임베디드 시스템(Raspberry PI 5, Galaxy Note 10+) 환경에서의 MobileViT 탑재 및 성능 검증에 관한 연구

김은호, 임채우, 이윤동, 유동관

동서울 대학교

ho1582@naver.com, dlacodn456@naver.com, solus1005@naver.com, dgyoo@du.ac.kr

A Study on the Porting and Performance Validation of MobileViT in Embedded Systems (Raspberry PI 5, Galaxy Note 10+) Environment

Kim Eun Ho, Im Chae Woo, Lee Yun Dong, Yoo Dong Kwan

Dong Seoul College

요 약

본 논문은 MobileViT 모델을 임베디드 시스템(Raspberry PI 5, Galaxy Note 10+)에 탑재한 후 성능 검증을 진행하였다. 첫 번째 실험은 Raspberry PI 5 환경에서 MobileViT 모델과 MobileNetv2 모델 간의 성능 비교를 수행하였다. 성능 비교에 사용한 데이터셋은 Imagenet-1k이다. 수행한 성능 비교 결과는 MobileViT 모델이 16% 더 높은 성능을 보였다. 두 번째 실험은 Galaxy Note 10+ 환경에서 MobileViT 모델과 MobileNetv2 모델에 대해서 주어진 이미지에 대한 추론 시간을 비교 평가하였다. 추론 시간을 비교한 결과 MobileViT 모델은 0.14초가 나왔고, MobileNetv2 모델은 0.05초가 나왔다. 본 논문의 실험 결과를 통해 MobileViT 모델이 성능 향상 값은 크고 추론 속도의 차이는 미미하므로 MobileNetv2 모델보다 온 디바이스 AI(On-Device AI)로 사용하는 것이 더 바람직함을 확인하였다.

Ⅰ. 서 론

기존에는 CV분야에서 CNN 계열 모델들이 많이 사용되었다. 또한, 트랜스포머(Transformer)가 등장하고 난 후 CV분야에 맞게 개량한 ViT(Vision Transformer) 모델이 등장하였고, ViT 모델을 계열로 여러 파생형 모델들이 등장하였다. 하지만 ViT 계열 모델들은 CNN 계열 모델 대비 성능은 우수하나, 파라미터의 수가 많다는 단점을 가지고 있다. 이유는 CNN에 존재하는 지역 귀납적 편향(Local Inductive Bias)이 ViT 계열 모델에는 부족하기 때문에 동일한 파라미터 수를 가지는 CNN 계열 모델보다 성능이 낮았고, 상대적으로 부족한 모델의 성능을 올리기 위해서 파라미터 수를 늘리는 방식을 선택한 것이다. 따라서 제약된 환경에서는 ViT 계열 모델을 사용하기 어려웠다[1]. 본 논문에서 소개하는 MobileViT는 ViT를 사용하지만, CNN의 특징인 지역 귀납적 편향을 많이 가지는 모델로 적은 양의 파라미터로 동작하므로 온 디바이스 AI로서 사용이 가능한 모델이다.

Ⅱ. 본론

1. MobileViT

MobileViT 모델은 널리 사용되는 CV모델인 CNN 모델과 ViT 모델의 장점을 합쳐 만든 모델이다. ViT 모델의 장점으로는 입력-적응형 가중치(Input-Adaptive Weighting)와 전역 처리(Global Processing)가 있다. 하지만, 높은 계산 비용과 많은 양의 데이터가 필요하다. CNN 모델의 장점은 상대적으로 많은 수의 지역 귀납적 편향과 데이터 증대에 둔감하다는 점이다. 그러나, 장거리 의존성 문제가 발생할 수 있다. CNN 모델과 ViT 모델의 장점을 합쳐 만든 MobileViT 모델은 크기가 작고 처리속도가 빠르며 ViT 계열 모델의 부족한 지역 귀납적 편향을 많이 가지고 있다[2].

2. MobileViT 블록

MobileViT 구조는 MobileNetv2 구조에서 MobileViT 블록이 3개 추가된 형태를 가지고 있다. MobileViT 블록은 지역 표현(Local Representation), 전역 표현(Global Representation), 퓨전(Fusion)으로 구성되어 있다. 지역 표현은 N x N 컨벌루션과 Point Wise 컨벌루션 연산을 수행한다. 지역 표현에서 컨벌루션 연산을 통해 지역 정보(Local Information)를 얻는다. 전역 표현은 Unfold와 Transformer, Fold로 구성된다. 전역 표현에서 지역 정보를 가진 입력을 패치로 나누어 Transformer 연산을 통해 전역 정보(Global Information)를 얻는다. 전역 표현의 출력은 지역 정보와 전역 정보를 동시에 가지고 있다. 또한, Unfold를 수행할 때 패치를 순서대로 펼치기 때문에 위치 정보에 대한 손실이 없어 순차 임베딩(Positional Embedding)이 불필요하다. 마지막 퓨전에서는 Point Wise 컨벌루션을 수행한 후, 초기 입력과 합친 다음, 입력 형태와 동일한 형태로 만들어주기 위해 N x N 컨벌루션을 수행한다. MobileViT 블록의 구조도는 다음의 그림 1과 같다.

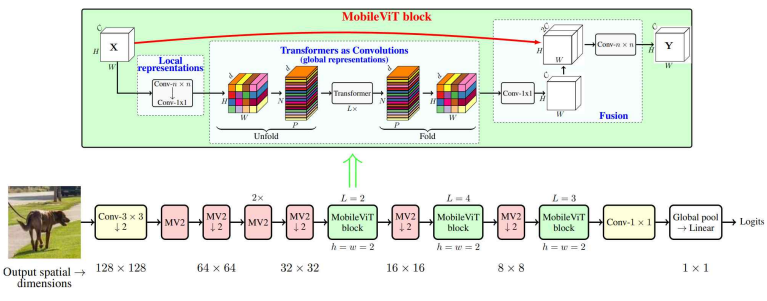


그림 1. MobileViT 블록의 구조도

4. 멀티 스케일 샘플러(Multi-Scale Sampler)

기존 ViT 계열 모델들은 다중 스케일 표현을 학습하기 위해서 미세 조정이 필수적이었다. 그 이유는 ViT 모델은 순차 임베딩을 사용하는데, 이미지의 크기가 달라지면 순차 임베딩도 달라지기 때문이다. 따라서, ViT 계열 모델들은 스탠다드 샘플러 방식으로 학습을 진행하였다. 하지만 MobileViT 모델은 순차 임베딩이 없기 때문에 멀티 스케일 샘플러 방식으로 학습할 수 있었다. 또한, 기존의 멀티 스케일 샘플러에 있었던 배치 사이즈에 대한 문제를 해결하여 적용하였다. 기존 방식은 제일 큰 이미지를 기준으로 배치 사이즈가 설정되어 작은 이미지를 처리할 때 GPU를 거의 사용하지 않았는데, 이를 현재 이미지의 크기를 기준으로 배치 사이즈를 유동적으로 변경하여 해결하였다. 스탠다드 샘플러와 멀티 스케일 샘플러의 구조는 다음의 그림 2와 같다.

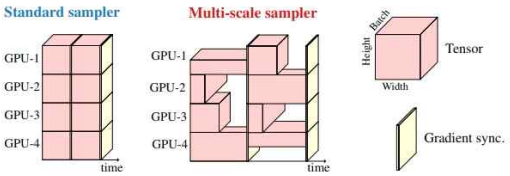


그림 2. 스탠다드 샘플러와 멀티 스케일 샘플러의 구조

Ⅲ. 실험

성능 비교 실험은 임베디드 시스템(Raspberry PI 5)에 2개의 라이브러리를 설치하고 Imagenet-1k 데이터셋을 사용하여 분류 태스크에 대하여 MobileViT 모델과 MobileNetv2 모델의 성능 평가를 진행하였다. 성능 측정 지표로는 Top-1, Top-5, Inference Time을 사용하였다. 임베디드 시스템(Raspberry PI 5, Galaxy Note 10+)의 구현 환경은 다음의 <표 1>, <표 2>와 같다.

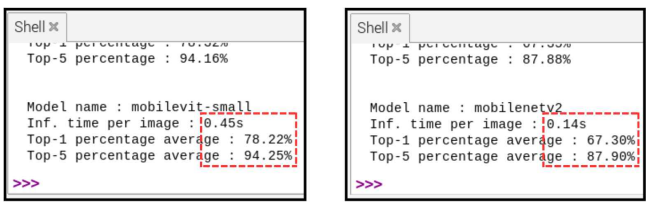
<표 1> 임베디드 시스템(Raspberry PI 5)의 구현 환경

항목		내용
H/W	CPU	BCM2712 (2.4GHz)
	GPU	VideoCore VII (800MHz)
	MEMORY	SDRAM 4267
	SD card	micro 카드 슬롯, SDR104 고속 모드 지원
S/W	O/S	Debian GNU/Linux 12
	Library	Pytorch=2.2.2, timm=0.9.16

<표 2> 임베디드 시스템(Galaxy Note 10+)의 구현 환경

항목		내용
H/W	CPU	Exynos 9825 (2.7GHz)
	GPU	Mali-G76 MP12 (754MHz)
	MEMORY	SDRAM
S/W	O/S	Android 12
	Program	Android Studio IDE (2023.2.1) SDK 24 이상 (Android 7.0 이상)

첫 번째 실험의 각 모델별 임베디드 시스템(Raspberry PI 5)에서 수행한 쉘의 동작화면은 다음의 그림 3과 같다.



(a)MobileViT (b)MobileNetv2

그림 3. Raspberry PI 5의 쉘의 동작화면

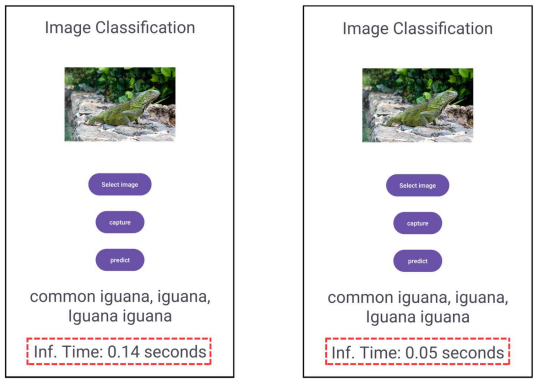
임베디드 시스템(Raspberry PI 5)에서 MobileViT 모델과 MobileNetv2 모델의 성능 평가를 진행하였다. 성능 및 추론 시간 실험결과는 다음의 <표 3>과 같다.

<표 3> 성능 및 추론 시간 측정 결과

Score / Model	Top-1	Top-5	Inf. time (second)
MobileViT	78.22	94.25	0.45
MobileNetv2	67.30	87.90	0.14

실험 결과, 임베디드 시스템(Raspberry PI 5)에서 MobileViT 모델은 0.45초, MobileNetv2 모델은 0.14초의 추론 시간을 기록하였으나, 성능 측면에서 MobileViT 모델은 Top-1, Top-5에서 78.22%, 94.25%를 기록하였고, MobileNetv2 모델은 Top-1, Top-5에서 67.3%, 87.90%를 기록하였다. MobileViT 모델이 MobileNetv2 모델 대비 Top-1, Top-5가 16%, 7% 더 높은 것을 확인할 수 있다.

두 번째 실험은 임베디드 시스템(Galaxy Note 10+)에서 MobileViT 모델과 MobileNetv2 모델을 구현하고 추론 시간을 비교하였다. 결과는 다음의 그림 4와 같다.



(a)MobileViT (b)MobileNetv2

그림 4. Galaxy Note 10+의 동작화면

임베디드 시스템(Galaxy Note 10+)에서 MobileViT 모델과 MobileNetv2 모델의 성능 평가를 진행하였다. 추론 시간 실험결과는 다음의 <표 4>와 같다.

<표 4> 추론 시간 측정 결과

Score / Model	Inf. time (second)
MobileViT	0.14
MobileNetv2	0.05

실험 결과, 임베디드 시스템(Galaxy Note 10+)에서 MobileViT 모델의 추론 시간은 0.14초, MobileNetv2 모델의 추론 시간은 0.05초를 기록하였다.

Ⅳ. 결론

본 논문에서 사용한 MobileViT 모델은 ViT 계열 모델임에도 불구하고 CNN 계열 모델들과 마찬가지로 지역 귀납적 편향을 많이 가지고 있으며, 크기가 작고 추론 속도 차이가 임베디드 시스템(Galaxy Note 10+)에서 0.1초 이내로 미미하기 때문에 임베디드 시스템(Raspberry PI 5, Galaxy Note 10+)에 온 디바이스 AI로 탑재하여 사용할 수 있는 모델임을 증명하였다. 따라서, 향후 온 디바이스 AI 분야 중에서도 로봇 등의 분야에서 활용될 수 있을 것으로 예상된다.

참 고 문 헌

[1] Peter W. Battaglia et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, arXiv:2010.11929

[2] Sachin Mehta et al., “MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer”, arXiv:2110.02178