Edge TPU의 온도 관리를 위한 Task-level 온도 모델링 기법

한창헌, 오상은*
아주대학교 AI융합네트워크학과, *아주대학교 소프트웨어학과

{ehwjs1914, *sangeunoh}@ajou.ac.kr

Task-level Thermal Modeling for Temperature Management of Edge TPU

Changhun Han, Sangeun Oh*
Department of Artificial Intelligence Convergence Network, Ajou Univ.
*Department of Software and Computer Engineering, Ajou Univ.

요 약

Edge TPU는 딥러닝 연산을 위한 고효율 저전력 가속기로서 다양한 엣지 컴퓨팅 응용 분야에서 활용되고 있다. 그러나 Edge TPU의 온도 상승은 성능 저하, 안정성 감소, 수명 단축 등의 문제를 야기할 수 있어 이를 관리하기 위한 온도 모델링이 필요하다. 본 논문에서는 Edge TPU의 온도 예측을 위한 task-level 온도 모델링 기법을 제안한다. 제안하는 기법은 다양한 딥러닝 태스크의 워크로드에 따른 CPU와 Edge TPU의 전력 소모량을 추정하고, 이를 기반으로 steady temperature 모델을 활용하여 Edge TPU의 최종 수렴 온도를 예측한다. 실험을 통해 제안된 기법이 다양한 워크로드에 대해 Edge TPU의 온도를 정확히 예측할 수 있음을 확인하였다. 평균 예측 오차는 0.6℃로 나타났다. 본 연구는 Edge TPU의 온도 특성을 고려한 효과적인 온도 예측 모델을 제시함으로써, 추후 온도 관리 기법 개발의 기반이 될 것으로 기대된다.

Ⅰ. 서 론

최근 딥러닝 기술의 발전과 함께 Edge TPU (Tensor Processing Unit)의 활용이 크게 증가하고 있다 [1]. Edge TPU는 딥러닝 연산에 특화된 저전력 고효율 처리 장치로, 엣지 컴퓨팅 환경에서 실시간 딥러닝 서비스를 제공하는 데 핵심적인 역할을 하고 있다. 기존에는 딥러닝의 높은 연산량과 긴 처리 시간으로 인해 엣지 디바이스에서의 AI 적용이 제한적이었으나, Edge TPU의 등장으로 스마트 팜, 스마트 팩토리, 헬스케어 등 다양한 분야에서 저전력 고성능 지능형 시스템의 개발이 가능해졌다.

Edge TPU의 활용이 증가함에 따라 온도 관리의 중요성도 함께 대두되고 있다. Edge TPU는 고효율 연산을 수행하기 위해 설계되었지만, 장시간 연속적으로 사용될 경우 칩 온도가 상승할 수 있으며, 특히 소형화, 경량화되는 엣지 디바이스의 방열 설계 제약으로 인해 온도 상승의 위험이 있다. Edge TPU가 고온에 장시간 노출될 경우 성능 저하, 실시간성 보장 어려움, 안정성과 신뢰성 저하, 수명 단축 등 다양한 문제점이 발생할 수 있어, Edge TPU의 온도를 적절한 수준으로 관리하는 것은 시스템의 성능, 안정성, 비용 효율성 측면에서 매우 중요하다 [1].

기존의 Edge TPU에서의 온도 관리는 주로 DVFS (Dynamic Voltage and Frequency Scaling) 나 쿨링 팬 제어 등의 방법을 통해 Edge TPU의 온도를 관리하고자 하였다 [2]. 그러나 이러한 방법들은 Edge TPU가 사용되는 다양한 환경과 워크로드 특성을 고려하지 않아 한계가 있다. 예를 들어, DVFS를 적용할 경우 실시간 태스크의 deadline을 맞추기 어려운 문제가 발생할 수 있으며, 팬 제어 방식은 추가적인 전력 소모 및 밀폐된 환경에서 열 발산 제한, 진동과 소음 등의 이유로 저전력 엣지 디바이스에 적용하기 어려울 수 있다.

본 연구에서는 이러한 한계점들을 극복하고 보다 효과적으로 Edge TPU의 온도를 관리하기 위한 기반 기술로서 온도 모델링 기법을 제안한다. 제안하는 기법의 핵심은 Edge TPU에서 수행되는 워크로드의 종류와 강도에 따라 발생하는 열을 정확히 예측하는 것이다. 이를 위해 Edge TPU 온도에 영향을 미치는 요인들을 분석하고 태스크 단위의 온도 예측 모델을 수립한다. 이렇게 개발된 온도 모델은 주어진 태스크 셋에 대해 Edge TPU의 온도가 임계치를 초과할 것인지 여부를 사전에 예측할 수 있게 함으로써, 적절한 온도 관리 전략을 수립하는 데 활용될 수 있다.

Ⅱ. Edge TPU 온도 모델링

Edge TPU의 온도를 효과적으로 관리하기 위해서는 다양한 딥러닝 태스크를 수행할 때 발생하는 열을 정확히 예측할 수 있어야 한다. 본 연구에서는 이를 위해 태스크 단위의 온도 모델링 기법을 제안한다. 제안하는 방법은 각 딥러닝 태스크를 다양한 utilization으로 수행하였을 때, CPU와 Edge TPU의 전력 소모량을 측정하고 이를 기반으로 steady temperature를 추정한다.

임베디드 보드 상에서 Edge TPU는 일반적으로 CPU와 함께 사용되며, 두 구성 요소는 서로 근접한

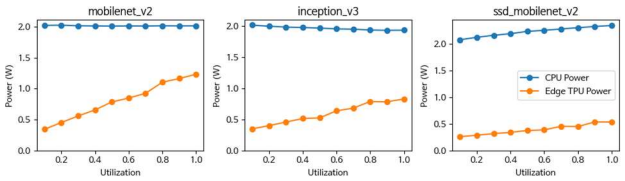

그림 1 태스크 별 CPU 와 Edge TPU 의 전력 소모량

위치에 배치되는 경우가 많다. 이로 인해 CPU에서 발생한 열이 Edge TPU로 전달되어 Edge TPU의 온도 상승에 영향을 미치게 된다. 따라서 Edge TPU의 온도를 정확히 예측하기 위해서는 CPU의 발열로 인한 열 간섭 효과를 고려해야 한다. 이를 위해 본 연구에서는 CPU와 Edge TPU의 전력 소모량을 모두 온도 예측 모델에 반영하였다.

우선 Edge TPU에서 수행되는 대표적인 딥러닝 태스크들을 선정하고, 각 태스크를 다양한 워크로드로 수행하며 전력 소모량을 측정한다. 이때 전력 소모량은 CPU와 Edge TPU의 개별적인 전력 소모량의 합으로 계산된다. CPU의 전력 소모량은 PMU (Performance Monitoring Unit)를 통해 측정되는 하드웨어 이벤트를 활용하여 추정하며, Edge TPU의 전력 소모량은 전체 시스템 전력에서 CPU와 다른 구성 요소들의 전력을 제외하여 계산한다.

이렇게 수집된 태스크 단위의 전력 소모량 데이터를 기반으로 steady temperature를 예측하기 위한 모델을 구축한다. 본 연구에서는 열 회로 모델[3]을 기반으로 steady temperature 모델을 구축하였으며, 태스크의 utilization에 따른 CPU와 Edge TPU의 전력 소모량을 입력으로 사용한다. 이를 통해 주어진 딥러닝 워크로드에 대한 Edge TPU의 최종 수렴 온도를 예측할 수 있다.

$$T_{\infty} = T_{amb} + P_{CPU} \cdot R_{CPU} + P_{TPU} \cdot R_{TPU} \quad (1)$$

여기서 T_{∞}는 steady temperature, T_{amb}는 주변 온도, R_{CPU} 와 R_{TPU} 는 각각 CPU와 Edge TPU의 열 저항, 그리고 P_{CPU} 와 P_{TPU} 는 각각 CPU와 Edge TPU의 소비 전력이다. 이 모델은 CPU와 Edge TPU 간의 열 간섭을 고려함으로써 보다 정확한 Edge TPU 온도 예측을 가능하게 한다.

Ⅲ. 실험

제안된 온도 모델링 및 예측 기법의 유효성을 검증하기 위해 실제 하드웨어 플랫폼에서 실험을 수행하였다. 실험에는 Coral Dev Board를 사용하였으며, 이는 NXP i.MX 8M SoC (Quad-core ARM Cortex-A53)와 Google Edge TPU로 구성되어 있다.

소비 전력 측정을 위해 Monsoon Power Monitor를 사용하였다. 이를 통해 Coral Dev Board 전체의 소비 전력을 높은 샘플링 레이트로 측정하였다. CPU 소비 전력의 경우 Linux의 Perf 도구를 활용하여 PMU 이벤트를 모니터링하여 추정한다. Edge TPU 온도는 센서 값을 제공하는 시스템 파일을 읽어와 측정하였다.

실험에서는 다양한 딥러닝 모델들을 태스크로 선정하였다. MobileNet V2, Inception V3 등 이미지

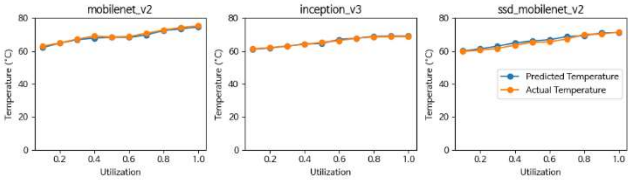

그림 2 예측된 steady temperature 와 실제 측정값 비교

분류를 위한 모델과, SSD MobileNet V2와 같은 객체 인식 모델이 포함되었다. 각 모델은 TensorFlow Lite로 변환되어 Edge TPU에서 실행되었다.

그림 1은 태스크 별 utilization에 따른 CPU와 Edge TPU의 전력 소모량을 보여준다. CPU의 경우 utilization의 변화에 크게 영향을 받지 않는 모습을 보인다. 반면 Edge TPU는 utilization에 따른 전력 소모량의 변화가 상대적으로 크다. 이는 태스크의 utilization이 증가함에 따라서 Edge TPU의 활용이 증가하기 때문이다.

그림 2는 예측된 steady temperature와 실제 측정값의 비교 결과를 나타낸다. 다양한 딥러닝 태스크와 utilization에 대해 예측 모델이 실제 수렴 온도를 잘 예측함을 확인할 수 있다. 예측 오차는 평균 0.6도로 나타났으며, 이는 제안된 온도 모델링 및 예측 기법이 Edge TPU의 열 특성을 잘 반영함을 의미한다.

Ⅳ. 결론

본 논문에서는 Edge TPU의 온도 관리를 위한 새로운 모델링 및 예측 기법을 제안하였다. 제안된 기법은 딥러닝 워크로드의 특성을 고려한 task-level 온도 모델링과 steady temperature 예측을 통해 Edge TPU의 발열 특성을 정확히 파악하고 예측할 수 있다. 다양한 딥러닝 태스크를 대상으로 utilization에 따른 CPU와 Edge TPU의 전력 소모량을 측정하고 분석하였으며, 이를 통해 각 태스크의 발열 특성을 정량화하였다. 또한, steady temperature 모델을 활용하여 Edge TPU의 최종 수렴 온도를 예측하였으며, 실험 결과 높은 정확도를 보였다. 본 연구의 결과는 Edge TPU를 활용하는 다양한 응용 분야에서 온도 관리 문제를 해결하는 데 기여할 수 있을 것으로 기대된다. 향후에는 제안된 온도 모델링 및 예측 기법을 실제 시스템에 적용하고, 온도를 일정 수준 아래로 낮추기 위한 메커니즘을 개발할 것이다.

참 고 문 헌

[1] T. Chantem, Y. Xiang, X. Hu, and R. P. Dick, “Enhancing multicore reliability through wear compensation in online assignment and scheduling,” in Proc. DATE, 2013, pp. 1373-1378.

[2] J. Kong, S. W. Chung, and K. Skadron, “Recent thermal management techniques for microprocessors,” ACM Comput. Surveys, vol. 44, no. 3, p. 13, 2012.

[3] T. L. Bergman, Introduction to Heat Transfer. Hoboken, NJ, USA: Wiley, 2011.