계층별 양자화 기반 초해상화 다중 스케일 잔차 네트워크 압축

*황지원 **배성호
경희대학교
*hjw5167@khu.ac.kr **shbae@khu.ac.kr

A Model Compression for Super Resolution Multi Scale Residual Networks based on a Layer-wise Quantization

*Jiwon Hwang **Sung-Ho Bae
Kyung Hee University

요 약

기존의 초해상도 딥러닝 기법은 모델의 깊이가 깊어지면서, 좋은 성능을 내지만 점점 더 복잡해지고 있고, 실제로 사용하는데 있어 많은 시간을 요구한다. 이를 해결하기 위해, 우리는 딥러닝 모델의 가중치를 양자화 하여 추론시간을 줄이고자 한다. 초해상도 모델은 feature extraction, non-linear mapping, reconstruction 세 부분으로 나누어져 있으며, 레이어 사이에 많은 skip-connection 이 존재하는 특징이 있다. 따라서 양자화 시 최종 성능 하락에 미치는 영향력이 레이어 별로 다르며, 이를 감안하여 강화학습으로 레이어 별 최적 bit 를 찾아 성능 하락을 최소화한다. 본 논문에서는 Skip-connection 이 많이 존재하는 MSRN 을 사용하였으며, 결과에서 feature extraction, reconstruction 부분과 블록 내 특정 위치의 레이어가 항상 높은 bit 를 가짐을 알 수 있다. 기존에 영상 분류에 한정되어 사용되었던 혼합 bit 양자화를 사용하여 초해상도 딥러닝 기법의 모델 사이즈를 줄인 최초의 논문이며, 제안 방법은 모바일 등 제한된 환경에 적용 가능할 것으로 생각된다.

1. 서론

초해상도 딥러닝 기법(Super-Resolution) 이란 저해상도의 사진을 고해상도의 사진으로 만들어내는 것이다. 딥러닝 기반 SR 방법들[1,2,3,4,5]이 좋은 복원 능력을 보여주고 있다. 하지만 좋은 성능을 내기 위해 모델 사이즈를 대폭 늘림에 따라, 실시간으로 적용하지 못할 만큼 많은 양의 메모리가 필요하고, 추론시간이 오래 걸리는 문제가 존재한다. 따라서 성능을 유지하면서 추론시간을 줄이도록 하는 연구가 필요 하지만, 많은 연구가 이루어지고있지 않다.

딥러닝을 사용한 이미지 분류에서는 모델 압축을 위해 가중치 양자화(quantization)[6,7,8], 가지치기(pruning)[9,10], 증류(distillation)[11,12], 등의 연구가 활발히 이루어지고 있다. Super-Resolution 에서도 양자화와 가지치기 방법을 시도하였다[13,14]. 가지치기 방법[13]은 성능을 향상시키며 파라미터 수를 10%로 감소시켜 압축에 성공한 반면, 양자화를 super-resolution 에 적용하는 연구[14]는 모델 사이즈를 20%로 압축시키지만 많은 성능 하락을 가진다. 따라서 우리는 super-resolution 에 양자화를 더 효율적으로 적용할 수 있도록, 레이어 별 최적 양자화 bit 를 찾아 모델 사이즈를 10% 이하로 압축시키면서 성능 하락을 최소화하고자 한다.

SR 모델은 feature extraction, non-linear mapping, reconstruction 세 부분으로 나누어져 모델 해상도를 향상시키며, non-linear mapping 부분에서는 skip-connection 을 가진 같은 블록이 여러 번 반복된다. Feature extraction 부분과 reconstruction 부분은 모델의 앞,

뒷부분으로 정보 전달에 있어 중요하고, non-linear mapping 내에서도 다른 레이어와의 연결 정도에 따라 레이어의 중요도가 달라진다. 따라서 모든 레이어를 같은 bit 로 양자화 하게 되면, 모델의 성능을 유지하는데 더 중요한 레이어와 덜 중요한 레이어를 구분하지 않아 성능 하락이 불가피하다. 이를 해결하기 위해 레이어 별 최적 bit 를 강화학습으로 찾아 성능을 유지한다.

이미지 분류 분야에서 이미 이 문제를 해결하기 위해 여러 방법들[15,16,17]이 제안되었으며, 우리는 [15]가 제안한 방법 중 모델 사이즈를 줄이는 방법을 super-resolution 도메인에서 사용할 수 있도록 최적화하여 제안한다.

2. 관련 연구

2.1 Super-Resolution

SRCNN[1]은 처음으로 CNN 을 사용하여 SR 을 수행한 모델로, 레이어를 3 개만 쌓아서 만든 비교적 간단한 모델임에도 불구하고 기존 전통적인 머신 러닝 기반의 SR 성능을 능가한다. 이후 VDSR[2]이 skip connection 을 사용하여 residual learning 기법을 도입하면서 20 개의 레이어를 사용하여 SRCNN 보다 성능을 향상시켰다. 이를 응용하여 더 좋은 성능을 가진 더 깊은 모델들[3,4,5]이 나왔다. 하지만, 800 개의 레이어를 사용하는 등[4] 모델의 추론시간, 실시간성을 크게 고려하지 않았다.

2.2 양자화

양자화 란, 아날로그 데이터를 디지털 데이터로 바꾸어 표시하는 것이다. 일반적인 양자화는 디지털 데이터로 표현할 때 사용할 bit 수를 줄여 모델 사이즈를 줄인다. 이를 딥러닝 모델 가중치 압축에 적용하기 위해 많은 양자화 방법들이 연구되고 있고, 이미 학습 된 모델의 가중치를 양자화 하는 방법[6]과 모델을 학습시키면서 가중치 값을 양자화 될 값에 맞춰가는 방법[7,8]이 존재한다.

이미 학습된 모델의 가중치를 양자화 하는 방법 중, 더 최적화된 압축을 하기 위해 레이어 별로 다른 최적 bit 를 찾는 연구(mixed precision)[15,16,17]가 영상 분류 분야에서 활발히 진행되고 있다. 이 방법은 미리 학습된 가중치를 레이어 별로 다른 bit 로 양자화 하여 성능 하락을 최소화하는 방법이다. 미분 가능하게 문제를 해결한 [17]과, 강화학습을 사용하는 HAQ[15], AutoQ[16]가 존재한다. HAQ[15]는 DDPG[18]를 사용하여 레이어의 가중치 및 활성함수 이후의 값 별 최적 bit 를 찾았고, AutoQ[16]는 강화학습 agent 를 계층적으로 사용하는 HIRO[19]를 사용하여 커널 및 활성함수 이후의 값 별 최적 bit 를 찾았다.

3. 제안 모델

성능이 뛰어나도 모델의 계산 속도가 느리다면, 실시간으로 사용을 할 수가 없다. 우리는 super-resolution 모델들 중, skip-connection 이 많이 존재하는 MSRN[5]을 레이어 별로 최적 양자화 하여 압축하였고, 이를 바탕으로 모델의 구조를 분석하였다.

이미지 분류모델 압축에서 사용되는 mixed precision 방법 중, HAQ[15]의 방법을 SR 에 최적화하여 사용하였다. HAQ[15]는 두가지 제한을 두고 모델을 압축하는데, 첫 번째는 실제 디바이스에 모델을 올렸을 때의 latency 를 제한으로 두고 모델의 가중치와 활성함수 이후의 값을 압축하는 것이고, 두 번째는 모델 사이즈를 어느정도 압축하냐는 제한을 두고 가중치를 압축한다.

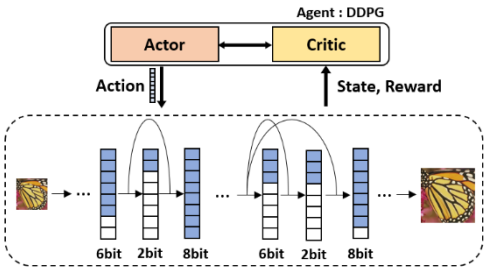

그림 1 제안하는 모델. 강화학습 에이전트인 DDPG 와 super-resolution 모델 간에 action, state, reward 를 주고받으며 레이어 별 최적 bit 를 학습. 그림 내 super-resolution 모델은 임의로 표현함.

그 중에, 모델 사이즈를 기준으로 한 압축 방법을 super-resolution 에 맞게 최적화하였다. 제안하는 모델은 그림 1 에서 볼 수 있다. 여러 강화학습 모델 중, DDPG[18]를 사용하여 각 convolution 레이어 별 비트를 찾는다. DDPG[18]는 정책을 근사하는 actor 와 action 을 뽑아내는 함수를 근사하는 critic 으로 구성되어 있다. Actor 가 현재 state 에 따라 정책에 맞는 bit 를 구하여 MSRN[5] 모델(environment)에 넘겨주면, 그에 대한 보상(reward)과 현재 상태(state)를 critic 에게 넘겨준다. Actor 는 critic 을 고려하여 또 다음 action 을 구하게 된다. 이 때, 모든 레이어들에 대한 bit 를 구하면, 하나의 에피소드가 완성된다. 각각의 에피소드마다 구해진 bit 로 양자화를 진행하여 보상을 구하고, 보상을 최대화하도록 이

[1] 본 논문은 2020 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2018R1C1B3008159)
[2] 본 논문은 과학기술정보통신부 및 정보통신산업진흥원의 '고성능 컴퓨팅 지원'사업으로부터 지원받아 수행하였음

과정을 반복한다.

이 때 양자화는 deep compression[6]에서 사용한 k-means clustering 방법을 사용하였으며, 총 600 에피소드를 반복하였다. Action 은 layer 별 bit 로 볼 수 있으며, State 는 비트를 찾을 때 고려하는 레이어 별로 입력 채널 수, 출력 채널 수, 가중치의 크기, 입력 feature-map 크기, 레이어 번호, 찾아진 bit 를 정규화 하여 사용하였다. DDPG[18]가 찾는 action 은 [0, 1]범위의 실수로, 아래 식을 사용하여 최종 자연수 bit 가 정해진다.

$$Bit = round(b_{min} - 0.5 + action \times (b_{max} \times b_{min} + 1))$$

MSRN	모델 압축률	Set5	Set14	B100	Urban100
full bit	0%	32.26	28.63	27.61	26.22
4bit 양자화	75%	31.85	28.40	27.47	25.85
Ours	90%	32.18	28.62	27.58	26.08

모든 레이어들에 대해 bit 를 찾아 하나의 에피소드가 완성되면 보상은 32 bit 부동 소수점으로 계산하였을 때의 PSNR 과의 차이에 람다 값을 곱해서 사용한다. 이 때, 람다 값은 0.1 로 사용한다.

우리는 많은 skip connection 을 가진 MSRN[5]을 베이스라인 모델로 사용하였고, 이 모델은 저해상도 이미지를 입력으로 받아 하나의 convolution 을 지나 8 개의 블록을 거쳐 고해상도 이미지로 reconstruction 된다. 이 때, 각각의 block 은 5 개의 convolution 레이어를 가지고 있으며, 서로 복잡하게 연결되어 있다.

4. 실험 및 결과

학습 데이터셋은 MSRN[5]과 같이 DIV2K[20]의 training set 800 장을 랜덤으로 잘라 16000 장으로 늘려 사용하였다. 강화학습 보상을 주기 위한 PSNR 은 DIV2K[20]의 training set 800 장 중, 1 번에서 80 번까지의 이미지의 PSNR 평균을 사용하였다. 테스트 데이터셋은 Set5, Set14, BSD100, Urban100 을 사용하였다.

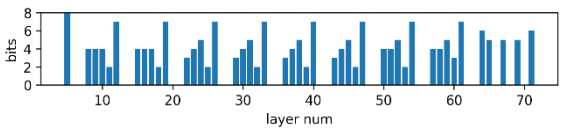

그림 2 MSRN scale 4 모델을 90% 압축할 때의 최적 양자화 bit

MSRN[5] scale 4 모델에 대해 제안한 방법으로 양자화를 시킨 결과는 그림 2 에서 볼 수 있다. 그림 2 에서 첫번째 레이어와 마지막 5 개의 레이어는 각각 feature extraction 과 reconstruction 부분이다. 가로축의 빈 공간은, ReLU 등 convolution 이 아닌 다른 레이어이기 때문에 bit 를 찾지 않는다(양자화를 하지 않는다). 중간에 비어 있는 공간을 기준으로 하나의 블록으로 볼 수 있으며, 실제로 각각의 블록 내 같은 위치의 레이어끼리 비슷한 중요도를 가지고 양자화되었음을 볼 수 있다. Feature extraction, reconstruction 및 각 block 의 마지막 layer 는 정보를 전달해야 하기 때문에 4bit 이상의 높은 결과가 나왔고, 각각의 block 내 네번째 layer 는 낮은 bit 만 사용해도 다른 연결들이 정보를 전달해주기 때문에 최종 성능을 하락시키지 않는 것을 볼 수 있다. 모델의 자세한 구조는 MSRN[5] 논문에서 볼 수 있다.

표 1 PSNR 비교표. 양자화 이전의 MSRN[5]성능, 4bit 로 모든 레이어를 양자화한 결과, 우리의 결과 순으로 나타낸 것.

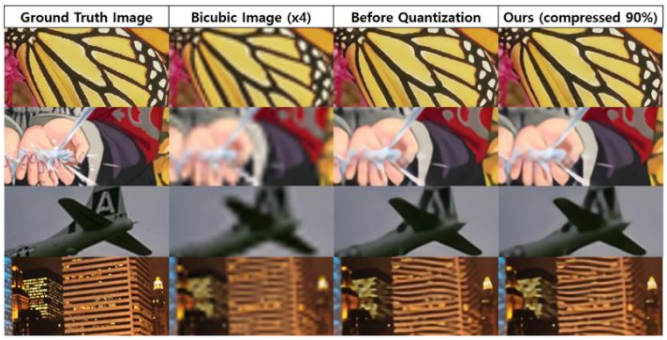

그림 3 제안한 방법의 성능. 맨 왼쪽부터 순서대로 원본 이미지, scale 4 bicubic 이미지, 양자화 이전의 기존 MSRN[5] 결과, MSRN[5]모델을 효율적으로 양자화 하여 모델 사이즈를 10%로 줄인 결과이다. 이미지는 위에서부터 각각 Set5, Set14, B100, Urban100 에서 랜덤으로 선택하였다.

표 1 은 제안한 방법을 사용하여 MSRN[5]의 모델 사이즈를 10%로 압축한 후의 결과를 full-precision 과 전체를 4bit 로 양자화 했을 때와 비교한 것이다. 시각적인 결과는 그림 3 에서 볼 수 있다. 결과를 보면, 가중치를 10%만 남겼음에도 불구하고 성능에 하락이 없는 것을 볼 수 있다.

5. 결론

Super-Resolution 딥러닝 모델들은 대부분 많은 skip-connection 을 가지고 있기 때문에, MSRN[5] 이외의 다른 모델들에 대해서도 같은 bit 로 모든 레이어를 압축하는 것 보다, 제안하는 방법을 사용하는 것이 압축률 대비 성능을 보존하는 방법이 될 수 있다. 따라서 이를 적절히 사용하여 큰 모델의 모델 사이즈를 성능을 유지하며 축소하여 실생활에 적용할 수 있을 것으로 기대된다.

참 고 문 헌

[1] Dong, Chao, et al. "Image super-resolution using deep convolutional networks." IEEE transactions on pattern analysis and machine intelligence 38.2 (2015): 295-307.

[2] Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee. "Accurate image super-resolution using very deep convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] Lim, Bee, et al. "Enhanced deep residual networks for single image super-resolution." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017.

[4] Zhang, Yulun, et al. "Image super-resolution using very deep residual channel attention networks." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[5] Li, Juncheng, et al. "Multi-scale residual network for image super-resolution." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[6] Han, Song, Huizi Mao, and William J. Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." arXiv preprint arXiv:1510.00149 (2015).

[7] Choi, Jungwook, et al. "Pact: Parameterized clipping activation for quantized neural networks." arXiv preprint arXiv:1805.06085 (2018).

[8] Esser, Steven K., et al. "Learned step size quantization." arXiv preprint arXiv:1902.08153 (2019).

[9] Li, Hao, et al. "Pruning filters for efficient convnets." arXiv preprint arXiv:1608.08710 (2016).

[10] Liu, Zechun, et al. "Metapruning: Meta learning for automatic neural network channel pruning." Proceedings of the IEEE International Conference on Computer Vision. 2019.

[11] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

[12] Zagoruyko, Sergey, and Nikos Komodakis. "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer." arXiv preprint arXiv:1612.03928 (2016).

[13] Hou, Zejiang, and Sun-Yuan Kung. "Efficient Image Super Resolution Via Channel Discriminative Deep Neural Network Pruning." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[14] Ma, Yinglan, et al. "Efficient Super Resolution Using Binarized Neural Network." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[15] Wang, Kuan, et al. "Haq: Hardware-aware automated quantization with mixed precision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.

[16] Lou, Qian, et al. "AutoQ: Automated Kernel-Wise Neural Network Quantization." arXiv preprint arXiv:1902.05690 (2019).

[17] Uhlich, Stefan, et al. "Mixed Precision DNNs: All you need is a good parametrization." arXiv preprint arXiv:1905.11452 (2019).

[18] Lillicrap, Timothy P., et al. "Continuous control with deep reinforcement learning." arXiv preprint arXiv:1509.02971 (2015).

[19] Nachum, Ofir, et al. "Data-efficient hierarchical reinforcement learning." Advances in Neural Information Processing Systems. 2018.

[20] Agustsson, Eirikur, and Radu Timofte. "Ntire 2017 challenge on single image super-resolution: Dataset and study." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017.