사물인터넷 기기 환경에서의 영상인식 기술 분석 연구

배은지, 이성진

동서울대학교 전자공학과

ejbae25@du.ac.kr, sungjinlee@du.ac.kr

Research on Performance Analysis of Image Classification Technologies in environments of IoT Devices

Bae Eunjee, Lee Sungjin

Dong Seoul University, Department of Electric Engineering

요 약

본 논문에서는 사물인터넷 기기 환경의 보편화로 이런 기기들에서 얻어지는 다양한 영상 데이터들을 활용한 영상인식 기술 서비스를 위한 성능 지표들을 제시하였다. 특히 전통적인 사물인터넷 기기 인 Rasberry Pi 3 (Model B+), 4 (Model B) 와 최근 Edge AI 기기로 각광받고 있는 NVIDIA Jetson Nano, Xavier AGX를 대상으로 경량 영상인식 기술들을 탑재하여 성능을 측정하였다. 성능 평가 metric으로 인식속도와 전력소모를 보였으며, 그 결과 Tensorflow에서 활용가능한 GPU를 탑재한 NVIDIA 계열이 그렇지 못한 Rasberry Pi 계열보다 인식속도 측면과 전력소모 효율 측면 모두에서 약 100여 배 정도의 이득을 관찰할 수 있었다.

Ⅰ. 서 론

4차 산업혁명 시대의 중심 축으로 작용하는 사물인터넷 기기 환경은 이제는 널리 보편화 되어 가면서 이런 다양한 기기들로부터 얻어지는 데이터들을 활용한 다양한 인공지능 서비스들이 새로운 부가가치 기술로서 여겨지고 있다. 특히 영상데이터를 기반으로 한 인공지능 영상인식 기술은 이런 사물인터넷 기기에서 On Device AI 라는 이름으로 해당 인공지능 연산들을 기기 자체에서 직접 수행하는 단계에 까지 이르렀고 이를 위한 대기업들의 다양한 노력들이 행해지고 있는 중이다.

하지만, 이런 영상인식 기술들은 그 막중한 연산량으로 인해 사물인터넷 기기에서 직접 수행하기는 어려워서 서버를 통한 클라우드 기술을 통해 풀어나가기도 하지만 보안 이슈, 통신 지연 이슈, 베터리 이슈 등으로 인해 오히려 On Device AI를 기술 경량 최적화로 해결하려는 연구가 주목받는 계기가 되었다. 이에 Google은 MobileNet [1] 이란 이름으로 Depthwise Separable Convolution 이라는 경량화된 영상인식 기술을 공개하였으며 이어서 개선된 MobileNet V2, V3 [2,3] 를 공개하며 Convolution Module 최적화에 대한 연구를 공개하였다. 또한 EfficientNet [4], MNasNet [5] 등이 공개되면서 AutoML을 통한 Convolution Module과 해당 Module들의 네트워크 구조 탐색의 최적화를 이룩하기도 하였고, FaceBook과 Berkely 대학은 ShiftNet [6] 을 공개하며 Convolution 연산의 부담을 최소화 한 연구를 공개하였다. 이밖에 SqueezeNet [7]은 영상 특성 추출기의 최적 설계를 위해 Filter Concatenation을 통한 Fire Module을 공개하였고, ShuffleNet [8]은 Filter들간의 Group Shuffle 기술을 통해 연산 Filter 수를 줄이고자 하는 연구를 공개하였다.

본 논문에서는 하드웨어적으로 제약이 있는 사물인터넷 환경에서 On Device AI 기술 실현을 위한 다양한 경량 영상인식 기술들을 분석해 보고 나아가야 할 방향들을 고찰해 보았다. 특히, 최근 Edge AI 로 주목받고 있는 NVIDIA 의 Jetson Nano, Xavier AGX 그리고 오랫동안 영상 用 사물인터넷 기기로 사용되던 Rasberry Pi 3 Model B+, 4 Model B를 대상으로 여러 경량 영상인식 기술들을 탑재하여 그 성능을 측정해 보았다.

2. 동작 하드웨어 분석

실험을 위해 사용된 사물인터넷 기기는 GPU 유무를 기준으로 GPU를 장착한 NVIDIA 의 Jetson Nano, Xavier AGX와 GPU가 없는 Low End 기기인 Rasberry Pi 3 Model B+, 4 Model B를 대상으로 하였다. 실험결과에 결정적인 영향을 미칠 수 있는 각 기기의 Hardware 사양은 표 1과 같다.

	CPU	GPU	RAM	가격	소비전력
Resberry Pi 3 Model B+	Broadcom BCM2837B0, Quad core Cortex-A53 64-bit SoC @ 1.4GHz	Broadcom VideoCore IV 400MHz (TensorFlow-GPU 사용불가)	1GB	$35	2.9 W, 6.4 W
Rasberry Pi 4 Model B	Broadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz	Broadcom VideoCore VI 500MHz (TensorFlow-GPU 사용불가)	4GB	$55	3.4 W, 7.6 W
NVIDIA Jetson Nano	Quad-core ARM A57 @ 1.43 GHz	128-core Maxwell	4GB	$99	5W, 10W
NVIDIA AGX Xavier	8-core ARM v8.2 64-bit CPU, 8MB L2 + 4MB L3	512-core Volta GPU with Tensor Cores	32GB	$999	10W,15W , 30W

표 1. 실험 기기의 Hardware 사양

3. 데이터 셋 및 성능 측정 Metric 선정
훈련 用 데이터 셋은 ImageNet을 대상으로 하였다. 성능평가 用 데이터 셋은 ImageNet의 validation의 이미지들을 사용하였다. 평가 Metric은 동작 속도로 ms (mile second) 와 소모 전력을 사용하였다.

4. 모델 변환
일반적인 딥러닝 개발 툴인 TensorFlow를 기반으로 해당 모델들을 각 모바일 기기 환경에 최적화된 포맷인 TF-Lite로 변환하여 탑재하였다. 해당 과정은 tensorflow 공식 홈페이지 [9]를 참고하여 수행하였다.

Ⅲ. 성능 측정

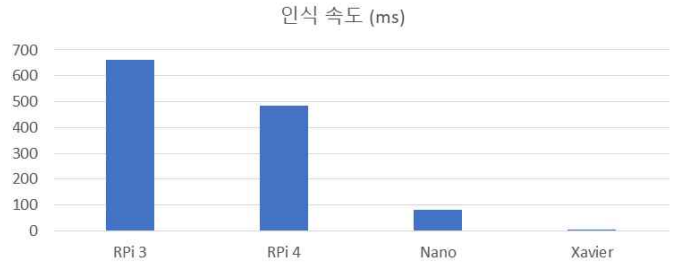

그림 1. 모델별 인식 지연 시간 비교

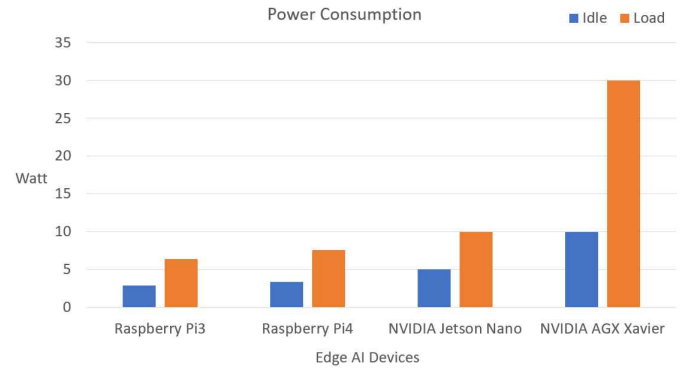

그림 2. 모바일 기기별 소비 전력 비교

그림 1은 Rasberry Pi 3, 4 기기와 NVIDIA Jetson Nano, Xavier AGX 기기에 MobileNet v2 버전모델을 탑재하였을 때 의 인식 지연 시간을 비교한 그래프이고 그림 2는 해당 기기들에 동일한 영상인식 모델을 동작시켰을 때의 소비전력을 나타낸 그래프 이다.
일단 첫 번째 발견으로 GPU의 사용여부에 따라 동작 시간이 많이 줄어든다는 것이다. 그림 1을 보면 Pi 3모델과 Xavier 모델 간에는 약 600배, Pi 4모델과 Nano 모델간에는 약 5배의 동작 시간 차이를 보인다는 것을 알 수 있다.
두 번째 발견은 Rasberry Pi 기기들의 전력 소모가 NVIDIA 계열의 전력 소모보다 크게 작지는 않다는 것이다. 특히 영상인식 모델이 실행될 때는 Full Load Power Profile을 사용한다는 가정 하에 그림 2에서 보듯이 Pi3 와 Nano 간에는 1.5배, Xavier 간에는 4.7배의 차이를 보이며, Pi 4를 기준으로는 각각 1.3배, 4배 의 차이를 보인다는 것을 알 수 있다. 하지만 그림 1의 동작 시간을 고려하면 전력 에너지 효율 측면에서 보면 NVIDIA 계열의 기기들이 약 4~140 배 정도 좋다는 것을 알 수 있다.

Ⅳ. 결론

본 연구는 대표적인 영상분류 모델들과 그 연산량 감소를 위해 개선된 경량화 구조 모델들, 그리고 기기들의 하드웨어 자원이 영상인식 성능, 전력 효율들에 미치는 영향들을 분석해 보았다. 그 결과 인식 속도에 결정적으로 영향을 미치는 요인으로 GPU의 유무와 그 성능이며 이는 전력효율에도 유사한 영향을 미칠 수 있음을 알 수 있었다.

ACKNOWLEDGMENT

이 논문은 2020년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기본연구사업임(No. NRF-2019R1F1A1062878)

참 고 문 헌

[1] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. CoRR,abs/1704.04861, 2017.

[2] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks, detection and segmentation. CoRR, abs/1801.04381, 2018.

[3] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019

[4] Mingxing Tan, and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning, PMLR 97:6105-6114, 2019.

[5] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile. CoRR, abs/1807.11626, 2018.

[6] Bichen Wu, Alvin Wan, Xiangyu Yue, Peter H. Jin, Sicheng Zhao, Noah Golmant, Amir Gholaminejad, Joseph Gonzalez, and Kurt Keutzer. Shift: A zero flop, zero parameter alternative to spatial convolutions. CoRR, abs/1711.08141, 2017.

[7] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. CoRR, abs/1602.07360, 2016.

[8] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shuf flenet: An extremely efficient convolutional neural network for mobile devices. CoRR, abs/1707.01083, 2017.

[9] https://www.tensorflow.org/lite/guide?hl=ko