Multi-task 수행을 위한 압축 심층신경망 기반 VCM

이해림, *이주영, 조승현
경남대학교, *한국전자통신연구원
haelimjigoo@gmail.com, *leejy1003@etri.re.kr, scho@kyungnam.ac.kr

VCM based on Compression Neural Network for Multi-task

Haelim Lee, *Jooyoung Lee, Seunghyun Cho
Kyungnam University, *ETRI

요 약

최근 기계 임무수행에 사용되는 데이터양이 증가함에 따라 기계를 위한 효율적인 영상 압축방식의 필요성이 높아졌다. 기존의 비디오 코덱은 HVS (Human Visual System) 특성을 고려한 기술이기 때문에 부호화 과정에서 기계 임무수행에 필요하지 않은 정보를 효과적으로 제거할 수 없다. 반면 심층신경망 기반 압축네트워크의 경우, 원본 영상으로부터 기계 임무수행에 필수적인 데이터만을 추출하여 부호화 하도록 학습할 수 있는 장점이 있다. 본 논문에서는 압축 심층신경망과 기계 임무수행 네트워크로 구성되는 VCM (Video Coding for Machine) 프레임워크를 제안하고 학습에 의한 압축효율 향상을 검증한다. 이를 위해 압축 심층신경망을 객체탐지 임무수행 네트워크와 함께 학습시킨 결과, VVC (Versatile Video Coding) 대비 평균 61.16%의 BD-rate 감소가 확인되었다. 뿐만 아니라, 학습된 압축 심층신경망은 객체분할 임무수행에서도 VVC 대비 평균 58.43%의 BD-rate 감소를 보여 다중 기계 임무의 효율적 수행이 가능함을 확인할 수 있었다.

1. 서론

최근 사물인터넷, 스마트시티, 자율주행 자동차 등의 응용 환경에서 다양한 영상 데이터가 수집되고 있으며, 기계에 의해 자동으로 영상을 인식하고 분석하는 기술들이 크게 늘고 있다 [1]. 인공지능의 발전과 함께 영상 데이터를 입력으로 하는 기계 임무수행 기술이 보편화되고 있는 것이다. 이러한 상황에서 영상 데이터의 전송 및 저장을 위해 기존의 비디오 코덱을 사용할 경우 압축률 대비 임무수행 정확도 측면에서 비효율적일 수 있으므로 기계 임무수행을 위한 효율적 비디오 부호화 방식의 연구에 대한 필요성이 제기되고 있다.

이에 부응하여 ISO/IEC JTC 1/SC 29/WG 2 - 이하, MPEG (Moving Picture Experts Group)은 VCM 기술의 표준화에 착수하여, 2021 년 1 월에 개최된 133 차 회의에서 CfE (Call for Evidence) [2]를 발행하고 가장 최근인 4 월 134 차 회의에서 CfE 응답에 대한 평가가 이루어졌다. 이 밖에도 134 차 회의에서는 평가체계 및 일부 평가 데이터셋의 조정과 이에 따른 VVC 앵커(Anchor) 실험결과의 변동사항이 업데이트되었으며 [3], 향후 추가적인 CfE 응답의 검토와 CfP (Call for Proposals) 발행에 대한 논의가 이루어졌다.

그간 MPEG VCM 표준화를 통해 논의되었거나 관련 연구를 통해 제시된 VCM 기술은 크게 분석 후 압축(Analyze then Compress) 방식과 압축 후 분석(Compress then Analyze) 방식으로 구분될 수 있다. 분석 후 압축 방식은 영상 부호화에 앞서 원본 영상에 대한 분석을 실시하고 그 결과를 일정 길이의 서술자(descriptor) 형태로 부호화 한다 [8]. 반면, 압축 후 분석 방식은 원본영상 또는 원본영상에서 추출한 특징(Feature)을 부호화 한다. 압축 후 분석 방식을 위해서는 기존의 표준 비디오 코덱뿐만 아니라 최근 활발한 연구를 통해 발전하고 있는 압축 심층신경망의 사용을 고려할 수 있다. MPEG VCM 표준화에서는 원본영상에 대한 VVC 코덱의 압축률 대비 임무수행 정확도를 기준으로 제안 기술들의 압축효율 평가를 진행하고 있다.

본 논문에서는 압축 후 분석 방식의 VCM 기술을 위해 원본영상을 압축 심층신경망을 통해 부호화 한다. 압축 심층신경망은 본 논문에서 제안하는 VCM 프레임워크 상에서 기계 임무수행 네트워크와 함께 학습되며 이를 통해 VVC 대비 대폭 향상된 압축효율을 달성할 수 있다. 또한, 이렇게 학습된 압축 심층신경망은 다른 기계 임무에서도 VVC 대비 압축효율이 향상되어 다중 임무수행이 가능한 VCM 기술의 실현 가능성을 보여준다. 본 논문의 실험결과는 134 차 MPEG 회의를 통해 기고된 바 있으며 [5], 본 논문에서는 제안 기술의 학습과정을 포함하여 보다 상세한 내용을 다룬다.

2. 배경기술

2-1. 영상 압축을 위한 심층신경망

심층신경망에 기반한 영상 압축 네트워크의 대략적인 구조를 그림 1 에 나타냈다 [6]. 부호화 (Encoder) 신경망은 입력 영상으로부터 특징을 추출해 은닉벡터 (Latent vector)를 생성하고, 추출된 은닉벡터에 대한 양자화 (Quantization)와 엔트로피 부호화 (Entropy encoding) 과정을 거쳐 비트스트림 (Bitstream)이 생성된다. 전송된 비트스트림에 대해 엔트로피 복호화 (Entropy decoding)와 역양자화 (Dequantization)과정을 거쳐 은닉벡터가 복원되며, 복호화 (Decoder) 신경망을 통해 은닉벡터로부터 영상이 복원된다.

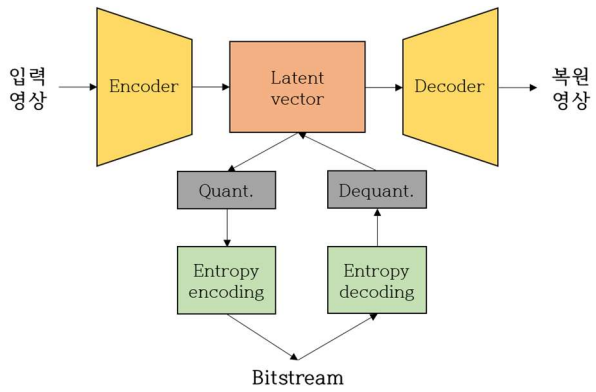

그림 1. 심층신경망 압축네트워크 구조

영상 압축 심층신경망에 대한 최근의 연구는 동일 비트율 (Bitrate)에서 VVC 와 유사한 PSNR (Peak Signal to Noise Ratio) 성능을 보이며 다양한 구현이 공개되고 있다. 본 논문에서는 Pytorch 기반의 CompressAI [7]에 구현된 mbt2018-mean 모델을 압축 네트워크로 사용하였다. 본 논문은 압축 심층신경망 자체의 성능이 아니라 학습에 의한 VCM 압축효율 향상을 목표로 하고 있기 때문에 가장 최신의 연구들에 비해 다소 낮은 압축효율을 보이지만 비교적 학습속도가 빠른 모델을 채택하였다.

2-2. MPEG VCM 의 기계 임무 및 평가 데이터셋

MPEG VCM 표준화 과정에서 압축성능을 평가하기 위한 기계 임무로 객체탐지 (Object detection), 객체분할 (Object segmentation), 객체추적 (Object tracking), 포즈추정 (Pose estimation), 행동인식(Action recognition)이 제안되었다 [3]. 임무 종류에 따른 평가 데이터셋과 임무수행 네트워크를 표 1 에 정리하였다. 본 논문에서는 MS COCO 학습 데이터셋으로 객체탐지 임무에 대해 임무수행 네트워크와 함께 압축 심층신경망의 학습을 진행하였으며, 마찬가지로 MS COCO 평가 데이터셋으로 객체탐지 및 객체분할 임무에 대한 압축효율 평가를 각각 수행하였다. 연구 진행 중에 해당 임무들에 대한 MPEG VCM 평가 데이터셋이 변경되어 이를 실험에 반영하지 못하였음을 밝힌다.

표 1. MPEG VCM 임무 종류, 평가 데이터셋 및 임무수행 네트워크

임무 종류	평가 데이터셋	임무수행 네트워크
Object detection	OpenImages-v6, FLIR Thermal Dataset, SFU-HW-Objects-v1, TVD	Faster R-CNN X 101-FPN
Object segmentation	OpenImages-v6	R50-FPN
Object tracking	HiEve-10	JDE-1088x608
Pose estimation	HiEve-10	HRNet
Action recognition	HiEve-10	Slowfast

3. 본론

3-1. 제안하는 VCM 프레임워크 구조

그림 2 에 나타낸 것과 같이, 본 논문에서 제안하는 VCM 프레임워크는 크게 송신부 (Transmitter)와 수신부 (Receiver)로 구성된다. 송신부는 이미지/비디오를 입력받아 공통의 비트스트림을 생성하는 부호화기이며, 수신부는 하나 또는 그 이상의 복호화기와 기계 임무수행기 (MVNet)로 구성된다. 그림 2 에서 압축 심층신경망 (CompNet)의 부호화기는 모든 기계 임무수행을 위해 공통적으로 사용되는 반면, 복호화기는 특정한 기계 임무 또는 사람 임무를 위해 사용되거나 여러 기계 임무 공통적으로 사용될 수 있다.

3-2. 압축 심층신경망의 학습

그림 2 에 나타낸 CompNet 과 MVNet 은 하나의 네트워크를 구성하는 부분 네트워크로 볼 수 있으며, End-to-end 방식으로 학습될 수 있도록 연결되었다. 즉, 그림 2 에 $\hat{x}$ 으로 표시한

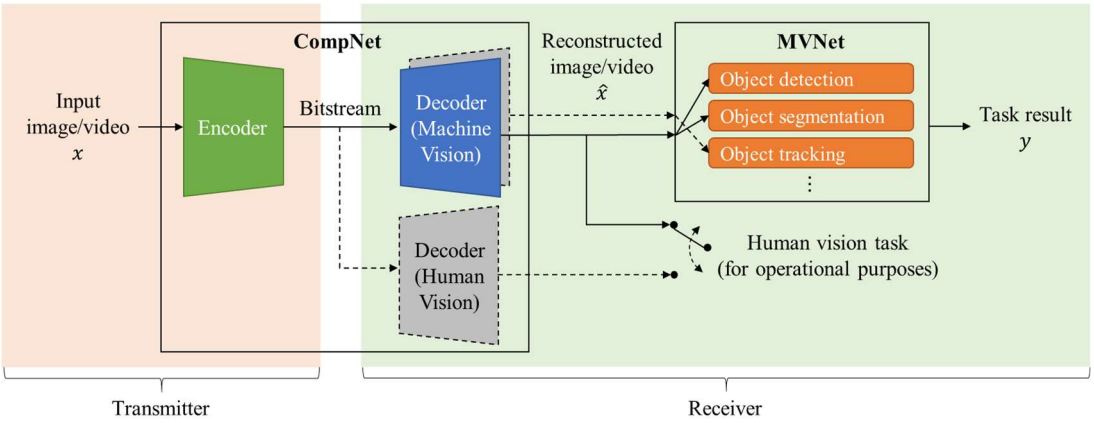

그림 2. 제안하는 VCM 프레임워크 구조

복원영상은 CompNet 의 복호화기에서 출력되어 MVNet 에 직접 입력된다. θ_e^{MV}와 θ_d^{MV}가 각각 CompNet 의 부호화기 및 복호화기 네크워크의 매개변수(Parameter)일 때, 제안된 VCM 프레임워크에서 CompNet 의 학습은 $\Theta^{MV} = \{\theta_e^{MV}, \theta_d^{MV}\}$인 최적의 Θ^{MV}를 찾는 과정이며, 아래의 수식 (1)과 같이 나타낼 수 있다.

$$\Theta^{MV} = arg \min_{\theta_e, \theta_d} \sum_{x_i \in X,\, y_i \in Y} \{(1-\lambda_{RA}) \cdot L_R(x_i) + \lambda_{RA} \cdot L_A\left(\Psi_{fix}(\Theta^{MV}(x_i)), y_i^{gt}\right)\} \quad (1)$$

여기서, $L_R(x_i)$은 주어진 입력 x_i에 대해 CompNet 에서 생성한 비트스트림의 예상 길이 이며, $L_A(y_i, y_i^{gt})$ 는 MVNet 의 임무수행의 정확도 손실(Loss)이다. Ψ 는 MVNet 의 매개변수를 나타내며, Ψ_{fix} 는 학습 과정에서 Ψ 가 업데이트되지 않음을 의미한다. MVNet 은 CompNet 과 독립적으로 미리 학습된 네트워크를 이용한다. λ_{RA} 는 L_R 과 L_A 을 절충하기 위한 가중치이다.

하나 이상의 기계 임무에 대해 상기의 CompNet 학습을 수행함으로써 제안한 VCM 프레임워크를 위한 공통 부호화기 및 일반적 기계 임무수행을 위한 복호화기를 얻을 수 있다. 이를 기반으로, 특정 기계 임무수행 또는 사람의 임무수행을 위해 보다 높은 압축효율의 개선이 필요한 경우에는 수식 (1)에서 Θ^{MV} 중 θ_e^{MV} 를 기 학습된 매개변수로 고정한 채 θ_d^{MV} 만을 추가로 학습시킬 수 있다. 이를 통해, 임무마다 서로 다른 비트스트림을 사용하는 대신 하나의 공통 비트스트림으로 다중 기계 임무 및 사람 임무를 효과적으로 지원할 수 있는 장점이 있다.

4. 실험 및 결과분석

제안하는 VCM 구조의 압축효율성 향상을 확인하기 위해, 그림 2 의 CompNet 으로 mbt2018-mean 을 사용하고 MVNet 으로는 표 1 의 Faster R-CNN X101-FPN [8]을 사용하여 객체탐지 임무에 대해 학습을 진행하였다. 학습의 시작에 앞서 Θ^{MV} 는 CompressAI [7]에서 제공된 매개변수로 초기화되었다. 넓은 범위의 비트율에 대한 효율적인 압축을 제공하기 위해, CompNet 은 서로 다른 λ_{RA} 값을 사용하여 1~6 의 품질 수준(Quality level)으로 각각 학습되었다. 학습과 평가를 포함한 모든 실험은 원본 입력영상 해상도에서 진행하였다. 학습완료 후 VVC 와의 압축효율 비교를 위해 각 품질 수준 별 CompNet 에 대해 BPP (Bit Per Pixel)와 mAP (mean Average Precision)를 측정하여 R-P (Rate-Performance) curve 를 그리고 이를 VVC(VTM-8.2) QP={47, 42, 37, 32, 27, 22}의 R-P curve 와 비교하여 BD-rate 변화를 측정하였다.

그림 3 과 그림 4 에 객체탐지 및 객체분할 임무에 대해 제안 방식의 압축효율을 VVC 와 비교하기 위한 R-P curve 들을 각각 나타냈다. 각 그림에서 파란색 점선은 VVC 앵커, 빨간색 점선은 제공된 매개변수로 초기화한 mbt2018-mean 모델, 빨간색 실선은 제안 방식으로 학습한 mbt2018-mean 모델의 R-P curve 를 나타낸다. 그리 고, 보라색 점선은 압축하지 않은 원본영상으로 측정한 mAP 값을 보인다. 그림 3 과 그림 4 에서, 제안 방식을 통해 학습된 압축 심층신경망을 통해 목적

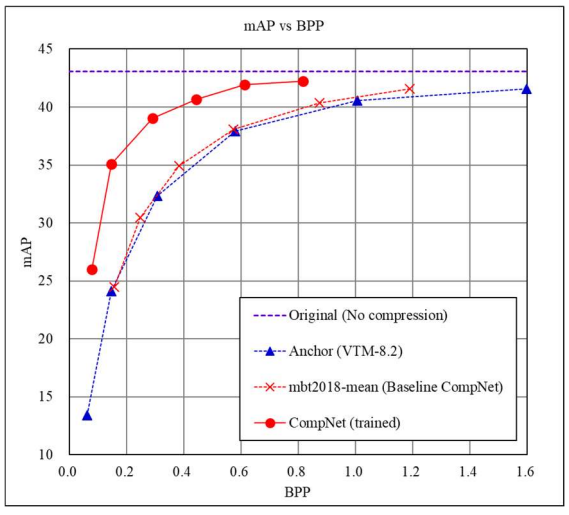

그림 3. 제안 방식의 VCM 객체탐지 압축효율 비교

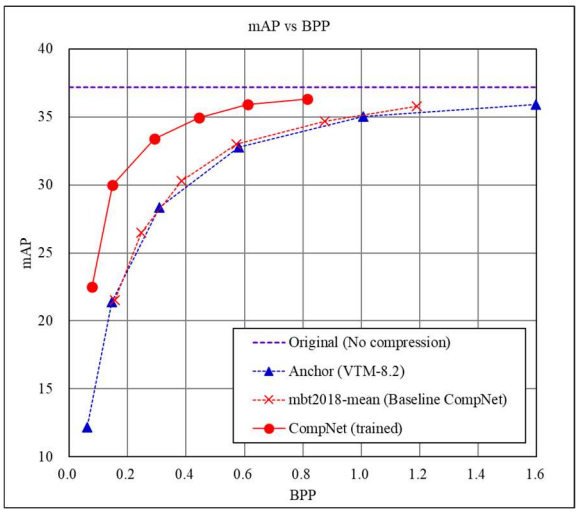

그림 4. 제안 방식의 VCM 객체분할 압축효율 비교

임무(Target task)인 객체탐지 뿐만 아니라, R50-FPN 을 통해 수행한 객체분할(표 1 참고) 또한 VVC 에 비해 압축효율이 크게 개선되었음을 알 수 있다. 표 3 과 표 4 에 제안 방식의 성능 평가 결과를 상세히 나타냈다. 제안방식은 VVC 앵커에 비해 객체탐지에서 평균 61.16%, 객체분할에서 평균 58.43%의 BD-rate 감소를 보인 것으로 확인되었다.

표 2. 제안 방식의 VCM 객체탐지 성능 평가 결과

	Anchor (VTM-8.2)		CompNet (trained)	
QP (Q. level)	BPP	mAP	BPP	mAP
47(1)	0.062	13.409	0.079	25.980
42(2)	0.147	24.133	0.147	35.095
37(3)	0.308	32.355	0.292	39.036
32(4)	0.580	37.933	0.444	40.645
27(5)	1.006	40.536	0.612	41.926
22(6)	1.598	41.558	0.816	42.246
BD-rate	-		-61.16%	

표 3. 제안 방식의 VCM 객체분할 성능 평가 결과

	Anchor (VTM-8.2)		CompNet (trained)	
QP (Q. level)	BPP	mAP	BPP	mAP
47(1)	0.062	12.181	0.079	22.519
42(2)	0.147	21.379	0.147	29.999
37(3)	0.308	28.335	0.292	33.399
32(4)	0.580	32.765	0.444	34.966
27(5)	1.006	35.057	0.612	35.940
22(6)	1.598	35.907	0.816	36.311
BD-rate	-		-58.43%	

5. 결론

본 논문에서 공통의 비트스트림을 기반으로 다중 기계 임무수행이 가능한 VCM 프레임워크가 제안되었다. 제안된 기술은 압축 심층신경망을 기계 임무수행을 위해 학습시킴으로써 기존 대비 매우 높은 압축효율을 달성할 수 있으며, 학습된 압축 심층신경망을 다른 기계 임무에 적용하였을 때에도 상당히 높은 압축효율 개선이 관찰되었다. 따라서, 제안된 방식은 VCM 문제해결을 위한 유효한 접근방법으로 판단 되었다.

감사의 글

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-00011, (전문연구실)기계를 위한 영상부호화 기술)

참고문헌

[1] Y. Zhang, M. Rafie and S. Liu, “Use cases and requirements for Video Coding for Machines”, ISO/IEC/JTC1/SC29/WG2/N43, Jan. 2021.

[2] M. Rafie, Y. Zhang, and S. Liu, “Call for Evidence for Video Coding for Machines”, ISO/IEC/JTC1/SC29/WG2/N42, April. 2021.

[3] M. Rafie, Y. Zhang, and S. Liu, “Evaluation Framework for Video Coding for Machines”, ISO/IEC/JTC1/SC29/WG2/N78, April. 2021.

[4] S. Cho, H. Lee, S. Jeong, J. Lee, Y. Kim, J. Do and J. Choi, “[VCM] Image compression neural network optimized for object detection”, ISO/IEC/JTC1/SC29/WG2/m56469, April. 2021.

[5] S. cho, Y. Kim, W. Lim, H. Kim and J. Choi, “A Technical Analysis on Deep Learning based Image and Video Compression”, JBE Vol. 23, May. 2018.

[6] CompressAI: https://github.com/InterDigitalInc/CompressAI

[7] Detectron2: https://github.com/facebookresearch/detectron2

[8] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video Coding for Machines: A Paradigm of Collaborative Compression and Intelligent Analytics," IEEE Trans. Image Process., vol. 29, pp. 8680-8695, Aug. 2020.