Chat-GPT 같은 대형 언어 AI 모델을 가속화 하기 위한 Outlier Quantization 프레임워크 개발 및 IP 사업화 추진

2024년 1학기 강의페어링 기계공학과 조영민, 지도교수 신호재

개요

강의 페어링 요소 : 본 프로젝트는 2024-1 강의페어링 과목의 일환으로 매틀랩 및 기본 프로그래밍 능력을 함양하는 어드벤처 디자인 과목과 IP 사업화 방법론을 다룬 미래산업과 기술창업론 과목을 융합하여 설계되었다.

배경 : 최근 2000억개 이상의 파라미터를 가지는 Chat-GPT와 같이 AI 모델의 크기가 점점 증가하고 있다.
이렇게 모델의 크기가 증가함에 따라 모델을 구동하는데 걸리는 시간과 전력도 마찬가지로 증가한다.

이를 해결하기 위해서 Quantization(양자화)라는 방법을 사용해서 AI 모델의 파라미터를 정확도의 손실 없이 압축하는 방법이 대두되고 있다.

특히 각 파라미터를 매우 작은 크기(4bit 이내)로 압축하는 경우 Outlier Quantization 이라는 방법이 주로 사용되는데

이는 모든 가중치를 같은 크기로 압축하는 것이 아니라 중요한 가중치는 압축하지 않고, 중요하지 않은 가중치만 압축하는 일종의 선별 압축 방법이다.

솔루션 : 본 솔루션은 Outlier Quantization된 가중치를 NPU(AI 프로세서)에서 효율적으로 연산하기 위해서 제시한 방법으로 이를 통해서 **AI 모델의 속도 향상, 전력 소비량 감소**를 달성할 수 있으며, 현재 이 방법은 아주대학교 지식재산 인재 양성 산업의 지원을 받아 **특허 출원을 완료**하였고, 향후 이를 통해 수익화를 기대하고 있다.

Outlier Quantization이란?

Outlier Quantization이란 기존 양자화에서 모든 가중치를 일률적으로 압축시킴으로 인해서 발생하는 정확도 손실을 개선하기 위해서 도입된 방법이다.

이 방법은 전체 가중치 중 결과값에 중대한 영향을 미치는 중요한 가중치는 전체의 1~5% 정도 밖에 되지 않으며 이를 Hessian Matrix 혹은 Activation에 근거한 관찰을 통해서 찾아낼 수 있다는 아이디어에 근거하여 상위 1%의 중요 가중치는 압축하지 않고 나머지 중요하지 않은 99%의 가중치는 압축하는 방법이다.

이러한 방법을 통해서 정확도를 손실하지 않고 모델 크기를 극한으로 압축시킬 수 있으며, 이는 곧 AI 모델 속도 향상, 전력 소비량 감소, 메모리 소비량 감소 등 여러 긍정적인 효과로 이어진다.

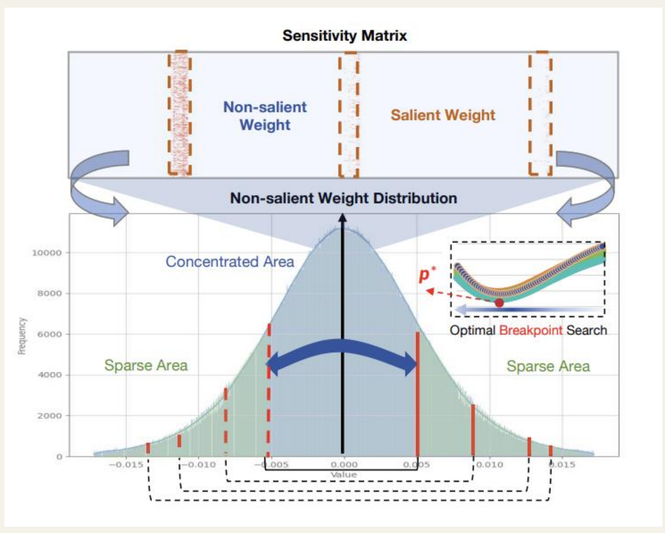

Figure 1 AI 모델에서 중요 가중치의 분포도

Outlier Quantization에서 HW 최적화 문제

Outlier Quantization은 모델 크기를 극한으로 압축시키면서도 정확도를 유지할 수 있다는 장점이 있지만, 압축된 가중치를 프로세서에서 불러 오면서 지연 현상이 생긴다는 단점도 있다.

구체적으로는 압축되지 않은 가중치와 압축되지 않은 가중치들이 혼재되어 있다 보니 HW에서 이를 처리할 때 각 데이터를 분리한 다음 처리해야 한다는 문제이다.

이는 연산 작업을 2번에 거쳐서 나눠 처리해야 한다는 비효율성을 야기한다.

이를 해결하기 위해서 본 연구에서는 CSC 기반 수정 압축 알고리즘과 그를 연산하기 위한 전용 HW 시스템을 제안한다.

구체적으로 각 가중치를 사전에 정해진 규칙에 따라 압축해 놓은 다음 Off-Chip Memory에서 On-Chip Memory로 불러올 때 바로 해독하여 전용 On-Chip Buffer 및 연산 데이터 패스를 통해서 처리하도록 하는 방법이다.

제안된 동적 압축 프레임워크

CSC 알고리즘은 기존까지는 가중치의 상당 부분 0으로 구성된 희소 행렬을 압축하기 위해서 사용되던 압축 방법으로 0이 아닌 데이터의 위치를 표기할 수 있는 알고리즘이다.

이를 통해서 별도의 처리가 필요한 특정 데이터를 Index 형태로 표현할 수 있다.

그러나 이는 일반적 G/NPU 기반 LLM에서 HW적 최적화가 덜 되어 있다.

구체적으로는 기존 CSC 알고리즘처럼 나머지 데이터를 0인 희소행렬로써 무시하는 것이 아니라, Low Quant된 Non-Outlier Data를 포함할 수 있도록 새로운 Data Vector를 포함시켜야 하며
전체 Weight Matrix가 아니라 각 Tile 별로 CSC를 적용하여 프로세서 런타임 동안 실시간으로 데이터를 재구성 할 수 있어야 한다.

본 연구에서는 가상의 가중치 행렬을 바탕으로 각 행렬의 특성에 따라 수정된 CSC 알고리즘 혹은 Run Time Length Encoding, Huffman Encoding을 동적으로 적용하여 G/NPU를 위한 새로운 압축 프레임워크를 제시한다.

이는 기존의 Outlier Quantization에서 사용하던 고정적 Bitmap 압축 방식에 비해 더 성능이 좋게 압축이 가능하다.

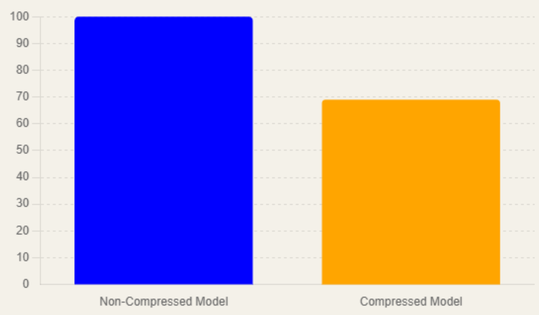

Figure 2 Tile별 CSC를 적용한 압축률 비교 (31% 압축)

HW 적용

또한 이를 잘 수행하기 위해서 전용 HW 유닛을 제안한다.
해당 HW 유닛은 크게 Outlier들을 따로 저장하는 Outlier Buffer와 Quantize된 데이터를 저장하는 Quantize Weight Buffer, 그리고 Position Data를 저장하는 Position Vector Controller로 구성되며 해당 데이터는 각 PE(Processing Element) 유닛의 연산 단위에 맞춰 비순차적 연산 (Out-of-Order) 방식으로 처리된다.

PE는 보통 Systolic Array 혹은 Adder Tree 기반으로 구성될 수 있으며, HW 특성에 따라 고유의 연산 단위를 가진다.

이 고유의 연산적 특성에 따라 본 프레임워크는 On-Chip Buffer를 위에서 제시한 방법대로 분할하고 각 HW의 연산적 특성을 인식하여 Out-of-Order를 실행시키는 방식을 사용함으로써 현형 G/NPU에 Outlier Inference 수행 능력을 향상시킨다.

GPU의 경우 CUDA로 구현될 예정이며, NPU의 경우 C++ 기반의 GCC Compiler Extension을 통해서 구현될 예정이다.

목표 적용 플랫폼	적용 방법
Nvidia GPU	Cuda
NPU	GCC Compiler Extension

Figure 3 타겟 플랫폼 및 적용 방안

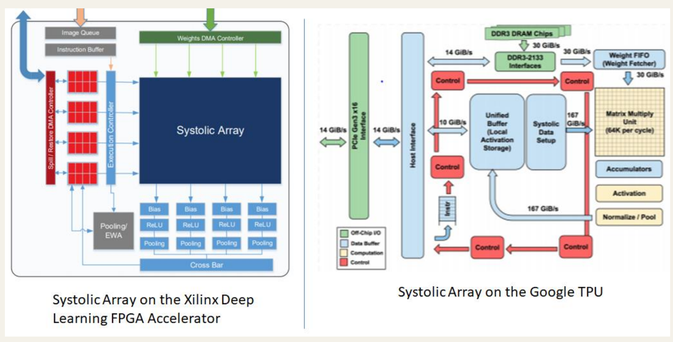

Figure 4 적용 가능 연산 유닛 예 (Systolic Array)

기대 효과

다양한 Data Precision과 압축 솔루션을 HW 레벨에서 제공할 수 있는 Outlier Quantization 프레임워크를 개발함으로써

기존 LLM Model을 더욱 압축하여 효율적인 AI 연산이 가능하게끔 한다.

특히 데이터센터의 전력 소비량은 전세계 전력 생산량의 1% 정도이고, 이는 Chat-GPT와 같은 초대형 AI 모델이 보편화 됨에 따라서 더욱 커질 것이다.

해당 압축 프레임워크가 보편화된다면 기존 대비 더 작은 모델로 효율적으로 AI 서버를 구동할 수 있어, 데이터 센터의 전력 부담 감소, AI 모델 반응 속도 향상 등 각종 긍정적 효과를 얻을 수 있다.

또한 GPU 뿐만 아니라 추론용 NPU 칩셋에서도 사용할 수 있게끔 함으로써 더 넓은 생태계에 적용이 가능하다.

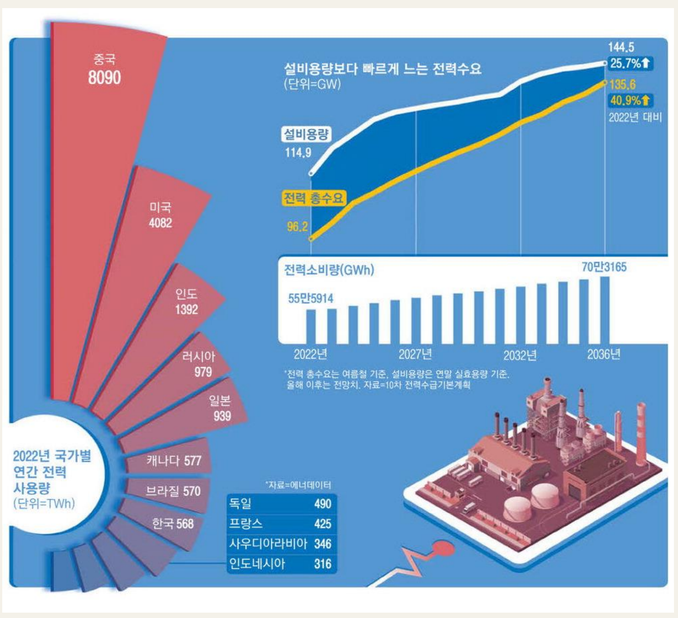

Figure 5 데이터센터 전력 소비량

사업화 전략 및 현황

현재 해당 프레임워크는 중소벤처기업부로부터 5천만원 R&D 개발 자금을 지원받았으며, 관련해서 특허 2건을 출원하였다.

현재는 LLM Model에서 SW 적으로 최적화 압축률을 확인하였으며, 향후 과제는 HW에서의 최적화 지원이다.

또한 현재까지 진행 상황을 IEEE Super Computing 2024 Conference에 출품할 계획이다.

해당 프레임워크는 24년까지 총 3개의 국내 특허 및 1개의 미국 특허를 출원할 예정이며 획득한 지재권을 바탕으로 다양한 AI 개발 기관에게 본 프레임워크 판매할 계획이다.

참고문헌

•Shang, Y., Yuan, Z., Wu, Q., & Dong, Z. (2023). *PB-LLM: Partially Binarized Large Language Models*. http://arxiv.org/abs/2310.00034

Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., Gan, C., & Han, S. (2023). *AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration*. http://arxiv.org/abs/2306.00978

Abdel-Aziz, H., Shafiee, A., Shin, J. H., Pedram, A., & Hassoun, J. H. (2021). *Rethinking Floating Point Overheads for Mixed Precision DNN Accelerators*. http://arxiv.org/abs/2101.11748

Ashkboos, S., Markov, I., Frantar, E., Zhong, T., Wang, X., Ren, J., Hoefler, T., & Alistarh, D. (2023). *QUIK: Towards End-to-End 4-Bit Inference on Generative Large Language Models*. http://arxiv.org/abs/2310.09259

후원 기관

본 연구는 아주대학교 지식 재산 인재 양성사업과 중소벤처기업부로부터 자금 지원을 받았습니다.