가중치 프루닝 기법을 이용한 새로운 초경량 초해상화 심층신경망 구조

배준기○ 양원영 연주은 배성호
경희대학교 소프트웨어융합대학 컴퓨터공학과
2013104081@khu.ac.kr annyeon@khu.ac.kr, shbae@khu.ac.kr

A Novel Light-Weight Deep Neural Network Architecture for Super Resolution using a Weight Pruning Method

Joon-Ki Bae○ Won-Young Yang Ju-Eun Yeon, Sung-Ho Bae
Kyung Hee University, The college of software convergence,
Department of Computer Engineering

요 약

최근 딥 러닝은 많은 분야에서 좋은 성과를 내고 있지만 성능에만 초점을 맞춘 나머지 하드웨어적 제한이 있는 환경에 적용하기 힘들다는 문제점이 발생한다. 이를 해결하기 위해 고안된 것이 모델 압축이다. 모델 압축 방법은 영상 분류 문제에서 상당한 모델 경량화 성능을 달성하였지만, 초해상화와 같은 영상복원 인공신경망 구조에는 아직 적용된 적이 없다. 본 논문은, 최초로 모델 압축 기법 중 가장 좋은 성능을 보이는 가중치 프루닝 기법을 초해상화용 인공신경망 구조에 적용하였다. 실험 결과, 가중치 프루닝 기법을 적용한 제안 초해상화 인공신경망은 기존 방법 대비 복원 성능(PSNR)의 저하 없이 최대 32배(약 97%)의 모델 압축 성능을 보였다. 이는 괄목할 만한 결과이며, 향후 본 연구가 모바일 응용과 같은 저전력/고성능 영상복원 응용에 매우 효과적으로 적용될 수 있음을 시사한다.

1. 서 론

최근 컴퓨터 비전 분야에서 딥 러닝이 가져온 성과는 대단하다. 그 중에서도 중요한 분야인 영상 초해상화(super resolution)는 저해상도 영상에서 고해상도 영상을 복원하는 영상처리 기술을 말한다. 이 기술은 의학 영상, 보안, 휴대용 멀티미디어 기기 등 실생활에 널리 적용될 수 있다[1].

하지만 이전까지의 초해상화 연구들은 성능에만 초점을 맞추어 진행되었기 때문에 상용화에 있어서는 아직까지 여러 가지 문제들이 존재한다. 하나의 예로, 일반적으로 딥 러닝 모델이 깊은 신경망으로 이루어질수록 좋은 성능을 보이는데, 깊은 신경망일수록 모델의 매개변수 수가 많아 저장 공간을 많이 차지하게 된다. 이와 같은 이유로 딥 러닝 기반 방법을 하드웨어적 제한이 있는 환경에 적용하기 힘들다[2]. 이를 해결하려 한다는 관점에서 모델 압축은 매우 중요한 연구 주제다.

현재까지 영상복원을 위한 딥 러닝 모델 압축 기법은 거의 연구되지 않았다. 즉, 기존 Deep Compression[2]기법은 영상 분류용 딥 러닝 구조에 적용되어 획기적인 압축 성능을 달성하였지만, 아직 영상 복원용 딥 러닝 구조에는 적용된 적이 없다. 이는, 영상 분류와 영상 복원을 각각 분류(classification), 회귀(regression)문제로 취급할 수 있는데, 일반적으로 회귀문제가 훨씬 더 어렵기 때문이다.

본 논문에서는 Deep Compression[2] 딥 러닝 모델 압축 기법을 이용하여, 영상복원용 딥 러닝 모델의 성능을 최대한 보존하면서 획기적으로 경량화 시키고자 한다.

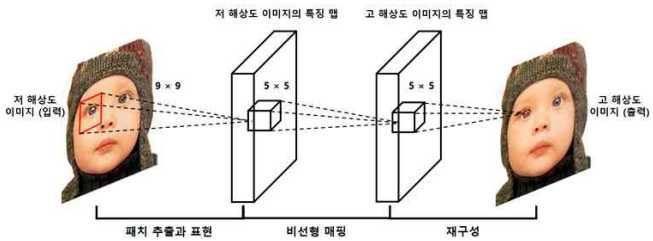


그림 1. SRCNN 모델 구조

2. 관련 연구

2.1 딥 러닝 초해상화 모델

본 논문에서는 단일 영상 초해상화(Single image Super Resolution) 문제에 최초로 딥 러닝을 적용한 SRCNN[1]을 기본 모델로 사용하였다. 그림 1과 같이 세 개의 층으로 이루어진 컨볼루션 신경망(Convolution Neural Network)이며, 세 가지 단계로 구성되어 있다: 1) 저해상도 이미지로부터 패치를 추출하는 단계, 2) 다차원 패치 벡터를 다른 다차원 패치 벡터로 매핑하는 단계, 3) 다차원 패치 벡터에서 고해상도 이미지를 생성하는 단계이다. 본 연구에서는 SRCNN 모델 중 가장 성능이 좋았던 9-5-5 모델[1]을 사용하였다. (9-5-5는 3개의 컨볼루션 층을 가지며 각 층의 컨볼루션 커널의 크기가 각각 9×9, 5×5, 5×5를 의미한다.)

2.2 딥 러닝 압축 기법

본 연구는 Deep Compression[2]의 기법을 사용한다. Deep Compression은 그림 2와 같이 세 가지 방법으로 파이프라인을 형성하여 수행된다: 1) 가지치기(pruning)는 컨볼루션 필터 내의 가중치(weight) 중 임계값 미만의 값을 가지는 가중치를 0으로 치환하여 더 이상 학습시키지 않는 방법이다[2]. 2) 양자화(quantization)는 모델의 가중치들을 비슷한 값을 가지는 가중치들끼리 군집을 형성한 후 각 군집의 대푯값만을 취하는 방식이다[2]. 3) 허프만 인코딩(Huffman coding)은 많이 쓰이는 값부터 순서대로 작은 bit를 할당하여 표현하는 방법이다.

이 중 양자화 및 허프만 인코딩 방법은 추가적인 연산을 유발하기 때문에[2], 본 연구에서는 추가적인 연산 없이 모델을 경량화 할 수 있으며 가장 높은 압축률을 보인 가지치기 기법을 SRCNN 모델에 적용하였다.

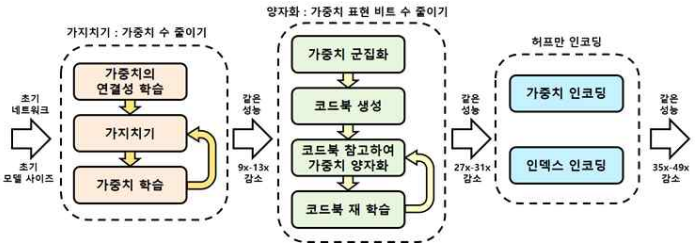


그림 2. Deep Compression 파이프라인

3. 제안 방법

그림 2를 보면, 가지치기를 수행한 후 다시 학습을 하는 과정의 반복이다. 첫 가지치기 후에 바로 종료하게 된다면 성능 손실이 크게 발생할 것을 직관적으로 알 수 있다. 재학습을 위한 장치로 마스크를 구현하였다. 그림 3은 기존의 컨볼루션 연산을 보여주고 있고, 그림 4는 가지치기를 위한 컨볼루션 연산을 보여준다. 여기서 마스크의 역할은 담당 필터 내의 가중치들이 연산에 관여할 것인지에 대한 연결 여부를 1 또는 0으로 저장한다.

기존 SRCNN[1]에 제시된 모델과 제안된 모델이 영 덧대기(zero-padding)가 추가된 점과 오차를 측정하는 방법이 달라, 성능을 재 측정하였다. 재 측정된 자료를 바탕으로 재학습을 수행할 모델을 택하였다.

가지치기 임계값으로는 가지치기를 수행할 각각의 컨볼루션 필터에 대해 가중치들의 표준편차를 구하여 가중치들의 절댓값과의 비교를 통해 가지치기를 수행했다.

4. 실험 결과 분석 및 토론

영상 손실 압축에서 화질 손실 정보를 평가할 때 사용하는 최대 신호 대 잡음비(PSNR)을 사용했다.

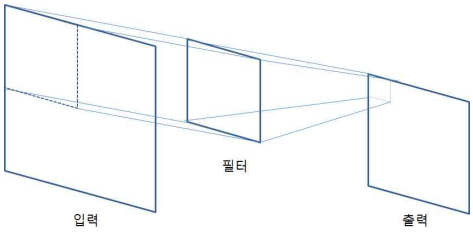


그림 3. 기존 컨볼루션 연산 형태

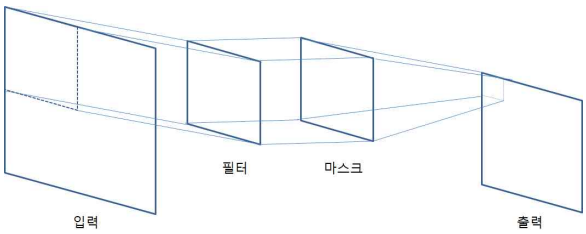


그림 4. 제안된 컨볼루션 연산 형태

4.1 가지치기 적용 모델 선정

가장 먼저 어느 시점의 SRCNN에 가지치기를 적용할 것인지에 대한 실험을 수행했다. 제안된 SRCNN[1]기반 모델을 학습시킨 후 특정 시점의 모델을 선정하였다. 오차 함수로 MSE(Mean Square Error)를 사용하였기 때문에 PSNR을 관찰함으로 학습의 진척을 평가할 수 있다. 1500세대 이후를 수렴상태로 간주하여, 해당 시점의 모델을 대상 모델로 선정하였다.

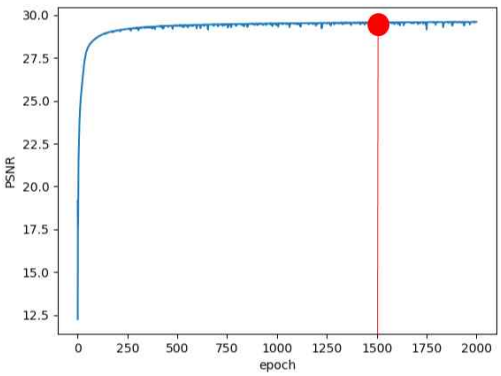


그림 5. 가지치기 적용 모델 선정

4.2 가지치기 적용

그림 6은 가지치기 적용 전의 모델과 가지치기 적용 후의 모델의 학습 상황을 비교한 것이다. 여기서 알 수 있는 중요한 사실은 첫 번째, 가지치기를 적용한 후에 재학습을 통해 기존의 성능과 거의 유사한 성능을 낼 수 있다는 점이다. 이는 컨볼루션 필터 내에서도 중요하게 다루어지는 가중치들이 있고, 그 가중치들로만 학습을 진행해도 기존의 모델과 거의 동등한 성능을 나타낼 수 있다는 것이다. 두 번째, 가지치기를 수행한 후에 기존의 성능을 따라잡기 위해서 더 많은 세대(epoch)수가 필요하다는 점이다. 이로부터 같은 조건 하에서는 가중치가 많은 쪽이 학습이 더 빠르다는 사실을 알 수 있다.

그림 7은 가지치기를 시행한 횟수에 따른 결과이다. 가

지치기가 거듭될수록 초기 PSNR이 높아지는데 이는 회를 거듭할수록 가지치기되는 가중치의 개수가 줄어듦에 따라 이전 모델의 최종학습 상태와의 차이가 줄어들기 때문이다.

표 1은 가지치기를 시행한 횟수에 따른 파라미터 수와 압축률의 변화를 나타낸 것이다. 원 모델의 파라미터 수는 69443개이고, 가지치기를 수행함에 따라 그 수가 1회 3201개, 2회 1876개, 3회 1350개, 4회 1070개로 줄었다. 첫 가지치기의 압축률은 95.39% 이고, 4회 수행 후에는 원 모델 대비 98.46%가 압축된다. 그럼에도 불구하고 성능은 가지치기 전 모델과 차이가 없다.

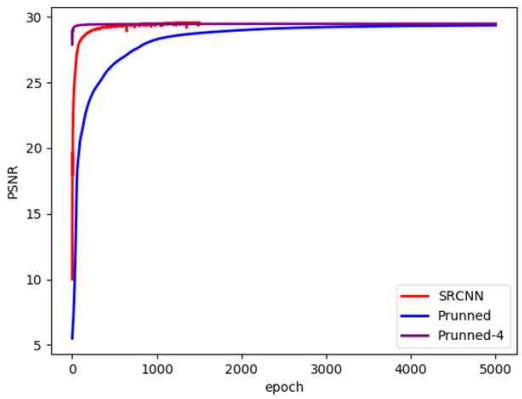


그림 6. 가지치기 적용 결과 비교

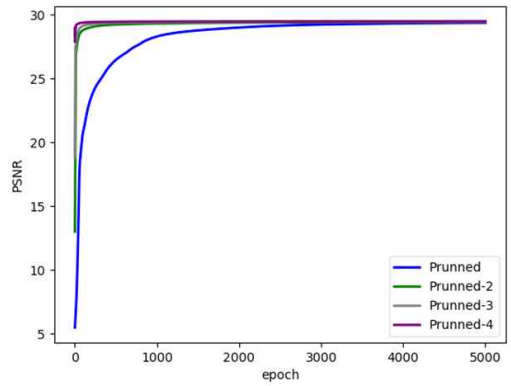


그림 7. 가지치기 횟수 별 비교

표 1. 가지치기 횟수 별 압축률

가지치기 수	PSNR	파라미터 수	압축률
0번	32.2881	694443	0.00%
1번	32.2171	3201	95.39%
2번	32.2171	1876	97.30%
3번	32.2222	1350	98.06%
4번	32.2239	1070	98.46%

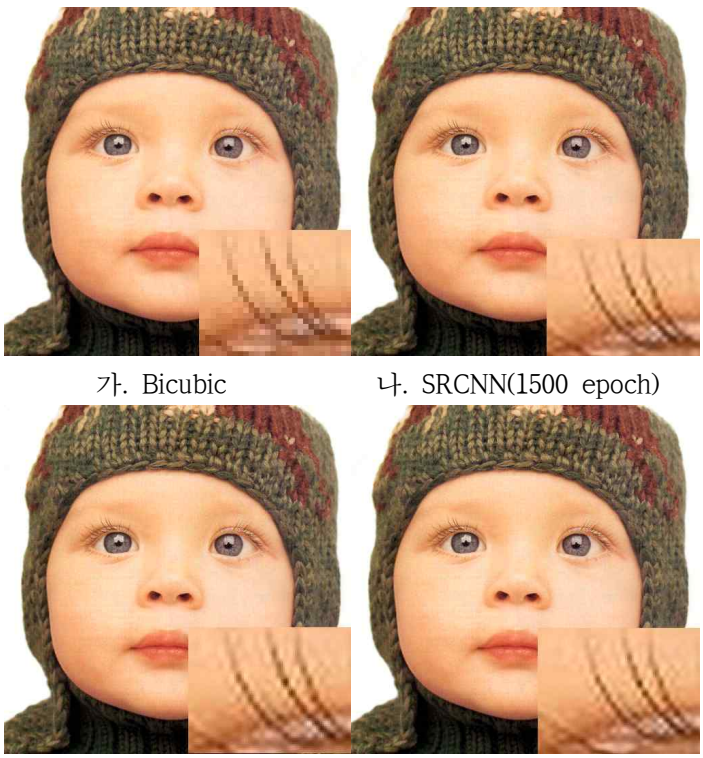
4.3 영상 비교

기존 초해상화 연구 중 하나인 SRGAN[3]에서 이미 PSNR의 목적 함수로서의 적절성에 대해 논의된 적이 있다. 이에 기반 하여 가지치기 후 재학습을 통해 보존된 성능이 PSNR 수치만으로 증명될 수 없다고 할 수 있다. 그렇기 때문에 직접 눈으로 복원된 영상을 보고 평가하는 과정이 필요하다.

그림 8은 각각의 경우에 해당하는 출력을 보여준다. 그림 8-가에 비해 딥 러닝 기술을 적용한 결과들이 뛰어난 복원력을 보였음을 알 수 있다. 또한, 가지치기를 수행하지 않은 SRCNN[1]과 비교해 보았을 때, 가지치기를 1회 수행한 것과 4회 수행한 것이 거의 비슷한 성능을 보였음을 알 수 있다.

5. 결 론

본 논문은 초해상화 분야에 Deep Compression[2] 딥 러닝 압축 기법 중 가지치기를 적용하여 초경량 모델을 얻을 수 있음을 보였다. 가지치기 기법을 적용했을 때, 학습에 필요한 시간은 늘어나지만, 결국에는 기존의 성능을 거의 보존해 낼 수 있다는 사실을 증명했다. 가지치기를 적용한 모델은 수치적으로 뿐만 아니라 시각적으로도 눈에 띄는 성능을 보였다. 이 결과를 통해 SRCNN[1] 뿐만 아니라, 다른 초해상화 모델들도 가지치기 기법을 통해 경량화시킬 수 있을 것이라는 희망을 얻었다. 우리는 더 나아가, 양자화(quantization) 및 허프만 인코딩(Huffman encoding)[2]을 포함한 파이프라인을 구축한 후 성능을 확인할 것이다.



가. Bicubic 나. SRCNN(1500 epoch)

다. 가지치기-1(5000 epoch) 라. 가지치기-4(5000 epoch)

그림 8. 실험 결과 비교

참고문헌

[1] Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang, “Image Super-Resolution Using Deep Convolutional Networks” IEEE 2015

[2] Song Han, Huizi Mao, William J. Dally, “Deep Compression: Compressiong Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding” ICLR 2016

[3] Ledig, C. et al, “Photo-realistic single image superresolution using a generative adversarial network” CVPR, 2017
[4] Barrett, Richard, et al. Templates for the solution of linear systems: building blocks for iterative methods. Vol. 43. Siam, 1994
[5] MacQueen, James. “Some methods for classification and analysis of multivariate observations.“ Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. No. 14. 1967