모바일 AI 에서 발열과 메모리가 딥러닝 기반 어플리케이션 성능에 미치는 영향 연구

최평준, 김정수, 곽정호
대구경북과학기술원

pyeongjun.choi@dgist, jeongsoo98@dgist.ac.kr, jeongho.kwak@dgist

A Study on the Impact of Heat and Memory on Performance of Deep Learning-based Application in Mobile AI systems

Choi Pyeong Jun, Kim Jeong Soo, Kwak Jeong Ho
Daegu Gyeongbuk Institute of Science & Technology

요 약

최근 몇 년 간 번역, 텍스트 추출, AI 비서 등 다양한 분야에서 딥러닝 기반 어플리케이션에 대한 소비자의 수요는 꾸준히 증가하고 있다. 모바일 프로세서의 비약적인 성능 개선과 딥러닝 모델의 소형화에 힘입어 많은 딥러닝 기반 어플리케이션들이 상용화되었지만 여전히 발열, 에너지 소비, 메모리 사용량 등 모바일 단말의 한계 때문에 소비자와 서비스 제공자 모두가 만족하는 최적의 성능을 내기는 어려운 상황이다. 우리는 다양한 모바일 AI 단말을 활용한 실험을 통해 단말의 온도와 가용 메모리가 딥러닝 기반 어플리케이션 성능에 미치는 영향을 보이고, 최적의 성능을 달성하기 위한 단초를 제공한다.

Ⅰ. 서 론

최근에는 스마트폰이나 XR 기기와 같은 모바일 기기에서 딥러닝 기반 서비스에 대한 수요가 늘어나고 있다. 스마트폰 제조업체는 기본 카메라 어플리케이션 내에서 텍스트 추출 및 얼굴 인식 기능을 제공하며, 음성 인식 기능이 있는 AI 비서를 제공한다. 딥러닝 모델의 소형화와 모바일 프로세서의 비약적인 성능 향상이 이러한 딥러닝 기반 어플리케이션들의 상용화를 가능케했다. 그러나 제한된 메모리를 사용하고 액티브 쿨링 시스템을 탑재할 수 없는 모바일 AI 단말의 특성상 딥러닝 기반 어플리케이션을 제공함에 있어 발열과 메모리 측면에서 한계가 있을 수밖에 없다.

딥러닝 기반 어플리케이션을 구동하기 위한 강력한 모바일 프로세서는 높은 CPU/GPU 주파수를 유지하기 위해 많은 에너지를 소비하고 많은 열을 발생시킨다. 이때, 모바일 AI 단말 별로 설정된 제한 온도를 초과하면 단말의 성능을 제한하여 추가적인 발열을 억제하는 써멀 쓰로틀링이 발생한다. 이 경우 사용자는 급격한 성능 저하를 경험하게 된다. 또한, 써멀 쓰로틀링 발생 이전에도 사용자가 직접 접촉하는 휴대폰과 같은 모바일 AI 단말의 경우 높은 온도 인해 불쾌함을 느끼게 된다.

메모리 사용량은 모바일 AI 기기에서 딥러닝 기반 어플리케이션의 성능을 저해하는 또 다른 요인이다. 모바일 AI 단말은 백그라운드에서 여러 어플리케이션을 동시에 실행하는데, 각 어플리케이션이 모두 풍부한 메모리 용량을 원하기 때문에 메모리는 항상 부족한 자원 중 하나이다. 메모리가 부족하면 모바일 AI 단말은 저장 장치를 메모리처럼 사용하는 스왑 메모리를 사용한다. 이 스왑 메모리는 원래 메모리보다 속도가 느리기 때문에 이를 사용하는 어플리케이션의 반응 속도와 성능 또한 저하된다.

Ⅱ. 본 론

우리는 모바일 AI 단말에서 발열과 메모리가 딥러닝 기반 어플리케이션 성능에 미치는 영향을 확인하기 위해 Samsung Galaxy S21 Ultra 와 nvidia Jetson TX2 를 사용한 실험을 진행했다. 먼저 발열에 의한 영향을 보기 위해 S21 Ultra 에서 YOLOv8l-cls [1] 모델을 이용해 15 분간 이미지 분류를 수행했을 때 단말의 변화를 관찰했다.

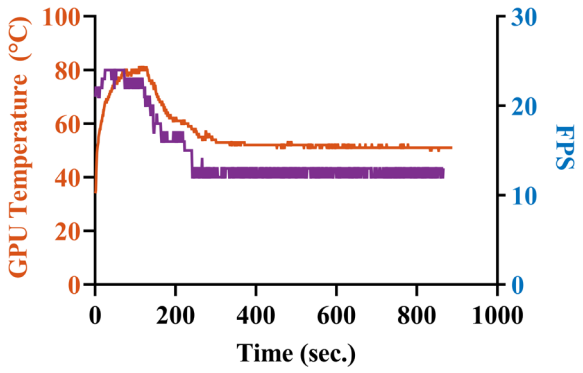

Figure 1. 시간에 따른 GPU 온도 및 FPS 변화

Fig. 1 은 시간에 따른 GPU 온도와 FPS 변화를 나타낸 그래프다. 약 150 초 이전까지는 20 이상의 FPS 를 보이며 준수한 성능을 보이다가 GPU 온도가 80 도에 도달하는 순간 성능을 크게 제한하여 점진적으로 12 FPS 까지 떨어지는 모습을 보인다. 제한된 성능 덕분에 발열은 어느정도 억제되는 모습을 보이나 이런 경우 사용자 경험은 크게 떨어지게 된다.

Fig. 2 는 해당 실험에서 이미지 한 장을 추론하는데 사

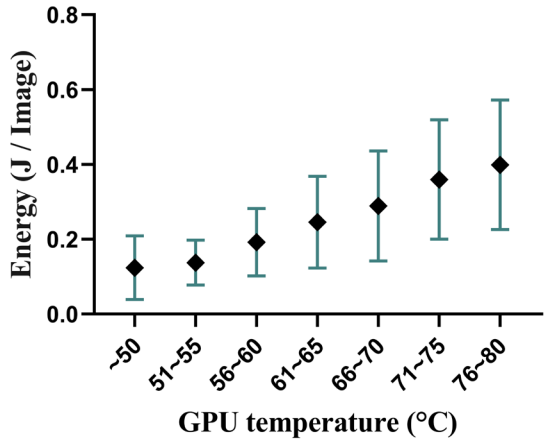

Figure 2. GPU 온도에 따른 이미지 당 에너지 사용량

용한 에너지를 GPU 온도 구간 별로 나눠서 평균과 표준 편차를 표현한 그래프다. GPU 온도가 높아질수록 평균적으로 이미지 한 장을 추론하는데 더 높은 에너지를 사용하는 것을 볼 수 있다. 특히 가장 낮은 온도 구간인 50 도 이하와 가장 높은 온도 구간인 76~80 도 구간은 평균적으로 3 배 이상의 차이를 보인다. 써멀 쓰로틀링이 발생하지 않더라도 GPU 온도에 따라 에너지 사용량이 크게 차이가 나므로 사용자의 선호에 따라 GPU 온도를 적정 수준으로 조절하는 것이 중요하다.

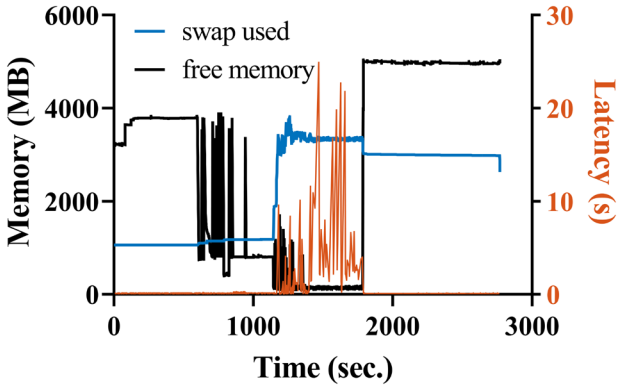

Figure 3. 가용 메모리 용량에 따른 이미지 추론 시간

단말의 가용 메모리 용량에 따른 이미지 추론 시간을 측정하기 위해 nvidia Jetson TX2 에서 메모리 부하를 조절해가며 Dynamic-OFA [2] 모델로 이미지 추론을 수행했다. 메모리 부하는 시작 10 분후에 3 GB 를 더하고, 20 분 이후 5 분 간격으로 1 GB 씩 추가로 더했으며, 30 분 이후엔 메모리 부하를 모두 해제했다. 이미지 추론에 걸리는 시간은 메모리 부하가 없는 경우 평균 0.1 초였으나 가용 메모리가 1 GB 아래로 줄었을 때 평균 0.15 초로 50% 증가하는 모습을 보였다. 특히, 스왑 메모리가 사용되는 구간 (즉, 20 분 이후)부터는 이미지 추론에 걸리는 시간이 급격히 증가하여 최대 250 배 늘어난 25 초까지 걸리는 것을 확인하였다. 정리하면, 이미지 추론에 걸리는 시간은 가용 메모리 용량에 반비례하게 증가할 수 있으며, 특히 스왑 메모리를 사용하는 경우 증가폭이 수백배에 이를 수 있게 된다. 때문에 모바일 AI 단말에서 딥러닝 기반 어플리케이션을 사용하는 경우 가용 메모리 용량을 적정 수준으로 유지하는 것이 중요하다.

Ⅲ. 결론

본 논문에서는 실험을 통해 모바일 AI 단말에서 발열과 메모리가 딥러닝 기반 어플리케이션의 성능에 미치는 영향을 알아보았다. 실험 결과에서 보인 것과 같이 모바일 AI 단말의 온도 및 가용 메모리 용량을 적정 수준으로 유지할 수 있다면 에너지 효율과 사용자 경험을 모두 만족할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1C1C1003030).

참 고 문 헌

[1] G. Jocher, A. Chaurasia, and J. Qiu, “YOLO by Ultralytics,” [Online]. Available: https://github.com/ultralytics/ultralytics

[2] W. Lou, L. Xun, A. Sabet, J. Bi, J. Hare, and G. V. Merrett, “Dynamic-ofa: Runtime dnn architecture switching for performance scaling on heterogeneous embedded platforms,” in Proc. of IEEE/CVF CVPR, Jun. 2021, pp. 3110– 3118.