**Weight Recycle을 이용한 모델 압축**

팀원 : 2014104117 양원영

**개요**

현재 사용되는 딥러닝 모델들은 엄청난 메모리와 연산량을 요구한다. 그렇기 때문에 모바일 기기등에 적용하기에는 너무나도 큰 용량을 차지한다. 또 앱스토어에서는 100MB가 넘는 앱은 Wi-fi로밖에 다운받지 못한다. 이 외에도 여러 이유때문에 Deep Learning Model Compression은 매우 중요한 문제이다.

그래서 딥러닝 모델의 Weight를 Recycle하여 용량을 압축하는 기법에 대해 연구해 보려고 한다.

1. **서론**
   1. **연구배경**

최근에 영상관련 연구주제에서 딥러닝이 가져온 성과는 대단하다. 하지만 좋은 성능을 보인 모델들은 매우 깊은 신경망으로 이루어져 있어 그 크기가 매우 크다는 문제가 있다. 이는 모바일 디바이스들과 같이 하드웨어 리소스적 제한이 있는 환경에 적용되기 힘들다는 것을 의미하는데, 이를 위해 고안 된 것이 모델 압축 기법이다.

본문에서는 Weight를 90도, 180도, 270도로 회전시킨 후 기존에 Weight에다가 합쳐서 사용하여 본래의 Weight 용량을 1/4로 줄이는 모델 압축 기법에 대해 연구해 보려고 한다.

* 1. **연구 목표**

큰 성능의 저하 없이 Weight Recycle을 통해 용량을 1/4로 줄이는 것을 목표로 한다. 기존의 연구가 없는 새로운 모델 압축 기법이다보니 결과가 미지수이다. 그렇기 때문에 많은 실험과 분석이 필요할 것이다. 그리고 분석 후 개선의 여지를 파악하고 성능 향상 방안을 모색한다.

1. **기존 연구**
   1. **기존 연구 1**

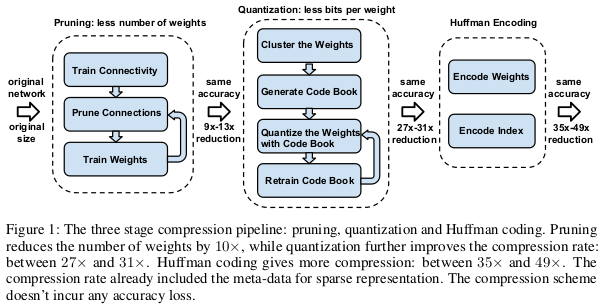
Deep Compressiong: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding

3가지의 Model compressing 기법에 대해 연구한 논문이다. Pruning, Quantization, Huffman coding 이 세가지 기법에 대해 연구되었다.

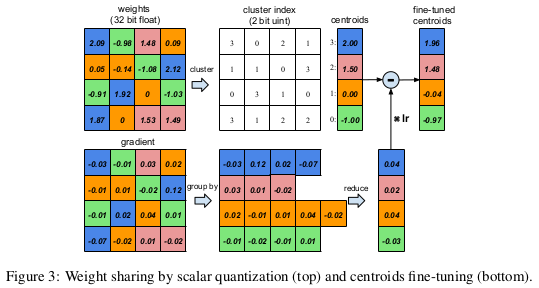
Pruning은 모델의 Weight들의 값을 조사한 다음 threshold를 정하고 그보다 0에 가까운 Weight들은 0으로 바꾼 뒤 더이상 학습시키지 않는 방법이다

Quantization은 Pruning 후 남은 모델의 weight들을 어느정도의 기준으로 묶은 다음에 한 묶음에 있는 weight 값들을 한가지 값으로 통일시키는 방법이다.

Huffman coding은 앞서 Pruning과 Quantization을 한 후에 많이 쓰이는 값부터 순서대로 작은 bit를 할당하는 방법이다.



(Deep Compression)



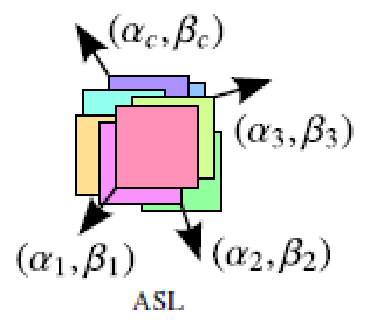
(Quantization and Weight Sharing)

* 1. **기존 연구 2**

Constructing Fast Network through Deconstruction of Convolution

Active Shift Layer을 사용한 모델 압축에 대해 연구된 논문이다.

Active Shift Layer이란 Weight의 Channel을 Shift 시켜 1x1 convolution layer로도 그보다 큰 convolution layer의 역할을 할 수 있게 만들어주는 기법이다.



(Active Shift Layer)

* 1. **기존 연구의 문제점**

기존 연구 1에서 설명한 Deep Compression의 Pruning같은 경우 0에서 먼 Weight가 많을 경우 성능의 저하 또는 압축률이 저조할 수가 있다. Quantization 기법 같은 경우도 Weight들의 거리가 멀면 큰 성능의 저하가 올 수 있다.

이처럼 앞서 설명한 기존의 모델 압축 기법들은 경우에 따라 큰 성능의 하락 또는 압축률이 저조한 현상이 생길 수 있다. 하지만 본 연구의 모델 압축 기법은 모든 경우에 Weight들을 1/4로 줄일 수 있다.

만약 성공적인 결과가 나오지 않는다 하여도 자세한 원인을 분석하여 차후 개선해 나가는 것 또한 목표이다.

1. **프로젝트**
   1. **기존 연구와 차이점 및 해결방안**

본 연구는 기존에 없던 새로운 모델 압축 기법을 고안해 내었다. 그렇기 때문에 기존에 참고할 만한 자료도, 구현된 코드도 없다. 또한 실험 후 결과가 좋을 수도 나쁠 수도 있다. 일단 기본적인 딥러닝 모델로 구현을 한 후에 많은 실험을 거쳐야 할 것이다. 실험에서 좋은 결과가 나왔을 경우 왜 좋은 결과가 나왔는지, 어떻게 개선할 것인지를 분석할 것이고, 나쁜 결과가 나왔을 경우에도 나쁜 결과가 나온 이유를 분석하고 분석된 부분을 어떻게 개선할 것인지를 도출해 낼 것이다.

* 1. **프로젝트 내용**

Weight를 90도, 180도, 270도로 회전시킨 후 하나로 합쳐서 모델의 크기를 1/4로 줄이는 모델 압축 기법에 대해 연구하는 것이 본 프로젝트의 내용이다.

1. **진행 일정**

|  |  |  |
| --- | --- | --- |
| 진행 주차 | 담당 | 내용 |
| 1~3 | 양원영 | 주제 선정 및 참고자료 조사 |
| 4~7 | 양원영 | 기술 적용 모델 구현 |
| 8 | 양원영 | 구현 모델 성능 테스트 |
| 9 | 양원영 | 실험 결과 분석 및 결론 |
| 10~16 | 양원영 | 4~9주차 루틴 반복 |

1. **결론**

딥러닝 기술은 모델의 크기 때문에 모바일 디바이스에서 동작하기에는 한계가 있었다. 하지만 이를 극복하기 위한 새로운 모델 압축 기법을 고안해 냈고 이를 구현하여 성능을 확인해 볼 것이다. 결과가 나오면 결과가 왜 그렇게 나왔는지를 분석하여 개선시켜 나갈 것이며 이 프로젝트가 성공한다면 모바일 기기와 같은 하드웨어적 제한이 있는 환경에서도 성공적으로 적용할 수 있을 것이다.

**참고 문헌**

[1] Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, [Song Han](https://arxiv.org/search/cs?searchtype=author&query=Han%2C+S), [Huizi Mao](https://arxiv.org/search/cs?searchtype=author&query=Mao%2C+H), [William J. Dally](https://arxiv.org/search/cs?searchtype=author&query=Dally%2C+W+J)

[2] Constructing Fast Network through Deconstruction of Convolution, Yunho Jeon, junmo Kim