트랜스포머 기반 자기 참조 인루프 필터링

이정경, 김나영, 강제원
이화여자대학교 전자전기공학과, 스마트팩토리융합전공
jungkyong1204@gmail.com, 12skdud21@gmail.com, jewonk@ewha.ac.kr

Transformer-based Self-Referential In-loop Filtering

Jung-Kyung Lee[1], Nayoung Kim[1], Je-Won Kang[1,2]
[1)] Department of Electronic and Electrical Engineering, Ewha Womans University
[2)] Smart Factory Multidisciplinary Program, Ewha Womans University

요약

다양한 미디어 서비스의 발전으로 비디오의 방대한 데이터를 효과적으로 압축할 수 있는 비디오 부호화 표준은 지속적인 발전을 하고 있다. 압축된 데이터를 다시 영상으로 복원하는 비디오 부복호화 과정에서 영상 데이터의 손실이 일어나고 그에 따른 다양한 형태의 열화가 나타나 영상의 화질을 저하한다. 이러한 열화들을 제거하여 원본 이미지에 가깝게 만들기 위해서 인루프 필터 과정을 비디오 부호화 표준에서 포함하고 있다. 이에 최근 영상처리 및 컴퓨터 비전 분야에서는 널리 사용되는 인공 신경망을 적용하여 효과적인 필터링을 하는 방법을 제시한다.

본 논문에서는 비디오 부호화 시 인루프 필터링에서 자기 참조를 통한 화질 개선 방법에 대해 연구하였다. 이를 위하여 트랜스포머 기반의 화질 개선 네트워크를 제안하고 기존 부호화 방법과 비교하였다. 인루프 필터링을 통해 화질을 향상하여 주관적 화질을 개선할 뿐만 아니라 객관적 부호화 효율을 증가시키는 방법을 개발하였다.

1. 서론

최근 인터넷 스트리밍 서비스나 소셜 미디어에서는 초해상도(UHD) 화질의 방대한 영상 데이터가 사용되고 있다. 이를 효과적으로 저장 및 전송하기 위해, 새로운 비디오 코딩 표준인 VVC(Versatile Video Coding)[1]에서는 HEVC(High Efficiency Video Coding)[2]의 기술을 기반으로 높은 부호화 효율을 제공하는 다양한 기술들을 채택하였다. 압축 비디오 품질을 높이고 비트 전송률을 줄이기 위해 인루프 필터(In-Loop Filter)에서 기존 HEVC에서 확장된 기술이나 새로운 알고리즘이 개발되었다. VVC에서는 비디오 부호화 시 생성되는 양자화 노이즈 혹은 블록형 아티팩트를 생성을 최대한 줄이고자 기존의 적응적 샘플 오프셋(SAO, Sample Adaptive Offset), 디블록킹 필터(DF, Deblocking Filter)와 새롭게 추가된 기술인 적응적 루프 필터(ALF, Adaptive Loop Filter), 크로마 스케일링을 사용한 루마 매핑(LMCS, Luma Mapping with Chroma Scaling) 필터를 도입하였다.

인공 신경망이 영상처리 및 컴퓨터 비전 분야에서는 효과적으로 사용되었기 때문에 비디오 압축 분야에서도 영상의 화질을 개선하기 위해 여러 인공 신경망 기반 인루프 필터가 제안되었다. 초기에는 필터 크기와 및 네트워크의 깊이를 조정하는 연구가 진행되었다. 가변 필터 크기와 잔차 이미지를 활용한 방법 [3,4] 이나 디코딩 복잡성을 줄이기 위해 루마와 크로마가 모두 공유하는 컨볼루션 3개의 층으로 구성된 네트워크가 제안되었다[5]. 이 외에도 레임 간의 시간적 상관관계를 활용하기 위해 다른 프레임을 참조하여 화질을 향상하는 방법이 개발되었다[5].

본 논문에서는 VVC의 인루프 필터링에서 자기 참조를 통한 화질 개선 방법을 제안한다. 먼저 트랜스포머[6,7] 기반의 화질 개선 네트워크를 설명하고 All-Intra 코딩 시나리오에서 기존 부호화 방법 대비 제안 방법의 성능 향상을 실험적으로 보인다.

2. 제안 방법

본 논문에서는 인루프 필터링 과정에서 현재 프레임을 참조를 하여 현재 프레임의 화질 개선을 하는 방법을 제안한다. 먼저, 특징 추출 네트워크를 통해 현재 프레임의 특징을 추출한다. 추출한 특징들을 유사도를 구해 유사도가 반영된 값을 가중합하여 구한다. 유사도를 구하기 위해 트랜스포머 기반 네트워크를 사용하여 특징에 대한 공간적 어텐션을 구하게 된다. 현재 프레임 내 다른 공간적 위치의 특징 정보를 활용하여 현재 프레임 내의 정보를 이용하는 자기 참조 구조가 제안 되었다.

그림 1은 제안하는 자기 참조 네트워크의 구조를 보인다. 인루프 필터 위치에서 복원된 현재 프레임(I_t)이 입력으로 사용되고 특징 추출 네트워크, 트랜스포머 기반 네트워크, 컨볼루션 레이어를 통과하여 출력 프레임 $\hat{I}_t$을 생성하는 것을 볼 수 있다.

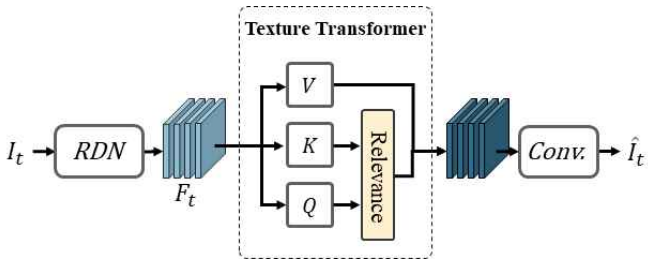

그림 1. 제안 네트워크 구조

A. 특징 추출 네트워크

특징 추출 네트워크를 통해서 현재 프레임의 위치별 텍스처 정보 및 특징을 구한다. residual dense network(RDN)[8] 구조를 사용하며 모든 컨볼루션 레이어는 64의 채널 사이즈와 3x3 크기의 커널을 가진다.

B. 트랜스포머 기반 네트워크

트랜스포머 구조의 기본적인 요소에는 키(Key), 쿼리(Query), 값(Value)가 있다[6]. 각각을 K, Q, V으로 나타내고 다음 식과 같이 현재 프레임(I_t) 을 특징 추출 네트워크인 RDN을 통해 추출된 특징맵으로 설정한다.

$$K, Q, V = RDN(I_t)$$

키, 쿼리, 값 모두 동일한 벡터의 집합에서 나온 벡터를 의미하기 때문에 셀프 어텐션 구조를 갖고 있다고 볼 수 있다.

추출된 특징맵의 서로 다른 위치에서 얻게 되는 벡터 간의 유사도를 통해 현재 위치와 가장 유사한 텍스쳐 및 픽셀 정보를 생성한다. 현재 프레임 내 다른 위치의 특징 정보를 활용하는 자기 참조 구조를 형성하게 된다. 유사도 $r_{i,j}$는 다음 식과 같이 Q 에 속한 특징 패치 q_i와 K 에 속한 특징 패치 k_j 간의 정규화된 내적을 통해 구한다.

$$r_{i,j} = \left\langle \frac{q_i}{\| q_i \|}, \frac{k_j}{\| k_j \|} \right\rangle$$

위 식에서 구한 유사도와 V 간의 가중치의 곱을 통해 최종 특징맵를 구한다. 이를 통해 다른 텍스처 정보를 현재 위치의 피쳐에 반영하여 텍스쳐를 더 정교화하게 되고 부호화의 프레임 열화 문제를 최소화할 수 있게 된다. 최종 특징맵은 입력 프레임과 크기와 차원을 맞추기 위해 컨볼루션 레이어를 통과하여 업샘플링을 수행하고 최종 출력 프레임을 구한다.

3. 실험 결과

본 논문에서 제안하는 자기 참조 인루프 필터의 성능을 평가하기 위하여 VVC 참조 소프트웨어인 VTM 10.0 [9]에서 공통 실험 조건 CTC (Common Test Condition) [10]을 참고하여, 제안하는 방법의 성능을 비교하였다. 실험은 All-Intra(AI) 코딩 시나리오를 사용하여 QP 22, 27, 32, 37에서 BD-rate를 측정하였다. CTC 조건에 따라 프레임 샘플링을 8로 설정하고, I-프레임 총 3장을 부호화하여 실험하였다. 부호화 시

2.10GHz Intel CPU와 4장의 RTX 1080 Ti GPU를 사용하였다.

실험 결과, 기존 부호화 방식 대비 깊이 자기 참조 인루프 필터링을 통한 부호화 방식은 평균적으로 약 2.3%의 BD-rate 감소 효율을 보였다. 클래스별 평균으로는 Class A1과 Class E에서는 각각 2.6%와 3.5%로 높은 부호화 성능 향상을 보였다. 특히 Class A1의 "FoodMarket4" 영상에서는 약 5.4%의 높은 성능 향상이 나타난다.

표1. 제안 방법의 BD-rate(%) 비교

Class	Video	BD-rate		
		Y	U	V
A1	Campfire	-0.5%	-0.1%	0.0%
	FoodMarket4	-5.4%	-0.2%	-0.4%
	Tango2	-1.8%	-0.3%	-0.4%
A2	CatRobot1	-2.3%	-0.1%	-0.1%
	DaylightRoad2	-1.7%	-0.2%	-0.6%
	ParkRunning3	-1.4%	-0.3%	-0.3%
B	BasketballDrive	-2.0%	-0.1%	-0.3%
	BQTerrace	-1.2%	0.0%	0.0%
	Cactus	-1.6%	-0.2%	0.1%
	MarketPlace	-1.4%	-0.2%	-0.1%
	RitualDance	-4.2%	-0.5%	-0.3%
C	BasketballDrill	-3.5%	-0.2%	-0.6%
	BQMall	-2.4%	-0.2%	-0.1%
	PartyScene	-1.0%	0.0%	0.0%
	RaceHorses	-1.0%	-0.1%	0.1%
D	BasketballPass	-3.0%	-0.1%	-0.2%
	BlowingBubbles	-1.7%	-0.2%	-0.1%
	BQSquare	-1.5%	0.1%	0.1%
	RaceHorses	-2.1%	-0.1%	-0.4%
E	FourPeople	-3.7%	-0.1%	-0.2%
	Johnny	-3.4%	-0.3%	-0.7%
	KristenAndSara	-3.3%	-0.4%	-0.3%
Average		**-2.3%**	**-0.2%**	**-0.2%**

4. 결론

본 논문에서는 인루프 필터링에서 인공 신경망 기반 화질 개선방법에 대해 연구하였다. 자기 참조가 가능한 트랜스포머 기반의 화질 개선 네트워크를 제안하고 기존 부호화 방법과 비교하였다. 제안 방법의 실험 결과, 기존 부호화 기법 대비 약 2.3% 성능 향상을 보인다. 화면 내 예측 상황에서 자기 참조를 통해 화질을 개선하였다는 점에서 큰 의의를 갖는다.

5. ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1A2C4002052).

참조문헌

[1] Sullivan, Gary J., et al. "Overview of the high efficiency video coding (HEVC) standard." IEEE Transactions on circuits and systems for video technology 22.12 (2012): 1649-1668.
[2] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, “Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC),” Proceedings of the IEEE,2021
[3] S. Zhang, Z. Fan, N. Ling, and M. Jiang, “Recursive residual convolutional neural network-based in-loop filtering for intra frames,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 7, pp. 1888-1900, 2019.
[4] Y. Dai, D. Liu, and F. Wu, “A convolutional neural network approach for post-processing in HEVC intra coding,” in International Conference on Multimedia Modeling. Springer, 2017, pp. 28-39
[5] R. Yang, M. Xu, Z. Wang, and T. Li, “Multi-frame quality enhancement for compressed video,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6664-6673.
[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[7] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, “Learning texture transformer network for image super-resolution,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5791-5800
[8] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image restoration,” TPAMI, 2020
[9] Versatile Video Coding Test Model (VTM), 10.0 [Online],“https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware VTM/-/releases/VTM-10.0.”
[10] F. Bossen, J. Boyce, X. Li, V. Seregin, and K. Sühring, JVET Common Test Conditions and Software Reference Configurations for SDR Video, document JVET-T2010, ITU-T/ISO/IEC Joint Video Experts Team (JVET), Oct. 2020.
[11]“Convolutional neural network loop filter.” Recommendation ITU-T
SG 16 WP3 and ISO/IEC JTC 1/SC 29/WG 11, 2019.