CNN 의 파라미터와 정확도간 상호 강인성 연구 및 파라미터 비트 연산 자동화 프레임워크 개발

이동인 [1], 김정헌 [2], 임승호 [3]
[1, 3] 한국외국어대학교 컴퓨터공학부
[2] 한국외국어대학교 정보통신공학과

dongeen1@gmail.com, kjh990705@naver.com, slim@hufs.ac.kr

Study the mutual robustness between parameter and accuracy in CNNs and developed an Automated Parameter Bit Operation Framework

Dong-In Lee[1], Jung-Heon Kim[2], Seung-Ho Lim[3]
[1, 3] Division of Computer Engineering, Hankuk University of Foreign Studies
[2]Dept. of Information Communication Engineering, Hankuk University of Foreign Studies

요 약

최근 CNN 이 다양한 산업에 확산되고 있으며, IoT 기기 및 엣지 컴퓨팅에 적합한 경량 모델에 대한 연구가 급증하고 있다. 본 논문에서는 CNN 모델의 파라미터 비트 연산을 위한 자동화 프레임워크를 제안하고, 파라미터 비트와 모델 정확도 사이의 관계를 실험 및 연구한다. 제안된 프레임워크는 하위 *n*- bit 를 0 으로 설정하여 정보 손실 발생시킴으로써 ImageNet 데이터셋으로 사전 학습된 CNN 모델의 파라미터와 정확도의 강인성을 비트 단위로 체계적으로 실험할 수 있다. 우리는 비트 연산을 수행한 파라미터로 InceptionV3, InceptionResnetV2, ResNet50, Xception, DenseNet121, MobileNetV1, MobileNetV2 모델의 정확도를 평가한다. 실험 결과는 성능이 낮은 모델일수록 파라미터와 정확도 간의 강인성이 높아 성능이 좋은 모델보다 정확도를 유지하는 비트 수가 적다는 것을 보여준다.

1. 서론

최근 CNN 이 다양한 산업에 확산되고 있으며, IoT 기기 및 엣지 컴퓨팅에 적합한 경량 모델에 대한 연구가 급증하고 있다 [1].

여러 선행 연구에서는 모델의 경량화를 위해 파라미터를 압축하는 방식의 양자화 기법을 적용하였다 [2]. 본 논문에서는 압축이 아닌 하위 *n*-bit 를 0 으로 만드는 비트 연산을 통해 상대적으로 영향력이 작은 비트에 대한 정보 손실을 발생시키며 모델의 정확도를 측정하였다. 압축하는 방식이 아닌 하위 *n*-bit 를 0 으로 만드는 실험을 한 이유는 파라미터의 비트 단위 정보가 정확도에 미치는 영향력을 알아보기 위함이다. 영향력이 있는 비트의 개수를 측정하고, 비트를 최소한으로 사용 가능하도록 하여 모델 경량화에 도움이 되고자 실험을 진행하였다.

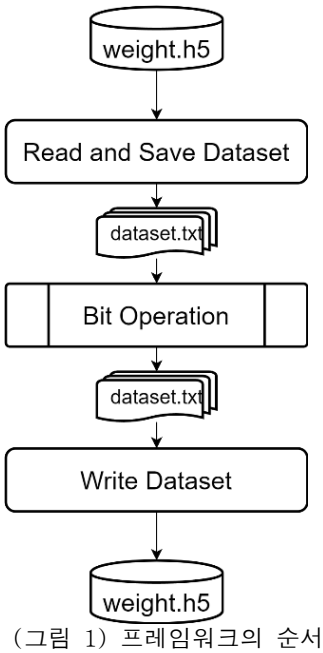
실험은 모델들의 파라미터의 비트를 자동으로 연산하는 프레임워크를 개발하여 활용하였다. 기존 모델 파라미터의 IEEE 754 32-bit floating point 를 하위 2-bit 씩 0 으로 만듦으로써 정보 손실을 발생시켰다. 성능 중심 모델 InceptionV3, InceptionResnetV2, ResNet50 과 경량화 중심 모델 Xception, DenseNet121, MobileNetV1, MobileNetV2 총 7 가지 모델을 선정하였다.

2. 파라미터 비트 연산 자동화 프레임워크

HDF5(Hierarchical Data Format 5)은 Tensorflow 에서 모델의 파라미터를 저장할 때 주로 사용하는 파일 형식이다. 우리는 HDF5 형식의 파라미터를 자유롭게 접근하고, 1-bit 단위로 쉽게 연산하여 경량화를 위한 실험을 손쉽게 하고자 하였다. 따라서 HDF5 형식의 파라미터를 자동으로 비트 연산을 할 수 있는 프레임워크를 개발하였다.

그림 1 의 전체적인 순서도를 보면 알 수 있듯이,

HDF5 형식으로 저장된 파라미터의 각 레이어 별 데이터셋을 txt 파일로 저장한 뒤, 비트 연산을 수행하고, 최종적으로 데이터셋을 각 레이어와 차원의 형태에 맞게 HDF5 형식으로 저장한다.



(그림 1) 프레임워크의 순서도

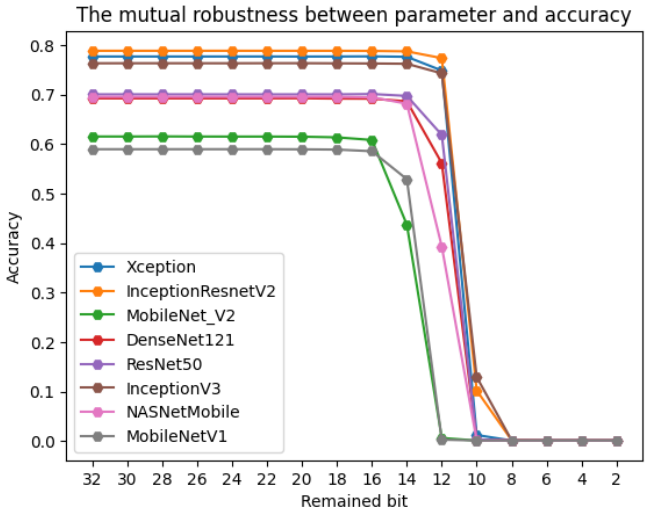
이 프레임워크에서 수행되는 비트 연산은 어떠한 비트 연산도 될 수 있지만, 본 논문에서 사용한 비트 연산은 파라미터와 정확도 간의 상호 강인성을 연구하기 위함이므로 32-bit floating point 의 하위 n-bit 를 0 으로 만드는 연산을 수행하였다. bit shift 연산을 통해 오른쪽으로 n-bit 만큼 shift 하고, 왼쪽으로 n-bit 만큼 shift 하여 n-bit 만큼의 하위 비트를 0 으로 만들도록 하였다.

3. 실험 및 결과 분석

본 연구는 ImageNet 데이터셋으로 사전 학습된 32-bit floating point 기반의 성능 중심의 모델 InceptionV3, InceptionResnetV2, ResNet50 과 경량화 중심의 모델 Xception, DenseNet121, MobileNetV1, MobileNetV2 로, 총 7 가지 모델을 선정하였다. 실험은 모든 모델에 대해 Optimizer 는 Adam 을 사용하였고, 손실 함수는 SparseCategoricalCrossentropy 를 사용하여 ImageNet 검증 데이터셋 50000 장으로 정확도를 평가했다.

우리는 하위 n-bit 의 정보 손실에 따른 정확도 변화를 관찰하기 위해 LSB 부터 2-bit 씩 0 으로 처리하였다. 결과적으로 32-bit, 30-bit, …, 4-bit, 2-bit 로 남은 비트의 파라미터를 사용하여 실험 결과로 나오는 정확도 변화를 관찰하였다.

그림 2 를 보면, 32-bit floating point 로 학습된 모델 중 약 60%의 정확도를 보인 MobileNetV1, MobilenetV2 의 경우 남은 비트가 16-bit 일 때까지 정확도가 유지되며, 14-bit 부터 정확도가 급격히 떨어졌다. 약 70%의 정확도를 보인 DenseNet121, ResNet50, NASNetMobile 의 경우 14-bit 까지 정확도가 유지되며, 12-bit 부터 정확도가 급격히 떨어지는 모습을 확인할 수 있었다. 약 80%의 정확도를 보인 InceptionResnetV2, Xception, InceptionV3 의 정확도는 12-bit 까지 유지되며, 10-bit 부터 상대적으로 조금 떨어졌다. 성능이 낮은 모델일수록 파라미터와 정확도 간의 강인성이 높아 성능이 좋은 모델보다 정확도를 유지하는 비트 수가 적다는 것을 실험적으로 확인하였다.



(그림 2) 네트워크의 실험 정확도 결과

4. 결론

본 논문은 파라미터를 1-bit 단위의 비트 연산을 자동화하는 프레임워크를 제안하며, CNN 모델들의 파라미터와 정확도 간의 상호 강인성에 대한 실험 및 연구 결과를 제시한다. ImageNet 데이터셋으로 사전 학습된 CNN 모델들의 파라미터의 하위 n-bit 를 2-bit 씩 0 으로 만들며 ImageNet 검증 데이터셋으로 평가해본 결과, 성능이 낮은 모델일수록 파라미터와 정확도 간의 강인성이 높아 정확도를 유지하는 비트 수가 적음을 발견하였다. 향후 여러 CNN 모델의 각 레이어 별 파라미터와 정확도 간의 상호 강인성을 연구할 예정이다.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (NRF-2021R1F1A1048026).

참고문헌

[1] Han Cai, Ligeng Zhu, Song Han “ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware” International Conference on Learning Representations (ICLR) Addis Ababa, Ethiopia 2019

[2] Song Han, Huizi Mao, William J. Dally “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding” IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Las Vegas, NV, USA 2016 1-9