열 및 메모리 제약 환경에서의 엣지 지원 온디바이스 인공지능 기법

김정수, 최평준, 곽정호
대구경북과학기술원

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

Edge-Assisted On-Device Artificial Intelligence in Heat and Memory Constrained Environments

Jeongsoo Kim, Pyeongjun Choi, Jeongho Kwak
DGIST

요 약

본 논문은 온디바이스 인공지능과 모바일 엣지 컴퓨팅 기술에서 모바일 장치의 제한된 리소스로 인해 발생할 수 있는 발열, 메모리 포화 등의 문제를 해결하고자 한다. 우리는 먼저 사전 실험을 통해 인공지능 어플리케이션 구동 시 네트워킹 및 프로세싱 자원과 다차원 심층신경망 (DNN) 모델 크기를 동적으로 제어했을 때 모바일 장치의 온도 및 메모리 제약에 미치는 영향을 자세하게 조사하였다. 이를 통해 온도 및 메모리 제약 조건이 있는 환경에서 목표 초당 프레임 처리 수 (FPS)를 유지하면서 추론 정확도를 최대화하는 것을 목표로 오프로딩, 동적 전압, 주파수 스케일링 및 DNN 모델 크기를 동시에 조정하는 임계값 기반 H&M알고리즘을 제안하고 실험을 통해 알고리즘의 성능을 검증한다.

Ⅰ. 서 론

최근 스마트폰 등 모바일 장치에서 인공지능 서비스에 대한 수요가 증가하고 있다. 이에 대응하여 스마트폰 제조업체에서는 문자 추출과 같은 온디바이스 인공지능 서비스를 제공하고 있다. 이와 같은 서비스는 모바일 장치의 컴퓨팅 성능 향상과 DNN 모델의 소형화 덕분에 가능해졌다. 그러나 추론 정확도가 높은 대규모 DNN 모델에는 여전히 상당히 많은 양의 컴퓨팅 자원과 메모리가 필요하여 모바일 환경에서는 실행 불가능하다.

모바일 엣지 컴퓨팅 (MEC)기술은 모바일 장치에서 높은 추론 정확도의 인공지능 서비스를 제공하기 위한 솔루션 중 하나로, 컴퓨팅 자원을 많이 요구하는 작업을 인근 MEC 서버로 오프로드하는 기술이다. MEC 기술을 통해 모바일 기기에서 정확도 높은 인공지능 서비스를 제공하기 위한 연구가 많이 진행되고 있다.

그러나 온디바이스 추론과 MEC 오프로딩 모두 FPS 등의 서비스 품질 (QoS)에 영향을 미치는 발열 문제를 겪고 있다. 온디바이스 인공지능을 위한 고성능 모바일 CPU/GPU 프로세서와 오프로딩을 위한 5G 모뎀은 모두 많은 열을 발생시킨다. 모바일 장치는 안전한 작동을 위해 장치의 온도가 열 임계값을 초과하면 장치 성능을 강제로 제한하여 추가적인 발열을 억제한다 [1].

모바일 장치는 메모리가 부족하면 NAND 플래시 등의 스토리지를 스왑 메모리로 사용하는데, LPDDR5와 같은 메모리보다 느려 처리 지연 시간이 길어진다. 일반적으로 DNN 모델의 크기가 커질 수록 인공지능 서비스의 추론 정확도도 높아진다. 그러나 대규모 DNN 모델을 모바일 장치에서 사용할 경우, 멀티태스킹에 의해 가용 메모리가 동적으로 변해 장치의 메모리가 부족해져 스왑 메모리를 사용하는 경우가 발생할 수 있고, 처리 지연 시간이 늘어날 수 있다.

우리는 사전 실험을 통해 모바일 장치에서 YOLOv8 모델을 사용한 이미지 분류 어플리케이션을 구동하여 네트워킹 및 프로세싱 자원과 다차원 DNN 모델 크기를 동적으로 제어했을 때 모바일 장치의 온도 및 메모리 제약에 미치는 영향을 자세하게 조사하였다.

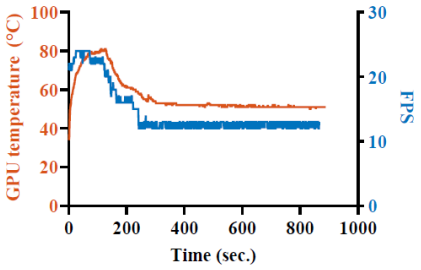

그림 1. GPU 온도 상승에 따른 FPS 변화

work	CPU max temp. (°C)	GPU max temp. (°C)	5G max temp. (°C)	Avg. FPS	Avg. power (W)
CPU-only	86	67	77	4.06	6.7
GPU-only	85	83	83	23.2	6.8
4K (G)	55	48	52	30	3.28
4K (B)	60	49	56	30	4.12

표 1. 모바일 장치 구성요소 사이의 온도 상관 관계

우선 그림 1을 보면 모바일 장치에서 GPU를 사용해 이미지 분류 어플리케이션을 구동하였을 때 GPU의 온도가 80°C에 도달하면 열에 의한 성능 제한으로 인해 FPS가 50% 저하되는 것을 알 수 있다. 표 1은 CPU, GPU, 5G 모뎀 중 하나의 구성 요소만 사용하여 이미지 분류 어플리케이션을 구동하였을 때 각 구성요소의 최대 온도를 나타낸 표이다. 이를 통해 하나의 구성 요소에서 발생하는 열이 다른 구성요소의 온도 상승에 영향을 주는 것을 확인하였다.

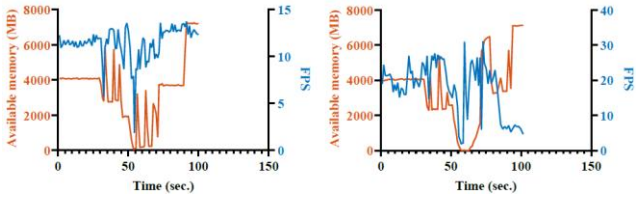

그림 2. 가용 메모리의 변화에 따른 FPS 변화

그림 2는 30초와 50초에 40초 동안 메모리 부하를 줬을 때, 가용 메모리와 FPS를 측정한 실험결과이다. 왼쪽 그래프는 각 지점마다 3500MB, 오른쪽 그래프는 각 지점마다 3700MB의 메모리 부하를 준 결과이다. 가용 메모리가 줄어듦에 따라 FPS가 감소하였고, 특히 가용 메모리가 거의 남아있지 않은 경우 FPS가 큰 폭으로 감소하는 것을 확인하였다.

결론적으로 모바일 인공지능 서비스는 제한된 컴퓨팅 성능 뿐만 아니라 발열 및 메모리 변동성에 의한 한계를 가지며, 이는 네트워크 인터페이스 사용, CPU/GPU 처리, 타 애플리케이션의 멀티태스킹 등 제어하거나 예측하기 어려운 요인이 얽혀 결정된다. 이에 우리는 5G 통신 인터페이스와 CPU/GPU 프로세서의 발열, DNN 모델의 메모리 사용량을 함께 고려한 동적 CPU/GPU 클럭 주파수, 오프로딩 속도, DNN 모델 세트 및 DNN 모델 선택 알고리즘을 제안한다.

Ⅱ. 본론

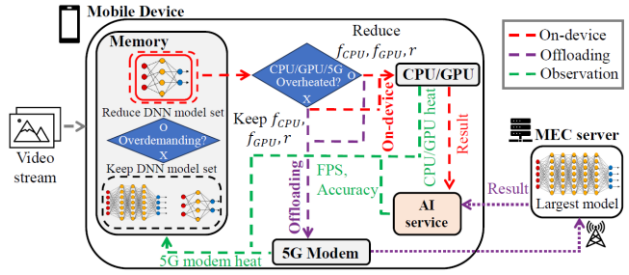

그림 3. 엣지 지원 온디바이스 추론 시스템을 위한 알고리즘 프레임워크

사전 실험에서 얻은 이해를 바탕으로 모바일 장치의 온도와 메모리 상황에 따라 작동하여 열 임계값을 넘지 않고 메모리 포화를 유발하지 않으면서 안정적인 목표 FPS와 정확도를 달성하는 H&M 알고리즘을 제안한다. 그림 3은 엣지 지원 온디바이스 추론 시스템을 위한 알고리즘 프레임워크이다. 우리는 모바일 장치에서 MEC 오프로딩과 온디바이스 추론을 모두 사용하여 인공지능 서비스를 제공하는 상황을 고려한다. MEC 오프로딩은 인공지능 어플리케이션의 워크로드를 스마트폰에서 MEC 서버로 전송하는 것으로, MEC 서버는 고성능 GPU를 통해 최대 규모의 DNN 모델로 추론하고 그 결과를 스마트폰으로 반환한다. MEC 오프로딩은 스마트폰에 비해 더 높은 추론 정확도와 더 빠른 추론 속도를 보장할 수 있지만, 시간에 따라 변하는 네트워크 상황에 의해 오프로드 할 목표 프레임 수가 항상 달성되지 않을 수도 있다. 온디바이스 추론은 메모리에 로드할 DNN 모델 세트를 결정, 메모리에 있는 DNN 모델 중 하나를 선택하고 CPU/GPU 클럭 주파수를 조절한다.

H&M 알고리즘의 동작순서는 우선 모든 결정 변수(예: CPU/GPU 클럭 주파수, 오프로딩 속도, DNN 모델 세트 및 DNN 모델)가 최대값으로 초기화된다. 매 타임 슬롯마다 H&M 알고리즘은 사용 가능한 메모리에 따라 메모리에 사전 로드된 DNN 모델을 선택한다. 다음으로, CPU, GPU, 5G 모뎀 중에서 최대 온도가 기준 온도를 초과하면 CPU/GPU 클럭 주파수와 오프로딩 속도를 감소시키고, 그렇지 않으면 증가시킨다.

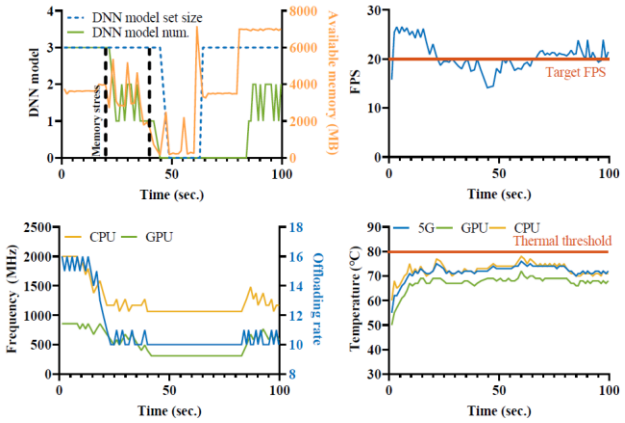

그림 4. 모바일 장치에서 수행한 H&M 알고리즘 실험 결과

그림 4는 모바일 장치에서 열과 메모리의 변화에 대한 H&M 알고리즘의 적응성을 보여준다. 온도가 상승함에 따라 H&M 알고리즘은 CPU, GPU 클럭 주파수 및 오프로드 속도를 낮추고 목표 FPS를 충족하기 위해 더 작은 DNN 모델을 선택한다. 또한, 가용 메모리가 줄어들면 DNN 모델 세트의 크기를 줄여 목표 FPS를 달성할 수 있도록 한다. 온도와 메모리 제약에 대해 여유로워지면 H&M 알고리즘은 추론 정확도를 높이기 위해 높은 CPU/GPU 클럭 주파수와 대규모 DNN 모델을 선택하는 경향이 있음을 알 수 있다.

Ⅲ. 결론

본 논문에서는 엣지 지원 온디바이스 추론 시스템에서 모바일 장치의 열과 메모리 제약이 미치는 영향에 대해 자세히 조사했다. 또한, 사전 실험에서 얻은 이해를 바탕으로 임계값 기반 H&M 알고리즘을 제안하고, 실제 스마트폰에서 이미지 분류 어플리케이션을 구동하여 가용 메모리가 동적으로 변하는 상황에서도 안정적인 FPS, 높은 정확도, CPU, GPU, 5G 모뎀의 온도를 열 임계값 이하로 유지할 수 있음을 실험을 통해 입증했다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] Seyeon Kim, Kyungmin Bin, Sangtae Ha, Kyunghan Lee, and Song Chong. 2021. zTT: Learning-based DVFS with Zero Thermal Throttling for Mobile Devices. In Proceedings of the 19th ACM Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '21). virtual, 41-53.