**Super Resolution Using Deep Compression Pipeline**

팀원 : 2013104081 배준기

2014104117 양원영

2015110477 연주은

**개 요**

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.

기존의 Image Classification에 적용되어 좋은 성능을 보였던 Deep Compression 기법을 Super Resolution에 적용한 후, 실험을 통해 분석을 하고자 한다.

1. **서론**
   1. **연구 배경**

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.

본문에서 사용하는 Deep Compression은 Pruning, Quantization, Huffman coding 총 세 가지 기법으로 이루어진 파이프라인이다. 기존의 연구에서 LeNet-300-100, LeNet-5, AlexNet, VGGNet에 적용한 결과, 뛰어난 압축력과 성능 보존 능력을 증명했다. 하지만 이는 Image Classification Problem에 적용하여 얻은 결과이며, 이를 Super Resolution Problem에 적용해도 같은 결과를 보일까 의문이 생겼다.

* 1. **연구 목표**

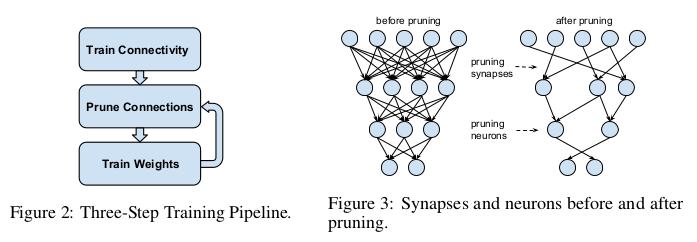
같은 Classification Problem에서는 모델을 바꾸어 가며 실험을 해도 동일하게 좋은 성능을 보였다. 하지만 Classification Problem에서 좋은 성능을 보였다고 하더라도 Regression Problem에서 동일한 성능을 보이는 것은 아니다. Deep Compression을 Classification Problem에 적용했을 때 보인 긍정적인 효과가 과연 Regression Problem에서도 나타나는지 확인하고 그 결과의 원인에 대해 분석하고 증명한다. 그리고 개선의 여지를 파악한 후 성능 향상 방안을 모색한다.

**2. 기존 연구**

2.1 기존 연구 1

Learning both Weight and Connections for Efficient Neural Networks

기본에 있던 Model compressiong 기법중에 하나인 Pruning에 대해 연구된 논문이다. Pruning이란 모델의 Weight들의 값을 조사한 다음 threshold를 정하고 그보다 0에 가까운 Weight들은 0으로 바꾼 뒤 더이상 학습시키지 않는 방법이다. 이 방법은 크게는 weight를 13배 줄이면서 정확도의 손실이 없게 만들었다.



(Pruning)

2.2 기존 연구 2

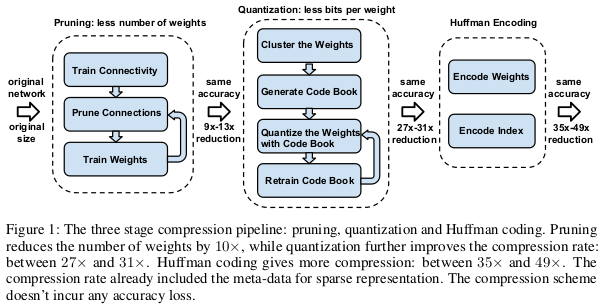
Deep Compressiong: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding

기존 연구 1에 나온 Pruning을 포함해 3가지의 Model compressing 기법에 대해 연구한 논문이다. 앞서 설명한 Pruning에 이어 Quantization과 Huffman coding에 대해 연구되었다.

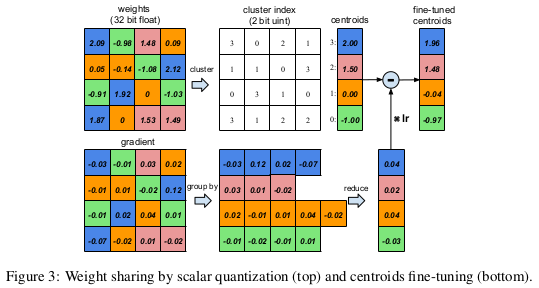
Quantization은 Pruning 후 남은 모델의 weight들을 어느정도의 기준으로 묶은 다음에 한 묶음에 있는 weight 값들을 한가지 값으로 통일시키는 방법이다.

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

위의 세가지 기법을 사용해서 크게는 weight를 49배까지 정확도의 손실 없이 줄일 수 있었다.

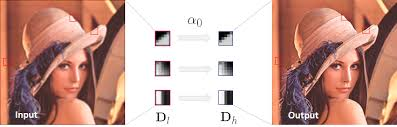


(Deep Compression)



(Quantization and Weight Sharing)

**2.3 기존 연구의 문제점**

****

(Example of Super Resolution)

앞서 설명한 기존의 연구들은 이미 여러 분야의 모델들에 적용되어 큰 성공을 이루어 왔다. 하지만 Super Resolution 분야에서는 아직 좋은 결과가 나오지 않았다. 다른 분야보다 Super Resolution은 Weight 하나하나의 값이 중요해서 Model compression을 적용하면 좋지 않는 결과가 나온다. 우리는 이 문제를 해결하고 Super Resolution 분야에 성공적으로 적용하는것이 최종 목표이다. 만약 성공적으로 적용하지 못한다고 하여도 자세한 원인을 분석하여 차후 개선해 나가는 것 또한 목표이다.

**3. 프로젝트**

**3.1 기존 연구와 차이점 및 해결방안**

Super-resolution은 저해상도(low resolution)의 영상을 고해상도(high resolution)로 복원 또는 재구성하는 딥러닝이 주로 쓰이는 영상 처리 기법 중 하나이다. 딥러닝은 4차 산업 혁명의 핵심 기술로서 컴퓨터 비전, 패턴 인식, 게임 등 많은 분야에 쓰이고 있다. 하지만 딥러닝은 모델의 복잡성으로 인해 모델 학습, 저장 및 실행에 많은 시간과 공간이 필요하여 스마트폰 등의 모바일 기기에서 사용하기 어렵다. 이러한 복잡성을 줄이는 여러 가지 기법이 존재하는데, 대표적으로는 Pruning, Quantization, Weight Sharing, Huffman coding이 있다. 기존의 연구에서는 Lenet-300-100, Lenet-5, AlexNet, VGG-16에 Pruning, Quantization, Huffman coding을 순차적으로 적용하는 압축 파이프라인을 사용하였고, 정확도의 손실 없이 모델을 35~49배 압축하는 데 성공하였다. 그러나 이는 classification에 적용한 결과이며 super resolution에도 좋은 압축률을 보일지는 모른다.

이를 확인하기 위해 먼저 Pruning, quantization, huffman coding을 super resolution model에 적용시켜 본다. 압축 후의 Accuracy와 weight 압축률을 구한다. 이 결과가 좋게 나오면 좋게 나온 이유를 분석하고, 결과가 좋지 않으면 Pruning의 threshold의 기준점과 Quantization의 나누는 정도 등을 조절해서 가장 좋은 Hyper-parameter을 구한다. 적절한 Hyper-parameter를 적용했을때 좋은 결과가 나오면 그 이유를 분석하고, 좋지 않으면 그 원인을 분석하고 더 나은 방법을 모색한다.

**3.2 프로젝트 내용**

Classification Problem에만 적용된 기존 Deep Compression 연구와 달리 Deep Compression 기법을 super resolution Problem에 적용하여 압축된 모델을 만드는 것이 본 프로젝트의 내용이다.

**4. 진행 일정**

|  |  |  |
| --- | --- | --- |
| 진행 주차 | 분류 | 내용 |
| 1-3주차 | - | 주제 선정 및 참고자료 조사 |
| 4주차 | Survey | 관련 논문 분석 및 정리 |
| 5주차 | Implementation | 기술 적용 모델 구현 |
| 6주차 | Experiment | 구현 모델 성능 테스트 |
| 7주차 | Analysis | 실험 결과 분석 및 결론 |
| 8주차 | Writing & Submission | 결론 취합 후 논문 작성 및 제출 |
| 9-16 | Review & Final | 4-8주차 루틴 반복 수행 |

**5. 결론**

딥러닝 기술은 모델의 크기 때문에 모바일 디바이스에서 동작하기에는 한계가 있었다. 하지만 이를 극복하기 위한 Model Compression 기법들이 소개되었고, Classification Problem에서 긍정적 효과를 입증하였다. 이를 Super Resolution Problem에 적용하여 효과를 확인해 볼 것이다. 이 프로젝트가 성공한다면 모바일 기기에서도 Super-resolution이 성공적으로 적용될 것이다.

**참고 문헌**

[1] Learning both Weights and Connections for Efficient Neural Networks, [Song Han](https://arxiv.org/search/cs?searchtype=author&query=Han%2C+S), [Jeff Pool](https://arxiv.org/search/cs?searchtype=author&query=Pool%2C+J), [John Tran](https://arxiv.org/search/cs?searchtype=author&query=Tran%2C+J), [William J. Dally](https://arxiv.org/search/cs?searchtype=author&query=Dally%2C+W+J)

# [2] 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)

추후 추가 예정