컨볼루션 신경망을 이용한 고효율 비디오 부호화에서의 인-루프 필터

박운성, *김문철
한국과학기술원, *한국과학기술원
pys5309@kaist.ac.kr, *mkim@ee.kaist.ac.kr

CNN (Convolutional Neural Network) based in-loop filter in HEVC

Woonsung Park *Munchurl Kim
KAIST *KAIST

요 약

본 논문에서는 고효율 비디오 부호화에서 채택하고 있는 인-루프 필터 중 SAO (sample adaptive offset)를 컨볼루션 신경망으로 대체하여 부호화 효율을 향상시키는 방법을 제안한다. SAO 는 양자화 에러를 줄이기 위해 인코더에서 디코더로 적절한 오프셋 값을 전송한다. 제안하는 컨볼루션 신경망을 사용한 인-루프 필터는 인코더와 디코더가 같은 컨볼루션 신경망을 사용하여, 추가적인 비트를 디코더로 전송할 필요 없이 양자화 에러를 줄일 수 있다. 컨볼루션 신경망의 구조는 두 가지를 각각 사용하였고, 각 컨볼루션 신경망의 구조에 대해서 입력 영상과 원래 영상의 평균제곱오차에 따라 다른 모델을 적용하였다. 따라서 제안하는 방법을 HEVC 에 적용하여 기존의 방법보다 더 적은 bit 로 더 좋은 화질의 영상을 얻어서 BD-rate 의 gain 을 얻을 수 있을 뿐만 아니라, 주관적인 화질의 비교에서도 더 좋은 결과를 보인다.

1. 서론

최근 고효율 비디오 부호화 (High Efficiency Video Coding, HEVC) [1]는 고급 비디오 부호화 (Advanced Video Coding, H.264/AVC) [2]와 같은 시각 품질을 유지하면서, 약 50%의 비트율 감소를 이뤄냈다.

고효율 비디오 부호화는 blocking artifacts, ringing artifacts, blurring artifacts 와 같은 압축으로 인한 결함을 줄이기 위해 2 가지 인-루프 필터링 기술을 사용한다. 첫 번째는 디블록킹 필터이고, 두 번째는 sample adaptive offset (SAO)이다.

SAO 는 샘플 왜곡을 줄이기 위해 먼저 복원된 샘플을 서로 다른 카테고리로 분류하고, 각 카테고리에 대한 오프셋을 얻은 뒤에 각 샘플에 더해진다. 각 카테고리의 오프셋은 인코더에서 계산되어 디코더로 전송된다. 그래서 SAO 는 평균적으로 3.5%의 BD-rate 감소를 이뤄냈다 [3]. 고효율 비디오 부호화 standard 가 발전하는 동안에 적응적 루프 필터(Adaptive loop filter, ALF)가 추가적인 부호화 효율 향상을 위해 제안되었다 [4]. ALF 는 필터 계수를 업데이트하고, 디코더로 그 정보를 전송한다. 두 가지 인-루프 필터 모두 디코더에 추가적인 비트를 보내주어서 양자화 에러를 줄인다는 특징이 있다.

최근에 컨볼루션 신경망(Convolutional Neural Network, CNN)은 영상 초해상화(image super-resolution, SR) [5,8], 영상 잡음제거 (image denoising) [6], 영상 분류 (image classification) [7]와 같은 분야에서 많은 연구가 이루어지고 있다. CNN 은 이러한 영상 처리나 영상 분류 문제에서 상당한 성능을 보여주고 있다. Dong *et al.*은 SRCNN [5]이라고 불리는 영상 초해상화 문제를 위한 CNN 구조를 제안하였다. SRCNN 은 저해상도의 영상에서 특징 맵을 추출하고 고해상도의 영상으로 매핑시킨다. SRCNN 은 비선형 매핑을 통해 저해상도 영상으로부터 잃어버린 고주파수 성분들을 복구하여 고해상도 영상을 만들어낸다. 최근에는 SRCNN 모델에서 확장하여, 더 깊은 층을 가진 CNN 구조에 대하여 영상 초해상화에 적용한 VDSR [8] 모델도 있다. VDSR 은 총 20 개의 층으로 되어 있고, 첫 번째와 마지막 층을 제외한 나머지의 층은 모두 같은 구조의 필터로 되어 있다. 각 64 개의 필터는 크기가 3×3×64 인 필터이다. 상당히 깊은 CNN 구조덕분에 VDSR 은 SRCNN 보다 PSNR 이 평균적으로 0.4dB 에서 1.2dB 까지 높게 나타났다.

최근에 Dong *et al.*은 AR-CNN (artifact reduction CNN)이라고 불리는 JPEG 압축 artifacts 를 줄이기 위해 CNN 을 사용했다 [9]. 이를 위해 SRCNN 구조와 비슷한 구조를 사용하였고, 훈련 과정에서 가중치를 초기화하기 위해 transfer learning 이라는 기법을 사용하였다. Transfer learning 을 통해 고화질의 압축 모델 (더 쉬운 모델)을 훈련해서 얻은 특징을 저화질의 압축 모델 (더 어려운 모델)에 적용하여 더 빠른 수렴을 하도록 하였다.

본 논문에서는 영상처리에 CNN 을 적용한 선행연구를 참고하여, HEVC 에서 인-루프 필터에 CNN 을 적용하는 방법을 제안한다. 구체적으로 인-루프 필터에 사용되는 CNN 모델을 훈련하는 방법을 제시하고, 얻어진 CNN 모델을 인-루프 필터에 적용하는 방법도 제안한다.

본 논문의 구성은 다음과 같다. 2 절에서는 본 논문에서 제안하는 인-루프 필터에 대해 살펴본 후, 3 절에서는 제안한

인-루프 필터를 HEVC 에 적용하여 얻은 결과에 대해 토의한다. 4 절에서는 본 논문에 대한 결론을 맺는다.

2. 제안하는 인-루프 필터

그림 1 에서 제안하는 CNN 기반의 인-루프 필터를 포함한 비디오 인코더의 구조를 보여준다. 기존의 HEVC 인코더는 두 번째 인-루프 필터로 SAO 를 사용한다. 즉, 본 논문에서는 SAO 를 대신하여 CNN 기반의 인-루프 필터를 사용함으로써 디코더에 추가적인 비트를 전송하지 않고 복원된 영상의 화질을 향상시킨다. CNN 에 사용되는 가중치는 오프라인으로 훈련되고, 훈련된 가중치는 인코더와 디코더에 같은 값으로 사용된다.

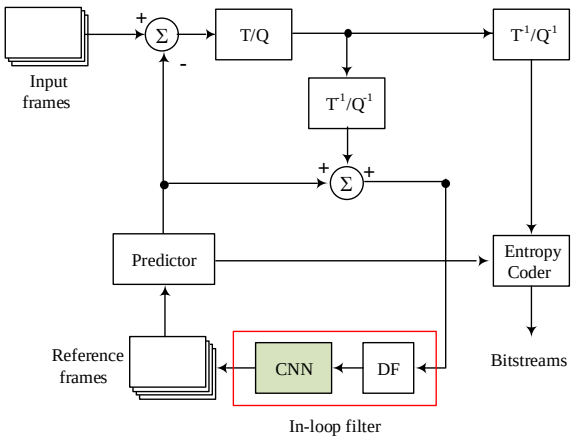
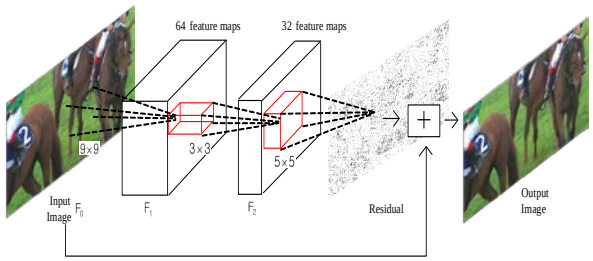


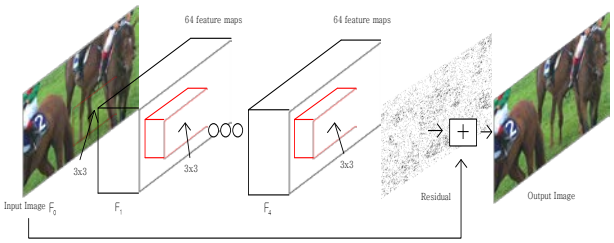
그림 1. 인코더에서의 CNN 기반의 인-루프 필터

제안하는 CNN 기반의 인-루프 필터의 구조는 SRCNN 구조를 기반으로 한다. SRCNN 구조는 상당히 많은 수의 매개변수와 비선형 활성화 함수 [10]의 사용덕분에 저화질의 영상에서 고화질의 영상으로 매핑하는 능력이 뛰어나다. 이는 압축으로 인해 잡음이 생긴 영상을 저화질의 영상이라고 생각하고, 원래 압축하기 전의 영상을 고화질의 영상이라고 생각한다면, 위와 같은 방법으로 인-루프 필터에서도 위의 구조를 적용할 수 있다.

그림 2 는 제안하는 인-루프 필터에 사용되는 두 가지 CNN 의 구조를 보여준다. CNN 구조 1 은 SRCNN 을 기반으로 구성되었다. SRCNN [5]은 저해상도 영상을 입력 영상으로 하여 고해상도 영상을 출력 하도록 훈련을 시킨 반면에, 제안하는 인-루프 필터에 사용되는 CNN 은 디블록킹 필터를 통과한 복원된 영상을 입력으로 하고, 원래의 압축되기 전의 영상과 CNN 의 입력 영상과의 차이 영상을 출력 영상으로 사용하여 CNN 파라미터를 학습 시킨다. 출력 영상으로 잔차 영상을 사용하는 이유는 학습과정에서 CNN 의 가중치가 더 빨리 수렴되도록 도와주기 때문이다 [8]. CNN 구조 2 는 VDSR [8]을 바탕으로 만들어진 구조이다. 입력 영상과 출력 영상의 정보는 CNN 구조 1 과 같다. 따라서 두 가지 CNN 구조에 대해서 최종적으로 인코더나 디코더에서 결과 영상을 얻으려면 학습을 통해 얻은 CNN 모델의 출력 영상(잔차 영상)과 CNN 의 입력 영상을 더해주면 된다. 이 때 CNN 모델은 입력 영상의 각 블록마다 적용되어 하나의 영상으로 합쳐지게 된다. 이번 실험에서 사용한 블록 사이즈는 128 로 하였다.



(1) 인-루프 필터에 사용되는 CNN 구조 1



(2) 인-루프 필터에 사용되는 CNN 구조 2

그림 2. 인-루프 필터에 사용되는 두 가지 CNN 구조

제안하는 인-루프 필터에 사용되는 CNN 의 가중치를 훈련시키기 위해 주어진 훈련 데이터에 대해서 CNN 모델에 의해 예측된 출력 영상과 원래 영상과의 평균제곱오차 (mean square error, MSE)를 최소화하는 방향으로 학습을 한다. 학습 방법은 일반적으로 CNN 의 학습으로 많이 사용되는 Stochastic Gradient Descent [11]의 방법으로 진행된다.

일반적으로 비디오 부호화에서 높은 양자화 매개변수 (quantization parameter, QP)는 많은 양자화 에러를 포함한 복원 영상을 만들어낸다. 반대로 낮은 QP 의 경우 적은 양자화 에러를 포함한 복원 영상을 만들어낸다. 따라서 CNN 모델이 넓은 범위의 QP 에 대해서 적용되려면 상당히 깊은 구조의 CNN 이 필요하다. 그러나 깊은 구조의 CNN 모델은 메모리뿐만 아니라 실행 시간에도 상당한 복잡도가 발생한다. 본 논문에서는 이러한 복잡도를 줄이기 위해 여러 CNN 모델을 사용하여 넓은 범위의 QP 에 대해서도 HEVC 의 인-루프 필터에 적용할 수 있도록 한다. CNN 모델을 나누는 기준은 QP 도 가능하지만, 같은 QP 에서도 영상의 특징에 따라 양자화 에러가 적은 부분과 많은 부분이 모두 존재할 수 있다. 따라서 본 논문에서는 양자화 에러가 포함된 복원된 영상과 원래의 압축되기 전의 영상과의 MSE 값을 기준으로 CNN 모델을 나눈다. 구체적으로 훈련 데이터로 총 9 개의 비디오 영상에 대해서 QP=18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42 의 복원된 영상을 만든 후에 32×32 블록으로 나누어서 MSE 값을 계산하여 총 4 개의 구간으로 나누어서 각 구간에 대해서 각 CNN 모델을 훈련시킨다. MSE 를 4 개의 구간으로 나눈 기준은 전체 MSE 분포에 대해서 25%, 50%, 75%의 위치를 기준으로 정한다.

제안하는 CNN 기반의 인-루프 필터의 효율성을 검증하기 위해 HEVC 검증 모델 (HM)의 All intra 모드에 대해서 실험을 한다. 그리고 간단한 연산을 위해 YUV 테스트 비디오 영상 중에서 Y 성분에만 적용한다. 성능 비교는 기존의 HM 16.0 과 SAO 를 CNN 모델로 대체하여 만든 HM 모델과의 비교로 이루어진다.

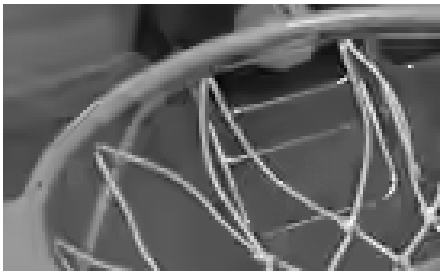
3. 실험 결과

본 논문에서 제안하는 방법에 대한 실험은 All intra 모드에 대해서 총 9 개의 비디오 영상으로 진행을 한다. 각 비디오 영상은 BasketballDrill (BD), BQMall (BQM), PartyScence (PS), BasketballDrillText (BDT), BasketballPass (BP), BQSquare (BQS), Blowingbubbles (B) 그리고 RaceHorses (RH)를 사용한다. 테스트에 사용된 영상은 훈련 데이터와 겹치지 않는 100 번째 프레임부터 199 번째 프레임으로 총 100 프레임을 사용한다. 실험 결과는 기존의 SAO 를 인-루프 필터로 사용하는 HM 16.0 의 결과를 기준으로 본 논문에서 제안하는 CNN 을 기반으로 하는 인-루프 필터를 사용한 HM 16.0 의 결과가 BD-rate 측면에서 얼마나 이득이 있었는지에 대해 퍼센트로 표시하였다.

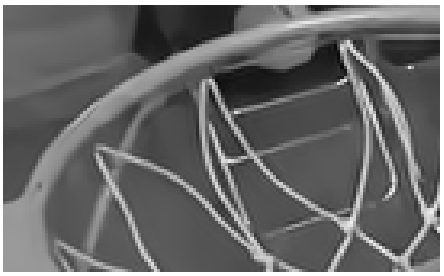
표 1. 제안하는 CNN 기반의 인-루프 필터와 SAO 와의 BD-rate 측면에서의 성능 비교

크기	영상 종류	All Intra 1	All Intra 2
		BDBR (%)	BDBR (%)
832×480	BD	-8.5	-9.0
	BQM	-3.0	-3.2
	PS	-1.3	-0.6
	BDT	-7.1	-7.5
	RH	-1.6	-1.5
416×240	BP	-2.8	-3.1
	BQS	-2.2	-1.3
	B	-1.5	-1.1
	RH	-2.4	-2.6
	평균	-3.37	-3.32

표 1 에 따르면 CNN 모델 1 을 사용한 인-루프 필터 (All Intra 1)는 SAO 에 비해 평균적으로 -3.37%의 BD-rate 이득이 있다. CNN 모델 2 를 사용한 인-루프 필터 (All Intra 2)는 SAO 에 비해 평균적으로 -3.32%의 BD-rate 이득이 있다. 그림 3 은 제안하는 방법과 기존의 HEVC 로 얻어진 영상의 Y 성분에 대한 주관적 화질의 차이를 비교한 결과이다.



(1) SAO, BasketballDrill, QP37



(2) CNN, BasketballDrill, QP37

(3) SAO, QP37, RH

(4) CNN, QP37, RH

그림 3. CNN 과 SAO 를 이용한 인-루프 필터의 결과 영상의 Y 성분에 대한 주관적 화질 비교

그림 3 에서 볼 수 있듯이, 인-루프 필터에 의해 복구된 영상은 CNN 을 이용한 방법이 에지 부분에서 더 선명한 모습을 보여주므로 주관적 화질 측면에서 더 좋은 결과를 보여준다.

4. 결론

본 논문에서는 HEVC 에서의 CNN 기반의 인-루프 필터를 제안한다. 제안하는 방법은 압축으로 인해 양자화 에러가 포함된 복구된 영상으로부터 원본 영상에 더 가깝게 매핑하는 필터를 SAO 라는 필터 대신에 CNN 모델을 사용한다. CNN 모델을 사용할 경우 디코더에 추가적으로 보내주는 비트가 없고, SAO 보다 원본 영상에 더 가깝게 만들어주기 때문에 BD-rate 측면에서뿐만 아니라 주관적 화질 측면에서도 더 좋은 성능을 보인다. MSE 에 따라 4 개의 CNN 모델을 학습하여 더 간단한 구조로 넓은 범위의 QP 에 제안하는 방법을 적용할 수 있도록 한다.

현재 낮은 QP 에 대해서 높은 QP 보다 낮은 성능을 보이는데, 낮은 QP 에 대한 CNN 모델을 잘 학습할 필요가 있다. 그리고 추후에 lowdelay P 모드나 random-access 모드에 대해서도 실험을 하여 인-루프 필터의 성능에 대해서 고찰할 것이다.

Acknowledgement

본 논문 연구는 연구재단 중견연구자사업 핵심연구(개인) 과제 (과제번호: 2014R1A2A2A01006642)로 수행되었습니다.

5. 참조

[1] G. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of High efficiency video coding (HEVC),” *IEEE Tr. Cir. Sys. for Video Tech.*, vol. 22, no. 12, Dec 2012.

[2] Joint Video Team (JVT) of ITU-T VCEG and ISO/IEC MPEG, “Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification,” *ITU-T Rec. H.264 and ISO/IEC 14496-10 AVC*, May 2003.

[3] C. M. Fu, E. Alshina, A. Alshin, Y. W. Huang, C. Y. Chen, C.

Y. Tsai, C. W. Hsu, S. M. Lei, J. H. Park and W. J. Han, “Sample adaptive offset in the HEVC standard,” *IEEE Tr. Cir. Sys. for Video Tech.*, vol. 22, no. 12, pp. 1755-1764, 2012.

[4] C.-Y. Tsai, C.-Y. Chen, T. Yamakage, I. S. Chong, Y.-W. Huang, C.-M. Fu, T. Itoh, T Watanabe, T. Chujoh, M. Karczewicz, and S.-M. Lei, “Adaptive loop filtering for video coding,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 6, pp. 934-945, 2013.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” ECCV 2014. Springer International Publishing, pp. 184-199, 2014.

[6] V. Jain and S. Seung, “Natural image denoising with convolutional networks,” Advances in Neural Information Processing Systems, 2009.

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *Int. Conf. Learning Representations*, 2014.

[8] J. Kim, J. K. Lee, and K. M. Lee, “Accurate Image Super-Resolution Using Very Deep Convolutional Networks,” arXiv preprint arXiv:1511.04587, 2015.

[9] C. Dong, Y. Deng, C. C. Loy and X. Tang, “Compression artifacts reduction by a deep convolutional network,” *IEEE Int. Conf. on Computer Vision*, pp. 576- 584, 2015.

[10] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” *Proc. 27th Int. Conf. on Machine Learning*, pp. 807-814, 2010.

[11] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner , “Gradient-based learning applied to document recognition,” *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.