Transformer의 Attention score를 위한 하드웨어 구조 설계

이승현, 이광엽, 김재현
서경대학교 컴퓨터공학과

E-mail: sunflower117@skuniv.ac.kr,

요 약

본 논문은 Transformer 모델의 Attention score를 하드웨어로 구현하는 방법과 그 성능 검증에 대한 연구를 제시하였다. 최근 자연어 처리, 기계 번역 및 컴퓨터 비전 분야에서 많은 성과를 거두고 있는 Transformer 모델은 Attention 메커니즘이 핵심 요소로 인정되며, 모델의 학습과 예측 성능에 큰 역할을 수행한다. 그러나 Attention 메커니즘은 복잡한 행렬 연산과 데이터 흐름으로 인해 많은 비용이 발생하는 단점이 있다. 본 논문에서는 이를 줄이기 위해 16bit 부동소수점 형식을 활용하여 하드웨어 모듈을 설계한다. Verilog HDL을 사용하여 곱셈기와 덧셈기, 내적 연산기를 구현하여 Attention score를 계산하였으며 Xilinx에서 제공하는 16bit 부동소수점 IP와 Pytorch에서 사용하는 float16 tensor 자료형을 활용하여 4x4 및 16x16 크기의 행렬에서 Attention score 의 연산 정확도를 비교하는 실험을 진행한다. Xilinx IP를 활용한 결과 평균 0.00095의 오차를 보였다. Pytorch 활용 결과, 평균 0.0013의 오차 값을 보였으며, 16x16 크기의 행렬에서는 평균 0.0718의 오차가 측정되었다.

Ⅰ. 서 론

최근 Transformer 모델[1]은 자연어 처리, 기계 번역 및 컴퓨터 비전 분야에서 많은 성과를 거두고 있다.[2] 특히 Transformer 모델의 핵심 요소 중 하나인Attention 메커니즘은 Transformer 모델의 학습과 예측 성능에 큰 영향을 미치는 중요 요소 중 하나로 인정되고 있다. 하지만 Attention 메커니즘은 많은 행렬 연산과 복잡한 데이터 흐름을 필요로 하기 때문에 많은 비용이 발생한다는 단점이 있다.[3] 이를 줄이기 위한 방법으로 본 논문에서는 Attention 메커니즘의 주요 특성 중 하나인 Attention score를 하드웨어를 통해 구현하는 방식을 제시한다.

본 논문에서는 Transformer 모델의 Attention socre를 구하기 위해 16bit 부동소수점 형식을 활용하여 하드웨어 모듈을 설계하였고, Xilinx에서 제공하는 16bit 부동소수점 라이브러리와 Pytorch에서 사용하는 float16 tensor 자료형을 이용하여 Attention score의 정확도를 비교하여 성능을 검증하였다.

Ⅱ. 본 론

1. Transformer의 Attention score

Transformer의 Attention은 입력 시퀀스의 각 위치마다 해당 위치와 다른 위치들 사이의 관계를 강조하는 메커니즘이다. 이러한 관계를 통해 모델은 입력 시퀀스의 각 단어나 토큰 간의 상호 의존성을 파악하고 주요 정보에 집중할 수 있도록 한다. Attention score는 입력 시퀀스의 다른 단어나 토큰과의 상호 관계를 반영하므로 모델의 중요 정보를 놓치지 않고 처리할 수 있도록 도와준다. Transformer 모델에서 정의하는 Attention score값은 식(1)을 이용하여 계산한다.

$$Attention\ score = \frac{Q \cdot K^T}{\sqrt{d^k}} \quad (1)$$

2. 16bit 부동소수점

본 논문에서는 하드웨어에서 수 처리를 하기 위해 10진으로 표현된 수를 [그림 1]과 같은 16bit 부동 소수점 방식을 사용하여 2진수로 변환하여 이를 이용한 곱셈기와 덧셈기, 내적 연산기를 Verilog HDL을 사용하여 설계 후 Attention score를 구하였다.

[그림 1] 16bit 부동 소수점 구조

3. 실험 과정 및 결과

Verilog HDL로 설계한 Attention score의 정확도를 검증하기 위해 Xilinx에서 제공하는 16bit 부동소수점 곱셈, 덧셈기 IP를 활용하여 구한 Attention score와 Pytorch에서 사용하는 16bit 부동소수점 tensor 라이브러리를 활용하여 구한 Attention score를 이용해 각각 그 값들을 비교하여 정확도를 확인하였다.

4x4 크기의 행렬로 Xilinx IP를 활용한 Attention score와 본 논문에서 설계한 Attention score의 값을 소수점 4번째 자리까지 비교한 결과 0 ~ ±0.0041 정도의 오차가 나왔으며 평균 0.00095 정도의 오차 값을 보였다. 또한 Pytorch를 활용한 Attention score와 비교한 결과 0 ~ ±0.0078 정도의 오차가 나왔으며 평균 0.0013 정도의 오차 값을 보였다.

16x16 크기의 행렬로 Pytorch의 라이브러리를 활용하여 구한 Attention score와 본 논문에서 설계한 Attention score의 값을 소수점 4번째 자리까지 비교한 결과 0 ~ ±1.1133 정도의 오차가 나왔으며 평균 0.0718 정도의 오차 값을 보였다.

Ⅲ. 결 론

본 논문에서는 Transformer 모델의 Attention score를 하드웨어로 구현하기 위한 방법과 구현한 모듈과 실제 사용되는 라이브러리들과 값을 비교하여 정확도를 측정하였다. Attention score를 구하기 위하여 16bit 부동 소수점 형식을 활용하여 Verilog HDL로 설계하였으며 Xilinx IP와 Pytorch를 활용하여 실험을 진행하였다.

실험 결과로 4x4 크기의 행렬에서 Xilinx에서 제공하는 16bit 부동 소수점 IP로 구한 Attention score와의 비교에서 평균 0.00095의 오차 값을 보였고, Pytorch에서 사용하는 16bit 부동소수점 라이브러리로 구한 Attention score와의 비교에서 0.0013정도의 오차값을 보였다. 이보다 크기가 큰 16x16 크기의 행렬과 Pytorch에서의 float16형태와의 비교에서 평균 0.0718의 오차로 더 큰 행렬에서도 좋은 정확도를 유지한다는 것을 보여준다.

※ 사사의 글

이 성과는 정부(과학기술정보통신부, 교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2019M3E7A1113102).

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. “Attention Is All You Need,” Jun. 2017.

[2] W. Ye, X. Zhou, J. T. Zhou, C. Chen, K. Li. “Accelerating attention mechanism on FPGAs based on efficient reconfigurable systolic array,” July. 2022.

[3] S. Lu, M. Wang, S. Liang, J. Lin, Z. Wang, “Hardware Accelerator for Multi-Head Attention and Position-Wise Feed-Forward in the Transformer,” Sep. 2020.