모바일 딥러닝을 위한 신경망 성능 평가에 관한 연구

신익희*[1,2], 박준용[2], 문용혁[2], 이용주[1,2]
[1] 과학기술연합대학원대학교 한국전자통신연구원스쿨
[2] 한국전자통신연구원 SW 콘텐츠연구소 지능정보연구본부
ihshin227@etri.re.kr*, junyong.park@etri.re.kr, yhmoon@etri.re.kr, yongju@etri.re.kr

A Performance Study on Lightweight Neural Network for Mobile Deep Learning

Ik Hee Shin[1,2], Junyong Park[2], Yong Hyuk Moon[2], Yong-Ju Lee[1,2]
[1]University of Science and Technology
[2] Electronics and Telecommunications Research Institute

요 약

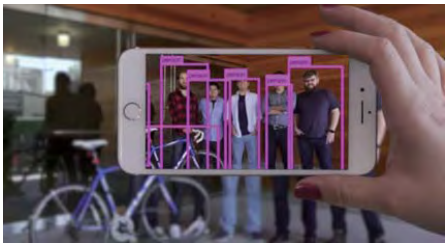
모바일 환경에서 다양한 AI 관련 응용을 수행하기 위해, 정확도에 기반한 크고 깊은 신경망 이외에, 정확도를 비교적 유지하면서 좀더 효율적인 신경망 구조에 대한 다양한 연구가 진행중이다. 본 논문에서는 모바일 딥러닝을 위한 다양한 임베디드 장치 및 모바일 폰에서의 성능 평가를 통해 경량 신경망의 비교 분석에 대한 연구를 담고 있다.

1. 서론

최근 들어, 이미지, 오디오, 텍스트 등 다양한 연구 분야에서 신경망 기반의 다양한 연구가 진행중이다. 일반적인 신경망 연구는 크고 깊은 신경망을 설계하여 학습 및 추론을 수행하는데 주 연구가 이루어지고 있었다. 하지만, 모바일 디바이스나 IoT 환경과 같이 컴퓨팅 능력이 제한적인 기기에서의 신경망을 적용하는 연구 또한 활발히 진행되고 있다. 이러한 모바일 딥러닝은 신경망의 연산을 비교적 적게 하여, 에너지 소모가 적게 하는데 그 목적이 있으며, 본 연구에서는 임베디드 장치와 모바일 폰에서의 기존 신경망 연구를 비교 분석하며, 앞으로 경량 신경망 연구에 대한 다양한 알고리즘 경량화 기법에 대한 연구를 설계한다.

2. 모바일 딥러닝

모바일 딥러닝은 기본적으로 추론이 모바일 기기에서 수행되는 특징을 가지고 있으며, 클라우드나 서버 기반의 네트워크의 트래픽이 필요하지 않으며, 민감한 개인 정보 보호 및 지연시간 감소 등의 다양한 이점을 가지고 있다. 하지만 모바일 환경에서의 추론은 에너지 소모와 신경망의 학습 결과를 사전에 기기에 저장 하여 사용해야 한다는 단점을 가지고 있으며, 동일한 또는 유사한 정확도를 갖기 위해 필요한 깊고 큰 신경망을 사용하기가 쉽지 않은 상황이다. 이를 위해 모델 압축과 같은 다양한 신경망 구조 변경 기술과 양자화(Quantization)와 같은 표현력을 줄이는 연구가 진행중이다[1]. 신경망 구조 변경은 기존의 합성곱의 필터를 보다 작게 설계하고, 표현력을 유지하기 위한 다양한 연구가 진행중이다. 레즈넷(ResNet), 덴스넷(DenseNet) 과 같은 기존의 평형망 형태가 아닌 블록을 설계하거나, 모바일넷(MobileNet)[2] , 셔플넷(ShuffleNet), 스퀴즈넷(SqueezeNet)과 같은 합성곱의 보다 효율적인 구조가 제안되었다. 양자화 연구는 기존의 32 비트로 표현되는 가중치의 부동소수점수를 8 비트 형태로 줄여, 모델 크기와 연산을 대폭적으로 줄이기 위한 연구이다. 현재 텐서플로와 같은 딥러닝 툴킷에서 제공되고 있는 양자화가 대표적인 예이다[3].



(그림 1) 스마트폰에서의 객체 인식 예

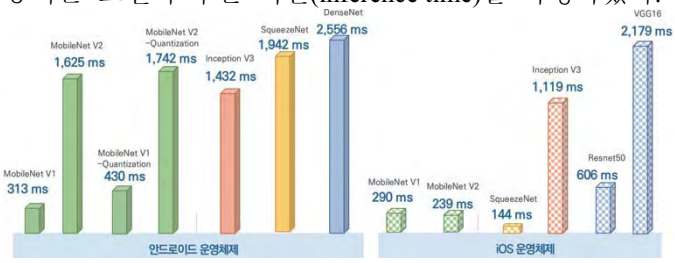
본 연구에서는 기존에 제안된 다양한 모바일에 적합한 신경망의 성능 평가를 통해, 그 특징을 알아보고 실제 적용 가능한 프레임워크를 제안한다. 제안하는 프레임워크는 크게 GPU 를 가진 클라우드에서 기존 신경망을 압축하여, 모바일 및 임베디드 장치에

적용 가능한 간소화 신경망을 생성하고, 모바일 추론을 위해 배포 및 계속적인 피드백을 통해 선순환되는 구조를 제안한다.

3. 모바일/임베디드 기기에서의 신경망 성능평가

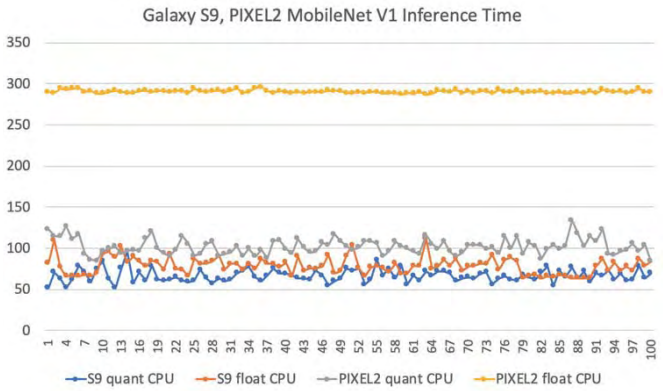
현재, 모바일 기기에서의 추론 가능한 딥러닝 모델은 구글과 애플에서 기기에 적합한 사전 훈련된 모델을 제공하고 있다. 비교적 작은 크기의 신경망으로 주로 객체 인식(object detection)이 가능한 카메라 앱 형태로 사용 가능하며, VGG16 부터 최근의 모델 들(SqueezeNet, DenseNet)등으로 다양하다. 또한 기본적인 모델 압축 기법인 양자화(Quantization)을 수행한 모델과 일반 부동소수점수를 사용한 모델 등이 복수로 제공되는 경우도 있다.

성능 평가는 크게 모바일 기기의 운영체제 별로 제공되는 모델의 추론 시간(inference time)을 측정하였다.



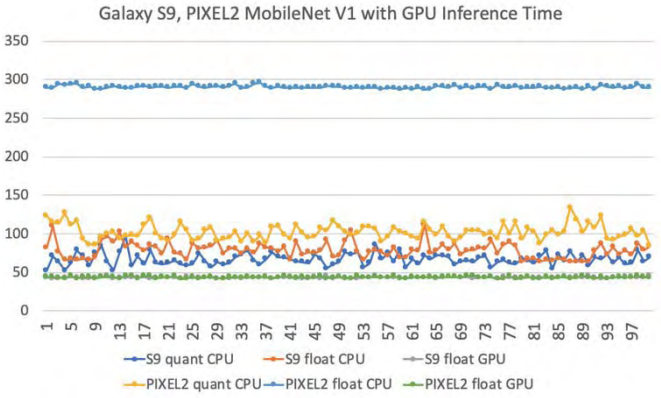
(그림 2) 안드로이드/iOS 기기에서의 모바일 딥러닝 모델 성능평가

(그림 2)에서와 같이, 안드로이드 운영체제에서는 모바일넷(MobileNet)이 가장 짧은 추론 시간을 가지며, 최근의 신경망인 스퀴즈넷(SqueezeNet)과 덴스넷(DenseNet)의 경우 비교적 긴 추론 시간을 보인다. 주목할 점은 동일한 모바일넷의 경우에도 기존 부동소수점수를 사용한 모델이 양자화를 거친 모델 보다 약간의 성능상의 이점을 보인다는 점이다. iOS 운영체제에서는 모바일넷과 스퀴즈넷이 가장 우수한 성능을 보이며, 비교적 이전 모델인 VGG16 과 인셉션이 느린 성능을 보였다. 동일한 연산을 수행하는 동일 모델에서도 안드로이드 운영체제와 iOS 의 운영체제에서 서로 상이한 결과를 나타내는 경우가 있으며 스퀴즈넷이 대표적이다.



(그림 3) 안드로이드 기기에서의 부동소수점연산(32비트)과 양자화를 거친 연산(8비트)에서의 성능 비교

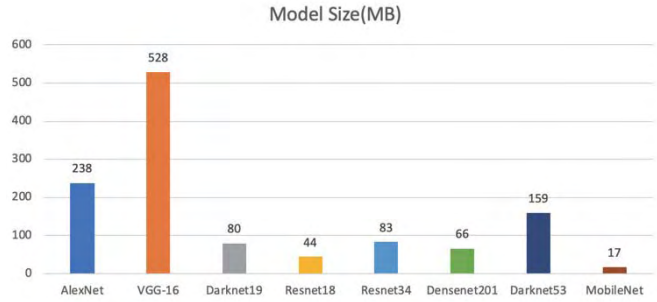
다음으로, (그림 3)과 같이 안드로이드 기기에서 대표적인 모바일 추론을 수행하는 모바일넷에서의 추론 시간을 비교하였다. 삼성 갤럭시 S9 모델과 구글 픽셀 2 모델에서의 추론 시간을 비교한 결과, 구글 픽셀 2 의 부동소수점 연산(32 비트)의 경우, 다른 실험 결과에 비해 비교적 긴 추론 시간을 나타낸다.



(그림 4) 모바일 GPU를 통한 안드로이드 환경에서의 추론 시간 성능평가

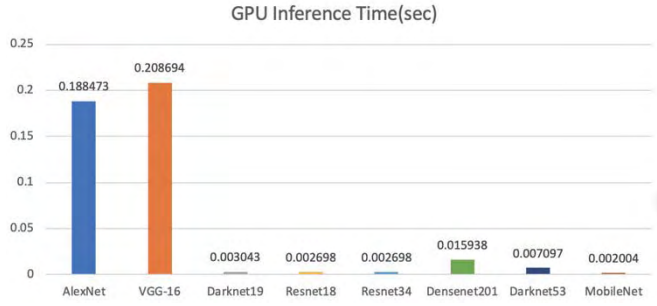
(그림 4)는 모바일 GPU 를 통한 추론 시간을 비교하였다. 픽셀 2 의 부동소수점 연산을 수행하는 경우를 제외하고는 모든 경우에서 비교적 균일한 추론 시간을 보인다. 모바일 GPU 를 사용하는 경우 양자화를 통한 추론 시간 보다 기존 방법의 추론 시간이 약간의 성능이 우수함을 볼 수 있다.

다음으로, 임베디드 장치인 NVIDIA Jetson TX2 에서 기존 신경망들의 성능을 측정하였다. (그림 5)는 기존 신경망 들의 모델 크기를 비교하였다.



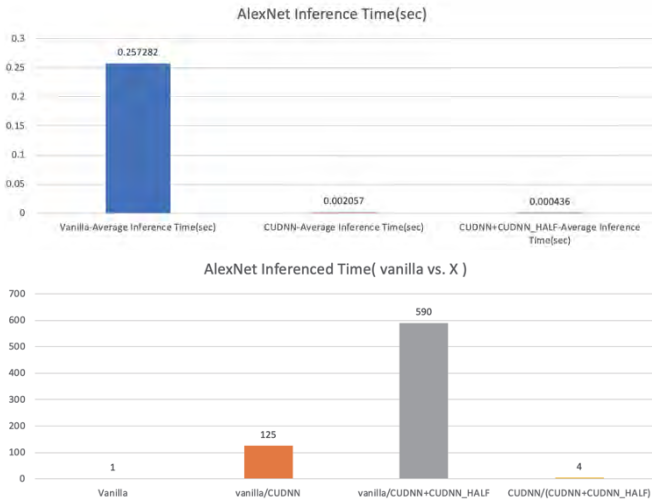
(그림 5) 기존 신경망 모델 크기 비교

모델 추론 시간의 측정은 이미지넷의 데이터를 통해 사전 훈련된 모델을 통해 5 만번의 추론 시간 측정의 평균값이며, 알렉스넷과 VGG16 를 제외하고 비교적 짧은 추론 시간을 가진다.



(그림 6) NVIDIA Jetson TX2에서의 신경망 모델의 모델 추론 시간 비교

기존의 모바일 추론 시간 이외에 현재 제공되는 CUDA 를 활용한 추론 시간도 비교하였다. 일반적인 추론 시간(Vanilla)과 NVIDIA 에서 제공되는 딥러닝 CUDA 패키지인 CUDNN 을 포함한 경우(FP32:floating point 32-bit, single-precision)와 16 비트 연산을 가능하게 하는 CUDNN_HALF 기능을 포함한 경우(FP16:floating point 16-bit, half-precision)의 3 가지 경우를 측정하였다. CUDNN+CUDNN_HALF 의 경우가 월등한 성능상의 이점을 보였다. 이러한 성능을 상대적으로 비교하면 Vanilla 의 경우를 1 로 보면 CUDNN 은 최대 125 배, CUDNN+CUDNN_HALF 은 590 배의 성능이 우수하며, CUDNN 과 CUDNN_HALF 는 4 배의 성능이 우수함을 보였다.

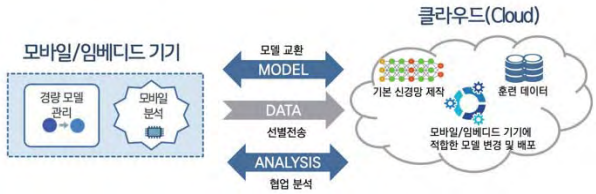


(그림 7) TITAN V에서의 알렉스넷 추론 시간 상대 비교

성능 측정 결과를 요약하면, 모바일 기기에서의 성능은 기존의 알렉스넷이나 VGG16 과 같은 전통적인 신경망 보다는 합성곱 필터 변경을 통해 단순화하는 기법인 모바일넷, 스퀴즈넷과 같은 형태가 우수하며, 모델 압축 기법은 양자화 기법의 비교적 성능에 큰 영향을 미치지 못하고 있다. 다만 모바일 GPU 를 활용하여, 기존 신경망의 가속화가 가능하게 된다. CUDA 가 지원되는 임베디드 장치의 경우는 CUDA 를 활용한 성능이 탁월함을 보인다. 라즈베리파이 3 와 같은 임베디드 보드에서 점점 모바일 추론이 가능한 CUDA 코어를 탑재한 NVIDIA Jetson[4], Myriad X VPU 를 사용한 Intel Movidius[5], Edge TPU 를 탑재할 예정인 Google Coral[6]이 대표적인 예로 볼 수 있다.

4. 모바일 딥러닝을 위한 클라우드 훈련 및 모바일 추론 방법 제안

현재까지, 모바일 딥러닝을 경량 신경망 모델인 모바일넷 등을 통해, 간단한 객체 인식 수준에 그치고 있지만, 앞으로 다양한 응용 분야에서 모바일 딥러닝이 가능할 것으로 보인다. 하지만, 모바일 기기에서의 훈련은 비교적 쉽지 않은 상황이므로, 클라우드에 기반한 딥러닝 모델 훈련 및 다양한 알고리즘을 경량화를 통해 모바일/임베디드 기기에서 추론이 가능한 형태로 진행될 것으로 예상된다. 모바일 추론에서는 클라우드에서 경량화된 딥러닝 모델을 받아서, 단순한 추론을 수행하는 형태이지만, 모바일 GPU 또는 임베디드 기기의 다양한 코어를 탑재한 기기에 출시됨에 따라 좀더 복잡한 추론이 가능 할 것으로 보인다. (그림 8)과 같이, 모바일 추론을 효율적으로 관리/수행하기 위해서는 클라우드 기반의 관리가 이루어져야 하며, 클라우드에서는 훈련 데이터 관리, 기본 신경망 제작 및 모바일/임베디드 기기에 적합한 모델 선정 및 변경/배포 기능을 내재 되어야 한다. 모바일/임베디드 기기에서는 모바일 분석 및 경량 모델에 대한 관리가 필요하다.



(그림 8) 모바일 딥러닝을 위한 클라우드와 연계 방법

5. 결론

본 논문에서는 현재 모바일 환경에서 수행 가능한 딥러닝 모델들에 대해 안드로이드/iOS 에서의 추론 시간 성능 평가, 양자화와 같은 모델 압축 기법을 적용한 추론 시간 비교, 임베디드 환경에서의 신경망 추론 시간, 다양한 CUDA 환경에서의 성능 향상을 비교 분석하였다. 이를 통해 모바일/임베디드 환경에서의 추론은 모델 구조/합성곱 필터 변경을 효율적으로 수행한 모델 형태로 발전하고 있으며, 모바일 GPU, 임베디드 기기의 다양한 가속화 유닛을 통해 점점 보다 복잡한 추론이 가능할 것으로 예상한다. 향후에는, 엣지 컴퓨팅과 같은 다양한 엣지 디바이스에 적용 가능한 효율적인 신경망에 대한 연구를 진행중에 있으며, 클라우드와 엣지를 통한 협업 분석에 대한 추후 연구도 진행할 예정이다.

* 이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행한 연구임(No. 2018-0-00278, 부하분산과 능동적 적시 대응을 위한 빅데이터 엣지 분석 기술 개발)

참고문헌

[1] Cheng, Yu, et al. "A survey of model compression and acceleration for deep neural networks." arXiv preprint arXiv:1710.09282 (2017).
[2] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H. “Mobilenets: Efficient convolutional neural networks for mobile vision applications.” arXiv:1704.04861 (2017)
[3] https://www.tensorflow.org/lite/performance/model_optimization
[4] https://www.nvidia.com/ko-kr/autonomous-machines/embedded-systems-dev-kits-modules/
[5] https://www.movidius.com/
[6] https://coral.withgoogle.com/products/