 정밀도만 Normal Float 4bit 가 아닌 Integer 8bit 로 구성된다는 점을 제외하면 QLoRA 의 설계 방식과 유사하다.[4] 학습 관련 주요 설정은 [표 3]에 정리하였다.

[표 3] Whisper 학습 핵심 설정 요약

Category	Value
Base Model	Whisper-large-v2
Model Setup	8-bit Quantized + LoRA
LoRA Config	r=32, α=64, dropout=0.1
Batch Size	16 x 4(accumulation)
Learning Rate	2e-5
Epochs	5

IV. 실험 및 결과

4.1 학습 데이터셋 기반 성능 평가

모델의 성능 평가는 문자 단위 오류율(CER)과 단어 단위 오류율(WER)을 기준으로 진행하였다. 한국어 기반 음성 인식에서는 발화의 세부 정확성을 판단하기 위해 WER 보다 CER 이 더 민감하게 작용하므로 본 연구에서는 CER 을 주요 평가 지표로 WER 을 보조 지표로 활용하였다. 먼저, 학습에 사용된 데이터셋에 대한 평가에서는 두 모델 간 CER, WER 수치가 유사한 수준으로 나타났고, [표 4]에 정리하였다.

[표 4] 학습 데이터셋 기반 성능 평가 결과

Model	CER(%)	WER(%)
Whisper-small	9.21	36.21
Whisper-large(Quant+LoRA)	10.03	37.62

표[4]에서 확인할 수 있듯, 학습에 사용된 데이터셋과 같은 출처인 테스트 데이터셋을 기준으로 평가한 결과, CER 과 WER 수치가 유사하는 것을 통해 두 모델 간 기본적인 학습 성능은 구조나 용량과 무관하게 거의 동일했다는 것을 알 수 있다.

그러나 학습에 사용되지 않은 외부 강의 음성에 대한 추론 결과에서는 명확한 성능 차이가 나타났다. [표 5]에 따르면, Whisper-large 는 whisper-small 에 비해 CER 기준 15.59% 높았으며 WER 기준 37.82% 높게 측정되어 일반화 성능에서 확연한 개선이 확인되었다.

[표 5] 외부 데이터셋 기반 일반화 성능 평가 결과

Model	CER(%)	WER(%)
Whisper-small	21.84	50.30
Whisper-large(Quant+LoRA)	6.25	12.48

이는 학습 데이터 기반 성능이 유사하게 수렴한 조건에서 도출된 결과로, Whisper-large(Quant+ LoRA) 모델이 수식 발화에 대한 인식 구조 일반화에 더욱 적합함을 보여준다.

4.2 VRAM 사용량

Whisper-small 모델은 학습 과정에서 약 22GB 의 VRAM 을 소모하였다. 반면, Whisper-large 에 Quantization 과 LoRA 를 적용한 구성은 약 7GB 수준으로 약 68% 가량 VRAM 사용량이 감소하였다. 동일한 학습 데이터를 처리하면서도 훨씬 적은 연산 자원으로 모델 학습이 가능하였고, 관련 수치는 [표 6]에 정리하였다.

[표 6] 모델 학습 시 VRAM 사용량 비교

Model	Peak VRAM Usage
Whisper-small	24.35 GB
Whisper-large(Quant+LoRA)	6.25 GB

V. 결론

Whisper-Large 모델에 Quantization 과 LoRA 를 결합하여 설계한 본 실험 구조는 기존 Whisper-small 모델 대비 약 68% 수준의 VRAM 만을 사용하면서도 외부 데이터셋에 대한 일반화 성능에서는 CER 기준 15.59%, WER 기준 37.82%의 개선을 달성하였다. 이처럼 VRAM 요구량이 낮아지면서 Whisper 기반 모델의 파인 튜닝은 보다 다양한 분야와 기관에서 접근하기 쉬울 것이다. 특히 전문 교육, 의료, 법률 등에서 맞춤형 적용이 용이할 것이다. 또한, 강의 영상 기반 정보를 신뢰성 있게 제공하는 RAG 기반 챗봇 시스템 구축에도 기여할 수 있어 LLM 의 환각 현상을 방지하는 데에 활용할 수 있다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음. (2022-0-00964)

This research was supported by the Ministry of Science and ICT (MSIT) and the Institute for Information & Communications Technology Planning & Evaluation (IITP) under the National Program of Excellence in Software (2022-0-00964).

참 고 문 헌

[1] Alec Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv preprint arXiv:2212.04356, Dec. 2022.

[2] Edward Zhen, William Chan, Daniel S. Park, et al., "Sub-8-bit Quantization for On-Device Speech Recognition: A Regularization-Free Approach," arXiv preprint arXiv:2210.09188, Oct. 2022.

[3] Edward Hu, Yelong Shen, Phillip Wallis, et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, Jun. 2021.

[4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, May 2023.

[5] Sungwon Park, Yoon Kim, "Advocating Character Error Rate for Multilingual ASR Evaluation," arXiv preprint arXiv:2410.07400, Apr. 2024.