Attention 모델을 이용한 단일 영상 초고해상도 복원 기술

*문환복 **윤상민
국민대학교 컴퓨터공학과 HCI 연구실
*helloiwsy@kookmin.ac.kr, **smyoon@kookmin.ac.kr

A Study on Single Image Super Resolution Using Attention Model

*Hwanbok, Mun **Sang Min, Yoon
HCI Lab., Computer Science, Kookmin University

요약

단일 영상 기반 초고해상도 복원은 컴퓨터 비전 및 영상처리 분야의 중요한 기초 및 응용 분야 중 하나이며, 딥러닝에 대한 연구가 발전됨에 따라 이를 이용한 다양한 연구들이 활발히 진행되고 있다. 기존 딥러닝 기반 연구들은 복원 성능을 높이기 위해서 다양한 구조의 네트워크를 설계하거나 네트워크를 학습하는 알고리즘들을 중점으로 연구되어 왔다. 최근 들어 네트워크 구조나 설계 이외에 네트워크를 통과하는 정보의 집합체인 특징 맵에 관한 연구들이 진행되고 있다. Attention은 특징 맵에서 채널 간의 관계를 이용하여 특정 채널을 강조하거나 또는 공간 정보를 강조하는 방식으로 특징 맵의 정보를 잘 활용하도록 하여 전체적인 네트워크의 성능을 향상시킨다. 본 논문은 단일 영상 기반 초고해상도 복원 네트워크를 기반으로 다양한 Attention방법들을 적용하고 성능을 비교 및 분석한다.

1. 서론

단일 영상 기반 초고해상도 복원은 손상되거나 정보를 잃어버린 저해상도의 영상을 고해상도 영상으로 복원하는 것으로 컴퓨터 비전의 한 분야이다. 이는 의료, 치안, 미디어 산업 등 다양한 산업과 응용 연구 분야에 적용될 수 있는 중요한 기초 연구 분야이다.

딥러닝이 발전하면서 SRCNN[1]은 기존의 보간법 혹은 사전에 정의한 맵핑 알고리즘의 성능을 넘어서면서 이를 시작으로 딥러닝을 활용한 다양한 연구들이 진행되고 있다. VDSR[2]은 SRCNN보다 더 깊은 네트워크를 쌓아서 복원 성능을 향상시켰다. 이후 EDSR[3]은 이전 방법들이 L2 손실함수를 이용해서 네트워크를 학습했던 것과 달리 L1 손실 함수를 적용하였고 Residual 네트워크와 Skip Connection을 사용해서 더욱 깊은 네트워크를 학습 할 수 있었으며 따라서 높은 성능을 보여주었다.

딥러닝에 대한 진행된 연구들의 일환으로 네트워크를 통과하는 특징 맵 자체에 관심을 갖는 연구들이 있었다. SE-Net[4]은 Squeeze모듈과 Excitation모듈을 이용해서 특징 맵들의 관계를 명시적으로 모델링하는 Channel Attention을 제안했고 네트워크의 표현 능력과 성능을 향상시켰다. SCA-CNN[5]은 Channel Attention과 특징 맵의 공간 정보를 활용하는 Spatial Attention을 제안했다.

Attention에 대한 연구들이 소개되면서 이를 적용한 초고해상도 복원 연구들이 진행되었다. RCAN[6]은 복원 성능에 영향을 주는 특징 맵을 더욱 강조하기 위해서 Channel Attention을 적용했으며 SAN은 이전의 Attention 방법들이 평균값이나 최댓값으로 채널의 대푯값을 정하던 것과 달리 특징 맵의 공분산으로 특징 맵의 상관관계를 파악하고 대푯값으로 사용하는 방법을 제안했다. 본 논문에서는 Channel Attention과 Spatial Attention을 단일 영상 초고해상도 복원 네트워크에 적용해서 실험하고 결과를 비교 및 분석한다.

2. 본론

2-1. 네트워크 구조

본 논문은 그림 1과 같이 초고해상도 영상 복원을 위해 Residual In Residual[6]구조를 사용하며 입력 영상으로부터 동시에 Structure와 Texture를 복원하고 이를 더해서 최종적으로 복원 영상을 출력한다. 네트워크 구조에 대한 세부 사항은 표 1과 같다.

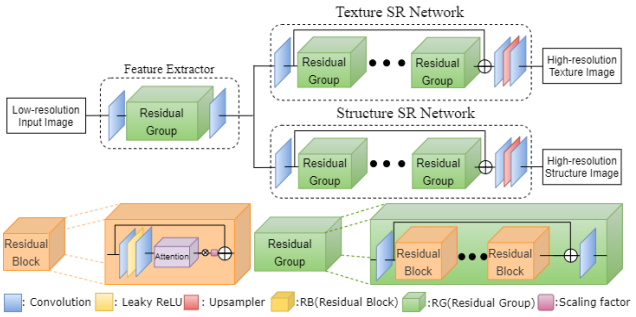

그림 1. 전체 네트워크 구조

네트워크의 입력은 저해상도 영상 I_{LR}이고 네트워크 F를 통해 복원된 Structure와 Texture는 각각 S_{SR}, T_{SR}이며 이를 더해서 최종적으로 복원한 영상 I_{SR}이다. 이를 수식으로 표현하면 다음 수식1과 같다.

$$I_{SR} = S_{SR} + T_{SR} = F(I_{LR}) \quad (1)$$

네트워크는 입력 영상으로부터 Feature Extractor를 이용해서 특징 맵을 추출한다. 추출된 특징 맵은 이후 두 개의 SR Network로 전달되고 각각 Structure와 Texture를 복원하기 때문에 풍부한 정보를 갖고 있어야한다. 추출된 특징 맵을 입력으로 받은 Texture, Structure SR Network은 Residual Block(RB)과 Residual Group(RG)을 통해서 복원된 Structure와 TextureS_{SR}, T_{SR}를 출력하고 두 영상을 더해서

표 1. 네트워크 세부사항

#Conv in RB	#RB in RG	#RG in SR Network	Local Skip Connection	Global Skip Connection
2	20	3	O	O

최종적으로 복원된 영상 I_{SR}을 출력한다. Feature Extractor와 SR Network은 Residual Group, Residual Blcok, Convolution layer로 구성된다. 각 Residual Group은 여러 개의 Residual Block으로 이루어져 있으며 Residual Block은 2개의 Convolution layer, 1개의 LeakyReLU 활성화 함수와 Attention 모듈로 구성된다. Residual Block에서 Convolution layer를 통과한 특징 맵은 Attention 모듈을 통해서 채널별로 혹은 공간적으로 강조된다. 또한 계산 비용을 낮추고 업샘플링 과정에서 발생할 수 있는 아티팩트를 줄이기 위해서 각 SR Network의 마지막에 Upsampler로 ESPCN[8]에서 제안한 PixelShuffler를 사용했다.

2-2. 손실 함수

네트워크를 학습하기 위해 사용하는 손실 함수는 수식2와 같다.

$$L = \lambda_T L_T + \lambda_S L_S \quad (2)$$

L는 전체 네트워크를 학습하는 손실 함수이며 Texture와 Structure SR Network에 대한 손실 함수인 L_T, L_S와 각 손실 함수에 대한 가중치인 λ_S와λ_T를 이용해서 계산된다. SR Network의 손실 함수의 수식은 다음 수식3과 같다.

$$L_T = \|T_{SR} - T_{HR}\|_1,\ L_S = \|S_{SR} - S_{HR}\|_1 \quad (3)$$

T_{HR}과 S_{HR}은 Ground Truth(GT)에 해당하는 Texture와 Structure영상이다. 네트워크는 두 개의 SR Network의 출력 영상 T_{SR}, S_{SR} 과 GT영상 T_{HR}, S_{HR} 간의 차이가 작아지도록 학습하며, 차이는 L1-Norm으로 계산한다. 본 논문에서는 제안하는 네트워크를 멀티 태스크 네트워크로 간주하고 적절한 가중치 값 λ_S와λ_T을 구하기 위해서 GradNorm[9]을 적용한다.

2-3. Attention Module

본 논문에서는 Channel Attention과 2개의 Spatial Attention을 적용하고 비교한다. 그림1과 같이 Attention 모듈은 Residual Block에서 Convolution layer 다음에 위치하며 입력 특징 맵 F_{in}으로부터 채널 또는 공간이 강조된 특징 맵 F_{out}을 출력한다.

2-3-1. Channel Attention

Channel Attention은 특징 맵에서 채널 간의 관계를 이용하여 특정 채널을 강조하며 구조는 그림 2와 같다.

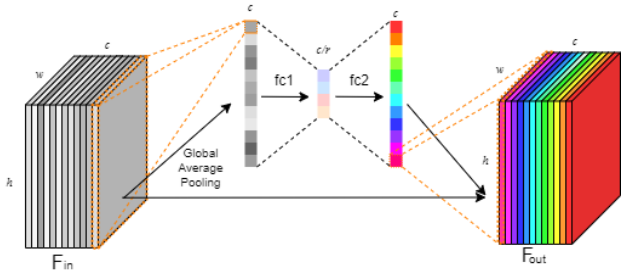

그림 2. Channel Attention Module

그림 2의 Channel Attention을 식으로 표현하면 다음 수식4, 5와 같다.

$$F_{out} = CA(F_{in}) \quad (4)$$

$$F_{out} = \sigma_2(fc2(\sigma_1(fc1(GAP(F_{in}))))) \cdot F_{in} \quad (5)$$

입력 특징 맵 F_{in}은 GlobalAveragePooling(GAP)으로 채널 크기의 벡터가 되며 각 값은 해당하는 채널을 대표한다. 이후 벡터는 첫 번째 Fully Connected Layer(fc1)를 통해서 유의미한 정보를 갖는 벡터로 압축되고 활성화 함수σ_1(Leaky ReLU)를 통해 비선형성을 갖는다. 이후 두 번째 Fully Connected Layer(fc2)와 활성화 함수σ_2(Sigmoid)는 압축된 벡터를 0 부터 1사이의 값을 갖는 채널 크기의 강조된 벡터로 만든다. 이 벡터의 각 값들은 입력 특징 맵에서 해당하는 채널을 스케일링해서 최종적으로 채널이 강조된 특징 맵 F_{out}을 출력한다.

2-3-2. Single Layer Spatial Attention

Spatial Attention은 특징 맵의 공간적인 특징을 강조하며 단일 층을 이용한 Spatial Attention은 다음 그림3과 같다.

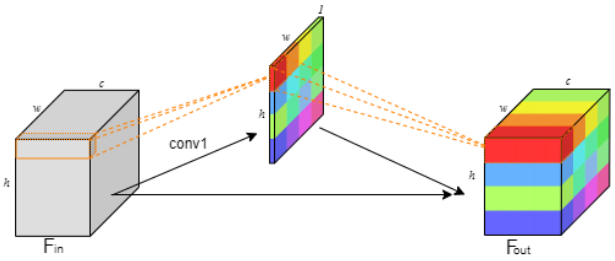

그림 3. Single Layer Spatial Attention Module

단일 층을 이용한 Spatial Attention은 입력 특징 맵 F_{in}에 1x1의 커널 사이즈를 갖는 Convolution layer를 적용해서 픽셀 별로 채널들의 정보를 취합하고 1개 채널의 강조된 특징 맵을 만든다. 이후 강조된 특징 맵으로 F_{out}을 스케일링하고 출력한다. 이를 수식으로 표현하면 다음 수식 6과 같다.

$$F_{out} = SA_{single}(F_{in}) = \sigma(conv_{1\times1}(F_{in})) \cdot F_{in} \quad (6)$$

$conv_{1\times1}$는 1x1커널을 갖는 Convolution Layer이며 σ는 강조 맵이 0부터 1사이의 값을 갖도록 만드는 Sigmoid 활성화 함수이다.

2-3-3. Multi Layer Spatial Attention

다중 층 Spatial Attention은 4개의 Convolution Layer를 사용해서 채널 정보 이외에 인접한 주변 픽셀들도 고려해서 강조된 특징 맵을 만든다.

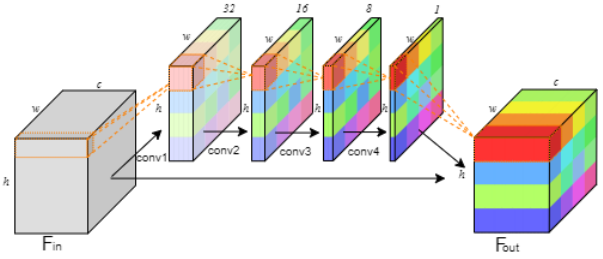

그림 4. Multi Layer Spatial Attention Module

인접한 주변 픽셀들을 수용하기 위해서 4개의 Convolution Layer를 사용하며, 각 Convolution Layer의 커널의 크기는

conv1부터 conv4까지 차례대로 7x7, 5x5, 3x3이며 마지막은 1x1이다. 또한 정보를 압축하기 위해서 각 Convolution Layer의 출력 채널의 크기는 점진적으로 32부터 16, 8을 거쳐 1로 줄어들며 이에 대한 수식은 다음 수식 7, 8과 같다.

$$F_{out} = SA_{multi}(F_{in}) \quad (7)$$

$$F_{out} = \sigma_4(conv_{1\times1}(\sigma_3(conv_{3\times3}(..F_{in})..)\cdot F_{in} \quad (8)$$

σ_4는 출력 특징 맵을 0부터 1사이의 값을 갖도록 하는 Sigmoid 활성화 함수이며 이를 제외한 σ_1부터σ_3은 Leaky ReLU활성화 함수이다.

3. 실험 방법

본 논문에서 제안하는 네트워크 및 Attention 모듈을 학습시키기 위해서 800장의 DIV2K셋과 SATV[10]를 이용해서 T_{HR}과 S_{HR}을 만들고 Adam을 사용했다. 또한 네트워크는 Tensorflow를 기반으로 NVIDIA V100을 이용해서 학습했다. 제안하는 네트워크의 성능을 검증하기 위해서 단일 영상 기반 초고해상도 기반 연구에서 널리 사용되는 Set5, Set14, BSD100, Urban100, Manga109로 성능을 평가하고 비교했으며 성능 평가 방법은 PSNR을 사용한다.

4. 실험 결과 및 분석

본 논문에서는 Attention 모듈의 성능을 평가하기 위해서 Attention 모듈을 사용하지 않은 모델을 기준으로 두고 3개의 Attention 모듈을 추가한 모델과 비교하였으며 검증 데이터 셋에 대한 실험 결과는 아래 표2와 같다. 가장 높은 PSNR은 **굵게** 표시하고 두 번째로 높은 PSNR은 밑줄로 표시한다.

표 2. Base 및 Attention 모듈 실험 결과(dB)

	Set5	Set14	BSD 100	Urban 100	Manga 109
Base	37.87	33.52	32.16	32.08	38.21
Base+CA	37.86	33.57	32.13	32.06	38.33
BASE+ Single SA	37.86	33.51	32.15	32.00	38.16
BASE+ Multi SA	**37.89**	**33.56**	**32.17**	**32.14**	**38.28**

표 2에서 Channel Attention을 추가한 모델의 경우 Set14와 Manga109에서 Base 모델보다 PSNR이 향상되었으나 나머지 데이터셋 Set5, BSD100과 Urban100에서는 오히려 감소한 것을 알 수 있다.이 실험 결과로 미루어보아 데이터 셋마다 성능 향상을 위해 강조되어야 하는 채널이 다르거나 혹은 잘못된 채널이 강조되어서 성능이 오히려 감소한 것으로 보인다. 단일 층 Spatial Attention은 모든 데이터 셋에서 Base모델 보다 낮은 성능을 보였으나 그 차이가 매우 작았다. 1 x1 커널의 Convolution으로 찾은 공간 특징 맵이 주변의 인접 픽셀을 고려하지 않고 채널 정보만 고려하고 한 개의 층만 사용해서 비선형성을 추가하지 못하고 한번에 한 개의 채널로 압축했기 때문에 성능이 감소한 것으로 보인다. 반면에 다중 Spatial Attention은 실험한 모든 모델 중에 가장 좋은 성능을 보였다. 이는 단일 층 Spatial Attention과 달리 다양한 크기의 커널 사이즈를 이용해서 인접 픽셀의 정보를 고려하고 또한 점진적으로 채널의 개수를 줄이는 동시에 3개의 비선형 활성화함수를 사용해서 특정 데이터 셋이 아니라 실험한 모든 데이터 셋에서 성능 향상시킬 수 있는 범용성을 갖는 강조된 공간 특징 맵을 찾은 것으로 보인다.

5. 결론 및 향후 연구

본 논문에서는 딥러닝을 이용한 단일 영상 기반 초고해상도 복원모델에 대해 다양한 Attention 모듈을 실험하고 성능을 평가 및 분석하였다. 그 결과 적절한 Attention 모듈을 설계한다면 성능을 향상 시킬 수 있는 것을 확인하였다. 향후 연구로는 Channel Attention에서 채널을 적절하게 표현있는 방법과 Channel Attention과 Spatial Attention을 동시에 설계하는 연구를 진행할 예정이다.

감사의 글

이 논문은 2016 년도 정부(교육부)의 재원으로 연구재단 기본 연구 지원사업의 지원으로 수행된 연구임. (NRF- 2016R1D1A1B04932889)

참고문헌

[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 CVPR. 2016.
[3]Lim, Bee, et al. "Enhanced deep residual networks for single image super-resolution." Proceedings of the IEEE conference on CVPR workshops. 2017.
[4]Hu, Jie, Li Shen, and Gang Sun. "Squeeze and excitation networks." Proceedings of the IEEE Conference on CVPR. 2018.
[5]Chen, Long, et al. "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning." Proceedings of the IEEE conference on CVPR. 2017.
[6]Zhang, Yulun, et al. "Image super-resolution using very deep residual channel attention networks." Proceedings of the ECCV. 2018.
[7]Dai, Tak, et al. “Second-order attention network of single image super-resolution,“ *Proceedings of the IEEE Conference on CVPR*. 2019.
[8]Shi, Wenzhe, et al. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network." Proceedings of the IEEE conference on CVPR. 2016.
[9]Chen, Zhao, et al. "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks." Proceedings of ICML. 2018.
[10]Song, Jinjoo, et al. "Structure adaptive total variation minimization-based image decomposition. " IEEE Transactions on Circuits and Systems for Video Technology 28.9 (2017): 2164-2176.