온 디바이스 국방 AI를 위한 PEFT 효용성 연구

배기민[O], 이학진[**], 김세옥[*], 이장형[*]
[O]육군미래혁신연구센터,
[*]육군미래혁신연구센터,
[**]육군지능정보기술단
e-mail: {baemin.dev[O], himit0131[**], bboyseiok[*]}@gmail.com, leejangh@mnd.go.kr[*]

Research on PEFT Feasibility for On-Device Military AI

Gi-Min Bae[O], Hak-Jin Lee[**], Sei-Ok Kim[*], Jang-Hyong Lee[*]
[O]Korea Army Research Center for Future and Innovation,
[*]Korea Army Research Center for Future and Innovation,
[**]Korea Army Intelligence and Information Technology Group

● 요 약 ●

본 논문에서는 온 디바이스 국방 AI를 위한 효율적인 학습 방법을 제안한다. 제안하는 방법은 모델 전체를 재학습하는 대신 필요한 부분만 세밀하게 조정하여 계산 비용과 시간을 대폭 줄이는 PEFT 기법의 LoRa를 적용하였다. LoRa는 기존의 신경망 가중치를 직접 수정하지 않고 추가적인 낮은 랭크의 매트릭스를 학습하는 방식으로 기존 모델의 구조를 크게 변경하지 않으면서도, 효율적으로 새로운 작업에 적응할 수 있다. 또한 학습 파라미터 및 연산 입출력에 데이터에 대하여 32비트의 부동소수점(FP32) 대신 부동소수점(FP16, FP8) 또는 정수형(INT8)을 활용하는 경량화 기법인 양자화도 적용하였다. 적용 결과 학습시 요구되는 GPU의 사용량이 32GB에서 5.7GB로 82.19% 감소함을 확인하였다. 동일한 조건에서 동일한 데이터로 모델의 성능을 평가한 결과 동일 학습 횟수에선 LoRa와 양자화가 적용된 모델의 오류가 기본 모델보다 53.34% 증가함을 확인하였다. 모델 성능의 감소를 줄이기 위해서는 학습 횟수를 더 증가시킨 결과 오류 증가율이 29.29%로 동일 학습 횟수보다 더 줄어듬을 확인하였다.

키워드: PEFT(Parameter-Efficient Fine-Tuning), 자동음성인식(Automatic Speech Recognition), QLoRA(Quantized Low Rank Adapters)

I. Introduction

현대 인공지능 모델들은 파라미터의 개수와 모델의 크기가 점점 증가하는 추세를 보이고 있다. 이러한 추세를 보이는 이유는 대용량 데이터를 학습해 마치 인간처럼 종합적 추론이 가능한 초거대 AI 개발을 목적으로 하기 때문이다.

초거대 AI는 인간의 뇌처럼 스스로 추론하고 창작할 수 있도록 방대한 데이터와 파라미터를 활용하는 인공지능(AI) 모델을 의미한다. 초거대 AI 모델 중 하나인 GPT, Whisper는 대규모 파라미터와 데이터 세트의 방대한 일반 데이터 학습만으로도 대부분의 작업을 완성도 높게 수행하고 있다.

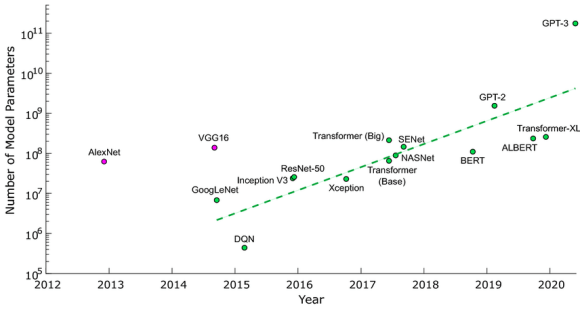

Fig. 1. Number of parameters in recent landmark neural networks [1]

하지만 초거대 AI를 구축하기 위해서는 대규모의 GPU 자원 확보가 필요하지만 수요가 공급을 초과하여 GPU 부족 현상이 발생되고 있다. 또한 이러한 현상으로 인해 증가되는 GPU 비용과 장기간의 구매 대기 시간 문제로 단일 GPU로의 학습은 점차 어려운 과제가 되고 있다.

양자화는 이러한 상황에서 초거대 AI 모델을 학습하거나 추론하기 위해 널리 사용되는 방법 중 하나로 모델의 파라미터를 저비트로 표현함으로써 계산 및 메모리 접근 속도를 향상시키는 경량화 기법이다. 양자화에는 학습 중 양자화를 고려하는 QAT (Quantization Aware Training)와 학습이 완료된 모델에 적용하는 PTQ(Post Training Quantization) 두 가지 접근법이 있다. QAT는 제한된 메모리와 연산 능력을 가진 임베디드 시스템이나 모바일 장치에서 모델을 효과적으로 학습하는 데 유용하며, PTQ는 리소스가 제한된 환경에서 실시간 추론이 필요한 경우에 적합하다. [2]

PEFT(Parameter-Efficient Fine-Tuning)는 사전 학습된 모델의 일부만을 미세 조정하여 새로운 데이터나 작업에 맞게 조정하는 방법으로 전체 모델을 처음부터 학습하는 것보다 훨씬 적은 수의 파라미터를 조정하면서 학습 시간과 연산 비용을 절약할 수 있다. [3]

국방 영역에서 AI를 사용하는 경우, 보안상의 이유로 데이터를 외부로 반출할 수 없으며, 기존 학습 데이터와의 도메인 차이를 고려해야 하므로, 사전 학습된 모델을 기반으로 한 전이 학습이 필수적이며 최근의 사전 학습된 모델들이 높은 하드웨어 요구량을 가지고 있다.

본 연구는 PEFT와 양자화를 통해 낮은 하드웨어 요구량에서도 최소한의 정확도 손실을 갖는 효과적인 모델을 개발하는 것을 목표한다.

II. Preliminaries

1. 신경망 경량화 기술

신경망의 경량화는 신경망의 불필요한 연산의 제거를 통해 메모리 사용량과 연산 복잡도를 감소시켜 학습 및 추론 처리시간을 단축하는 최적화 방법론으로 양자화(Quantization), 가지치기(Pruning), 경량 자동화(AutoML)등으로 구분된다. [4]

양자화는 학습 파라미터 및 연산 입출력에 데이터에 대하여 32비트의 부동소수점(FP32) 대신 부동소수점(FP16, FP8) 또는 정수형(INT8)을 활용하는 경량화 기법으로 정확도 손실을 최소화 하기 위해 데이터 포맷과 비트 수(Bit-Width)를 세밀하게 조절해야한다.

가지치기는 신경망에서 불필요한 부분을 제거하는 방법으로, 데이터의 개별 요소를 대상으로 할 때는 Weight Pruning 분야로, 행렬의 행과 열 또는 그룹 단위로 제거하는 경우 Structured Pruning 분야로, 특정 레이어나 모듈을 전체적으로 제거하는 경우에는 Layer Pruning 분야로 분류된다. 데이터의 개별 요소를 대상으로하는 Weight Pruning은 0을 많이 표현하게 되므로 이를 효율적으로 처리하기 위해 CSR(Compressed Sparse Row)과 같은 방법을 사용하여 하드웨어상에서 처리 효율을 높일 수 있다. 가지치기는 신경망의 일부 데이터를 구조적으로 제거하여 별도의 하드웨어 변경 없이도 신경망의 속도를 빠르게 할 수 있다.

경량 자동화는 데이터의 형식, 비트 수, 행렬의 행과 열 등 어떤 부분을 어느 정도 줄일지 자동으로 결정한다. 이러한 결정은 강화학습이나 특정 규칙을 기반으로 이루어지며 처리 속도, 메모리 사용량, 에너지 사용량 등을 최소화하도록 고려된다. 이러한 방법은 경량화 과정에서 불필요한 시행착오를 줄이고, 최적화 시간을 단축하여 경량화한 후의 성능을 최대한 끌어올리기 위해 사용된다.

2. 전이학습 효율화 기술

PEFT(Parameter-Efficient Fine-Tuning)은 대규모 사전 훈련된 언어 모델들을 특정 작업에 효율적으로 적응시키는 기법으로 모델 전체를 재학습하는 대신 필요한 부분만 세밀하게 조정하여 계산 비용과 시간을 대폭 줄일 수 있다. PEFT는 모델의 전체 매개변수 수를 유지하면서도, 작업 특화 매개변수만을 업데이트하여 모델의 성능을 최적화한다. 이는 특히 대규모 모델을 실시간 시스템이나 메모리가 제한된 환경에서 사용할 때 중요하다.

LoRA(Low-Rank Adaptation)는 PEFT의 대표적인 방법 중 하나로, 기존의 신경망 가중치를 직접 수정하지 않고 추가적인 낮은 랭크의 매트릭스를 학습하는 방식을 채택한다. 이 추가 매트릭스는 원래 모델의 가중치와 결합되어 새로운 작업에 특화된 출력을 생성한다. LoRA의 핵심은 기존 모델의 구조를 크게 변경하지 않으면서도, 효율적으로 새로운 작업에 적응할 수 있도록 하는 것이다. 이 방식은 기존 가중치에 작은 변화만을 추가함으로써, 새로운 작업에 대한 미세 조정이 가능하면서도 전체적인 모델 구조와 매개변수의 규모는 유지한다.[5]

III. The Proposed Scheme

본 연구의 실험은 음성 인식 문제에서 많이 사용되는 OpenAI의 Whisper large v3을 사용하여 PEFT와 양자화가 적용되지 않은 모델과 적용된 모델을 비교

Table 1. System Environment

Item	Value
CPU	Xeon® E5-2683 v4
Memory	64GB
GPU (Vram)	A6000 (48GB)

다음의 표 1은 실험에 사용한 환경 정보이다. PEFT와 양자화가 적용 안된 모델을 실험하기 위해 NVIDIA의 A6000 GPU를 사용하였다. 이를 통해 PEFT와 양자화가 적용에 따른 GPU 메모리 사용량과 이에 따른 모델 성능을 비교한다.

1. 활용 데이터

실험에는 AIHUB에서 제공하는 한국어 음성 데이터 세트를 사용하였다. 한국어 음성 데이터 세트는 대화형 음성인식 성능 개선을 위한

음향모델(Acoustic Modeling)용 한국어 자유발화 음성데이터로 조용한 환경에서 2,000여명이 발성한 한국어 대화음성 1,000시간이 구축되어 있다. 대화 음성은 두 사람이 다양한 주제(예: 일상, 쇼핑, 정치, 경제, 날씨, 취미 등)로 자유롭게 대화하는 음성을 녹음하고 발성내용을 ERTI전사규칙(예: 간투사, 머뭇거림 등)에 따라 철자전사가 되어있다. [6]

2. 데이터 전처리

제공되는 음성 파일은 PCM 형태로 Whisper 모델의 입력을 위해 mp3 파일로 변환하였다.

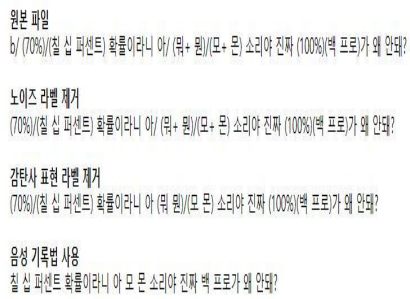

Fig. 2. Processing results for transcription files

그림2는 제공되는 전사 파일의 가공 과정 결과로 첫 번째로 b/, n/, / 등의 노이즈 레이블을 삭제하였으며 두 번째로 감탄사 표현에 사용되는 '/', '*', '+' 등의 라벨을 삭제하였다. 또한 음성 기록 방식을 숫자와 특수 기호를 발음 표기대로 사용하였다. 또한 문자 단위로 가공하였다.

3. 실험 결과

실험에는 Whisper Large V3 모델을 사용하였으며 먼저 각 기법 적용에 따른 GPU 메모리 사용량을 측정하였다.

Table 3. Experiment Result

Model	GPU VRAM usage
Whisper Large V3	32GB
Whisper Large V3 + LoRA	11GB
Whisper Large V3 + LoRA + 8Bit	6.9GB
Whisper Large V3 + LoRA + 4Bit	5.7GB

표 3은 각 기법 적용에 따른 GPU 메모리 사용량 측정 결과로 기본 모델은 32GB의 GPU 메모리 사용량을 확인하였다. PEFT 기법의 LoRA를 적용하는 경우 11GB의 GPU 메모리 사용량을 확인하였으며 기본 방식보다 65.62% 감소 하였으며 LoRa와 양자화를 8Bit로 적용하는 경우 78.44% 감소, 4Bit를 적용하는 경우 82.19% 감소함을 확인하였다.

Table 4. Model Performance

Model	CER
Whisper Large V3	11.459 (100 Step)
Whisper Large V3 + LoRA + 4Bit	17.571 (100 Step) 14.815 (200 Step)

표 4은 기본 모델과 LoRa와 4Bit 양자화가 적용된 모델을 동일한 조건에서 학습 한 후 CER(Character error rate)을 측정한 결과이다. CER은 인식된 문자열(한 글자)과 정답 문자열(한 글자) 사이의 문자 오류 비율을 나타내는 지표로 인식된 문자열에서 잘못된 문자의 개수를 총 문자의 개수로 나눈 비율로 계산할 수 있다. CER이 기본 모델에 비해 53.34% 상승하였으며 이는 LoRa와 4Bit가 적용되어 발생한 효과이며. 이러한 모델 성능의 감소를 줄이기 위해서는 학습 횟수를 더 증가시켜야 한다는 사실을 확인하였다.

IV. Conclusions

현대 인공지능 모델들은 파라미터의 개수와 모델의 크기가 점점 증가하는 추세를 보이고 있다. 이는 대용량 데이터를 학습해 마치 인간처럼 종합적 추론이 가능한 초거대 AI 개발을 목적으로 하기 때문이다. 하지만 초거대 AI를 구축하기 위해서는 대규모의 GPU 자원 확보가 필요하지만 수요가 공급을 초과하여 GPU 부족 현상이 발생되고 있으며 GPU 비용 증가와 장기간의 구매 대기 시간 문제로 단일 GPU로의 학습은 점차 어려운 현실이 되고있다.

국방 영역에서 AI를 사용하는 경우, 보안상의 이유로 데이터를 외부로 반출할 수 없으며, 기존 학습 데이터와의 도메인 차이를 고려해야 하므로, 사전 학습된 모델을 기반으로 한 전이 학습이 필수적이며 최근의 사전 학습된 모델들이 높은 하드웨어 요구량을 가지고 있다.

본 연구에서는 PEFT와 양자화를 통해 낮은 하드웨어 요구량에서도 최소한의 정확도 손실을 갖는 효과적인 모델을 개발하기 위해 신경망 경량화 기술인 양자화와 PEFT 기법을 적용하였다.

적용 결과 학습시 요구되는 GPU의 사용량이 32GB에서 5.7GB로 82.19% 감소함을 확인하였다. 동일한 조건에서 동일한 데이터로 모델의 성능을 평가한 결과 동일 학습 횟수에선 LoRa와 양자화가 적용된 모델의 오류가 기본 모델보다 53.34% 증가함을 확인하였다. 모델 성능의 감소를 줄이기 위해서는 학습 횟수를 더 증가시킨 결과 오류 증가율이 29.29%로 동일 학습 횟수보다 더 줄어듬을 확인하였다.

본 연구를 통해 이를 통해 초거대 AI 모델의 학습 플랫폼에 따른 PEFT 및 양자화 기법 적용과 그에 따른 모델 성능 하락 감소를 위한 학습 횟수 증가를 통해 온 디바이스 초거대 AI 모델 학습이 가능할것으로 판단된다.

향후 연구로 LoRa 이외의 새로운 PEFT 기법 적용 또는 더 많은 학습 횟수 증가를 통해 모델 성능과 PEFT 및 양자화 적용의 관계의 연구를 추가로 진행할 예정이다.

REFERENCES

[1] Freely scalable and reconfigurable optical hardware for deep learning - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Number-of-parameters-ie-weights-in-recent-landmark-neural-networks1-2-31-43_fig1_349044689 [accessed 18 Dec, 2023]

[2] Nagel, Markus, et al. "A white paper on neural network quantization." arXiv preprint arXiv:2106.08295 (2021).

[3] Mangrulkar, Sourab, et al. "PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods." 2022. GitHub, https://github.com/huggingface/peft.

[4] H.J. Kim, C.G. Lyuh. "Trends in Lightweight Neural Network Algorithms and Hardware Acceleration Technologies for Transformer-based Deep Neural Networks." Electronics and Telecommunications Trends 38.5 (2023): 12-22.

[5] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).

[6] Bang, Jeong-Uk, et al. "Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition." Applied Sciences 10.19 (2020): 6936.