신경망 비디오 표현 기술 및 MPEG-INVR 표준화 동향

□ 방건 / 한국전자통신연구원

요약

최근에 주목을 받고 있는 암시적 신경망 비디오 표현(INVR: Implicit Neural Visual Representation)은 2D/3D 비디오를 새로운 방식으로 표현하기 위한 기술로 빠르게 발전하고 있다. 이 분야에서는 신경망 학습을 통해 2D와 3D 비디오를 함축적으로 표현하는 연구가 주로 진행되고 있다. 최근 MPEG Video Coding(ISO/IEC SC29 WG4) 그룹에서는 이와 같은 최신의 연구 결과들을 바탕으로 압축 관점에서 바라보며 표준화를 준비하기 시작하였다. 이를 위해 INVR AhG를 구성하였으며, 관련 전문가들은 최신 기술 동향을 조사하고, 표현 능력과 압축 성능을 평가하기 위한 탐색 실험을 진행하고 있다. 본 기고문에서는 최신 암시적 신경망 표현 기술과 MPEG INVR AhG에서 진행된 2D와 3D 비디오의 신경망 표현에 대한 탐색 실험 내용을 정리하였다.

I. 서 론

최근 10년간 영상 처리 분야에서 딥러닝 기술은 많은 발전을 이루었다. 딥러닝은 영상 인식과 분류, 영상 생성 및 변환, 영상 압축 등 다양한 작업에서 좋은 결과를 보여주고 있다. 특히, Convolutional Neural Networks(CNN), Generative Adversarial Networks(GAN), Variational Autoencoders(VAE) 등의 다양한 딥러닝 모델들은 영상 처리 분야에서 주목받고 있는 기술들이다. 일례로 딥러닝을 이용한 영상 인식 작업에서는 영상의 특징을 자동으로 추출하여 분류하거나 객체를 감지하는 등의 작업을 수행할 수 있으며, GAN이나 VAE와 같은 모델을 사용하면 영상 생성, 영상 스타일 전이, 영상 복원 등 다양한 영상 생성 및 변환 작업을 수행할 수 있다. 뿐만 아니라, 딥러닝을 영상 압축에 적용하여 더 효과적인 압축 방법을 개발하는 연구도 활발히 진행되고 있는 분야 중 하나이다.

이러한 영상 분류, 생성, 복원 등을 다루는 딥러닝 기술과는 다르게 2020년 초부터 영상 자체를 학습된 모델로 표현할 수 있는 방법에 대한 연구도 이루어지고 있는데, 이를 보통 암시적 신경망 비디오 표현(Implicit Neural

Visual Representation) 기술이라 한다. 전통적인 방법에서 2D 비디오는 각 픽셀의 (x, y) 좌표값과 해당 좌표의 색상값으로 표현하고, 3D 비디오는 (x, y, z) 좌표와 해당 좌표의 색상값으로 표현하는 것이 일반적이다. 하지만 암시적 신경망 비디오 표현 방법은 좌표 기반 색상값 표현이 아니라, 신경망 학습을 사용하여 영상을 함축적인 모델로서 표현하는 것이 특징이다. 이 방법은 딥러닝 모델을 활용하여 영상의 공간적인 특성과 시간적인 변화를 학습하여 픽셀의 좌표보다는, 암시적 신경망으로 학습된 가중치와 특징 벡터를 통해 영상을 재구성하거나 생성할 수 있는 표현 방법이다. 여기서 학습된 가중치와 특징 벡터로 구성된 것을 암시적 신경망 비디오 표현 모델이라고 부른다. 암시적 신경망 비디오 표현 방법은 여러 개의 레이어가 완전 연결(Fully connected)된 형태의 단순한 신경망 구조를 기반으로 하고 있으며, 신경망은 비디오 표현에 필요한 특징들을 학습할 수 있게 설계된 것이 특징이다. 이를 통해 기존의 좌표계-색상값 표현 방식보다 더 효율적으로 비디오를 표현할 수 있기 때문에, 영상 생성, 변환, 압축 등 2D/3D 비디오 분야에서 새로운 표현 방식으로 받아들여질 것으로 기대된다.

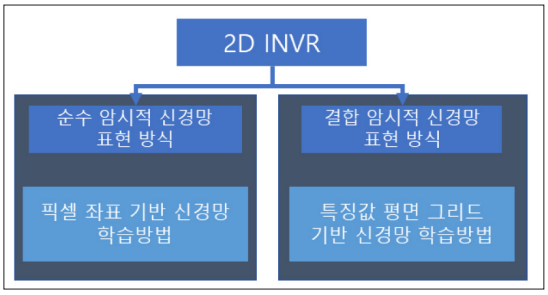
최근 MPEG Video Coding(ISO/IEC SC29 WG4) 그룹에서는 이러한 연구의 결과들에 대해 관심을 가지고 압축 관점에서 표준화를 준비하기 시작하였다. 2022년 7월부터 INVR AhG를 구성하여 관련 분야의 전문가들이 모여 최근 발표된 다양한 신경망 비디오 표현 기술들을 이용한 2D 및 3D 비디오의 표현 능력과 압축 성능에 대한 실험을 진행해 오고 있다.

본 기고에서는 지금까지 발표된 암시적 신경망 기술들을 2D와 3D 비디오 표현 방법들로 분류해서 소개하고 MPEG에서 진행되었던 실험 내용을 설명하고자 한다.

II. 2D 신경망 비디오 표현 방법 소개

암시적 신경망을 사용한 2D 비디오 표현 방법은 2D 비디오를 신경망을 통해 학습시킴으로써 함축적인 모델로 2D 비디오를 나타낼 수 있기 때문에 비디오 압축 분야에 활용할 수 있다. 이 기술은 〈그림 1〉과 같이 두 가지 방식으로 정리할 수 있다. 하나는 2D 비디오 전체를 신경망 학습만을 이용해 하나의 모델로서 표현하는 방법으로 이를 순수 암시적 신경망 표현(Full Implicit Neural Representation(FINR))이라고 한다. 이는 2D 비디오의 픽셀 좌표를 입력 받아서 신경망 학습을 통해 비디오를 표현하는 방법이 있다.

결합 암시적 신경망 결합 표현(Hybrid Implicit Neural Representation(HINR))에서는 2D 비디오를 그리드 형태로 정렬된 특징 집합으로 변환한다. 예를 들면, 비디오 프레임을 그리드 형태의 행렬로 표현하고, 각 행렬의 픽셀값을 특징으로 사용할 수 있다. 이렇게 생성된 특징 집합은 신경망에 입력되어 비디오 프레임을 학습할 수 있다.



<그림 1> 2D 신경망 비디오 표현 방법 분류

1. 순수 암시적 신경망 표현 방법들

2D 비디오를 위한 순수 암시적 신경망 표현(FINR) 방법 중 대표적인 SIREN[1] 기술은 픽셀의 좌표(x, y)와 프레임을 구분하기 위한 인덱스(t)를 신경망의 입력으로 받는다. 신경망 학습은 각 프레임의 픽셀 정보를 암시적으로 표현할 수 있으며, 이 정보는 학습된 신경망 모델의 형태로 매우 함축적으로 표현되기 때문에 2D 비디오를 압축하는 효과를 얻을 수 있다.

NeRV[2] 기술은 신경망 학습을 위해 프레임 인덱스(t)만을 입력으로 하여 프레임 전체를 학습시키는 방식을 취하고 있다. NeRV는 Convolution Network을 포함하고 있는 신경망 구조를 갖고 있기 때문에 시간 t에 대해 프레임을 학습할 수 있는 특징을 가지고 있다. PS-NeRV[3]는 프레임 인덱스와 프레임 내의 픽셀값들을 패치 단위로 구분한 위치값을 입력하여 학습시키는 방식이다.

2. 결합 암시적 신경망 표현 방법들

결합 암시적 신경망 표현(HINR) 방식은 2D 비디오를 평면 그리드로 구성하여 여기에서 얻어진 값을 신경망으로 학습하는 방법이다. 암시적 신경망은 입력과 출력 사이의 매핑을 직접적으로 명시하여 정의하지 않고, 신경망을 사용하여 매핑을 학습하는 방법이다. 이를 통해 비디오의 복잡한 특징을 추출하고 비디오 데이터의 공간적인 구조를 인식할 수 있다. 예를 들면, 신경망은 그리드에 가까운 위치에 있는 픽셀값을 보간법을 통해 계산할 수 있도록 그리드의 특징값을 학습시킬 수 있다.

이런 방법 중에 대표적으로 NVP[4]는 픽셀 좌표(x, y)와 시간 t에 대해 (x, y), (x, t)와 (y, t)에 대한 3개의 평면 그리드를 구성하고 마지막으로 (x, y, t)에 대해 3D 그리드를 구성한다. 이들 각각은 신경망을 통해서 학습되고 학습된 특징값은 JPEG 압축 방식을 이용하여 압축한다.

또 다른 방법인 FFNeRV[5]가 있는데, 이는 공간 해상도와 다양한 변이를 가질 수 있는 시간 해상도를 표현하는 그리드로 구성하여 신경망 학습을 수행한다.

HNeRV[6]는 비디오의 픽셀 정보를 저차원 형태의 벡터로 표현한 특징값을 갖고 복원 과정에서 벡터 정보로부터 원래 정보를 얻게 되는 일종의 오토인코더 개념과 유사한 신경망 구조를 갖고 있다. 따라서 이 신경망 구조는 모든 프레임에 대해 저차원 백터값에 대해 과적합 학습을 수행하기 때문에 속도면에서 NeRV보다 우수하다는 평가를 받고 있다.

III. 3D 신경망 비디오 표현 방법 소개

암시적 신경망 비디오 표현 방식에는 3D 비디오 즉, 다시점 비디오를 입력으로 임의의 시점을 생성하고 이를 사용자에게 보여줌으로써 자연스러운 운동 시차를 제공하는 3D 신경망 비디오 표현 방법들이 있다. 3D 신경망 비디오 표현 방법에는 크게 고정된 장면을 표현하는 정적 3D 신경망 비디오 표현 방법과 시간에 따라 변화하는 움직임 장면을 다루는 동적 3D 신경망 비디오 표현 방법으로 구분할 수 있다.

1. 정적 3D 신경망 비디오 표현 방법

가장 유명한 방법으로 알려진 NeRF[7]는 입력으로 3차원 위치(x, y, z)와 시점 방향(θ, φ)을 사용하여 해당 ray-point에서의 밀도값과 색상값을 MLP(Multilayer perceptron) 구조의 신경망을 이용하여 매개 변수화시키는 학습 방법이며, 학습된 모델은 렌더링을 위해 밀도값과 색상값을 추정할 수 있다. 여기서 ray-point는 3차원 공간상에 빛이 지나가면서 보일 수 있는 샘플링된 점을 의미한다.

MLP는 입력으로 샘플의 3차원 위치와 시점 방향을 받아들이고 샘플의 특성을 학습하여 해당 위치에서의 밀도와 색상을 예측하게 된다. 렌더링 과정에서는 Ray의 시작부터 끝까지 샘플들을 결합하여 최종적인 색상을 생성한다. 이때 고주파 영역에 있는 세밀한 부분을 표현할 수 있도록 positional encoding 방식이 사용되고 있다. 추가적으로 NeRF++[8]은 NeRF를 확장하여 전경과 배경 두 개로 분리하여 두 개의 MLP 구조를 연계시켜 학습을 시킴으로써 360도 장면에 대해 학습할 수 있는 기술도 있다.

하지만 최근에는 장면 표현의 요소로서 계산 복잡도가 높은 ray 샘플링에 기반한 순수 MLP를 사용할 필요가 없는 방식도 소개되고 있다. 이 방법은 3D 비디오를 먼저 복셀(voxel) 그리드로 표현함으로써 비교적 짧은 시간 내에 3D 비디오에 대한 학습을 마칠 수 있도록 하는 방법이다.

대표적인 기술인 Plenoxel[9]은 우선 시점에 따라 변할 수 있는 밀도값과 구면조화(Spherical harmonic)계수들로 구성된 희소 복셀 그리드를 명시적으로 구성한다. 여기서 신경망은 학습 과정에서 간단한 정규화 함수를 사용하여 복셀 그리드 내의 잡음을 완화해 가면서 3D 비디오의 3차원 복셀 구조를 완성해 나아가는 방법을 사용하고 있다.

하지만 메모리 용량은 복셀 그리드 해상도의 증가와 함께 기하급수적으로 증가하기 때문에 하드웨어 비용이 높아지는 단점이 있다. 이를 극복하고자 여러 개의 개선 방안이 제안되었으며, 대표적으로 Instant-NGP[10], TensorRF[11]와 같은 방법이 알려져 있다.

2. 동적 3D 신경망 비디오 표현 방법

3D 영상을 표현하는 대표적인 방법인 NeRF는 정적인 장면에서는 매우 좋은 결과를 얻을 수 있지만, 움직이는 객체나 장면에 대해서는 좋은 결과를 얻을 수 없다. 하지만 현실 세계에서는 일반적으로 동적인 변화를 포함하고 있으므로, 시공간을 포함한 3D 비디오를 표현하는 기술은 매우 중요하다.

대표적인 방법으로는 D-NeRF[12]가 있는데, 이 기술은 동적 볼륨 데이터를 표현하기 위해 시간축상에서 변형가능한 동적 영역 필드와 고정된 부분을 나타내는 정적 영역 필드를 나누어 사용하는 방법이다. 이 방법은 단안 3D 비디오 시퀀스를 입력으로 받기 때문에 매 프레임이 각기 다른 시점으로 취급되며, 각 프레임마다 동적 영역 필드에서는 변형 가능 좌표를 3D 변환으로 매개 변수화하여 표현해 주어야 한다. 동적 영역 필드의 매개 변수는 관찰 공간의 관찰점(x, y, z)에서 정규 공간의 해당 지점까지의 상대적인 위치 이동(dx, dy, dz)을 나타내며, 이를 통해 관찰점의 좌표를 정규 공간으로 변환할 수 있게 된다. 이때 신경망에서는 변환된 정규 공간 좌표를 입력 받아 NeRF와 유사한 방법으로 체적 밀도와 색상 정보를 학습한다.

다른 방법으로 DyNeRF[13]는 암시적 함수를 위한 입력으로 이산적인 시간 변수 대신 일련의 시간에 종속된 잠재 코드를 학습에 사용하도록 제안했다. 이 방법에서는 단순 MLP 아키텍처를 사용하기 때문에 장면 표현 모델은 간결해지지만, 학습 및 렌더링 시간 모두 많이 소요되는 단점이 있다.

IV. MPEG INVR 표준화 동향

MPEG은 비디오/오디오/그래픽 데이터의 기기 간 상호 호환성을 보장할 수 있는 압축 및 저장 방법에 대한 표준 개발을 주로 하고 있는 국제표준화 단체이다. 최근 MPEG은 암시적 신경망 표현 방법에 대한 관심을 가지고 있으며, 2022년 7월부터 MPEG INVR AhG을 구성하여 표준화를 위한 탐색 실험을 진행하고 있다.

이를 위해 MPEG INVR에서는 2D 비디오의 암시적 신경망 표현에 대한 탐색 실험과 3D 비디오의 암시적 신경망 표현에 대한 탐색 실험으로 나누어 진행하고 있다. 이번 장에서는 지금까지 MPEG INVR에서 진행되었던 다양한 암시적 신경망 표현에 대한 영상 품질과 압축에 대한 탐색 실험 결과를 정리하였다.

1. 2D INVR 탐색 실험

2D INVR에서는 모델의 압축 시 품질에 미치는 영향에 대한 탐색을 진행하고 있다. 2D INVR에서 진행하고 있는 탐색 실험의 목표는 2D 비디오를 암시적 신경망 표현으로 학습시킨 모델에 대한 복원 영상 품질 및 압축 성능에 대한 탐색 실험을 진행하고 있다.

본 장에서는 프레임 간의 연관성 없이 압축하는 방법인 All-Intra coding과 프레임 간의 중복성을 고려하면서 압축하는 방법인 Inter-like coding 실험에 대한 탐색 결과를 정리하였다.

1) All Intra mode 압축 성능 탐색 실험

2D INVR 방법으로 학습된 모델을 신경망 모델 압축 기술

을 사용하여 실험한 탐색 실험 결과를 살펴보면 다음과 같다. 여기서 사용된 신경망 모델 압축 기술인 NNCodec[14]은 MPEG에서 표준화된 NNC(Neural Network Coding/ISO/IEC 15938-17)를 근간으로 구현된 모델 압축을 할 수 있는 소프트웨어이다.

이 소프트웨어를 사용하여 서로 다른 QP 설정을 통해 다른 압축률을 갖는 압축 모델을 얻을 수 있으며, 실험은 이때 복원된 모델의 복원 품질을 측정하는 실험을 진행하였다. 이 품질의 비교를 위해 Quantization Aware Training(QAT)을 사용한 결과를 함께 비교하였다. 탐색 실험에 사용된 2D INVR 방법은 INRIC를 사용하여 모델을 학습시켰으며, 이 기술은 가중치 초기화를 위해 COIN [15] 방법을 사용하고 positional encoding을 위해 SIREN 방법과 결합한 방법이다.

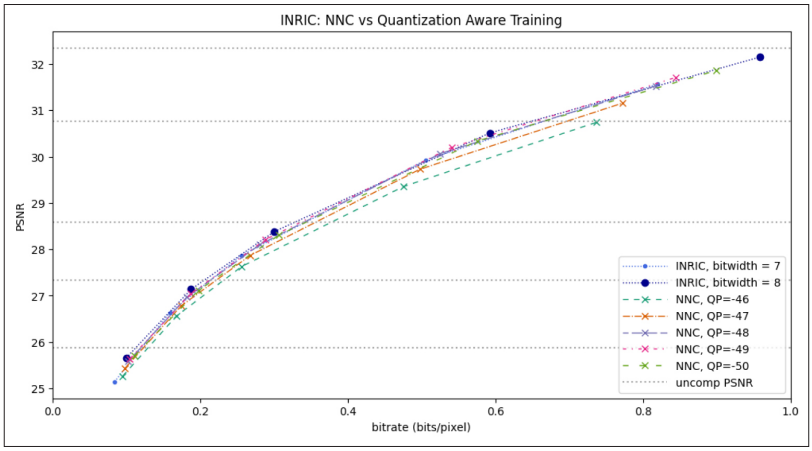
우선 실험에서는 INRIC를 사용 시, 아래의 〈표 1〉과 같이 동일 영상에 대해 다른 학습 구조를 사용하여 다른 품질의 모델 5개를 준비하였다. 품질은 Q5일 때가 가장 좋고 Q1일 때가 가장 나쁜 것을 의미한다. 압축되지 않은 모델의 비트레이트는 1.3 ~ 14.3 bits/pixel 구간을 가질 수 있도록 학습하였다.

<표 1> INRIC 모델 구조별 품질 구분

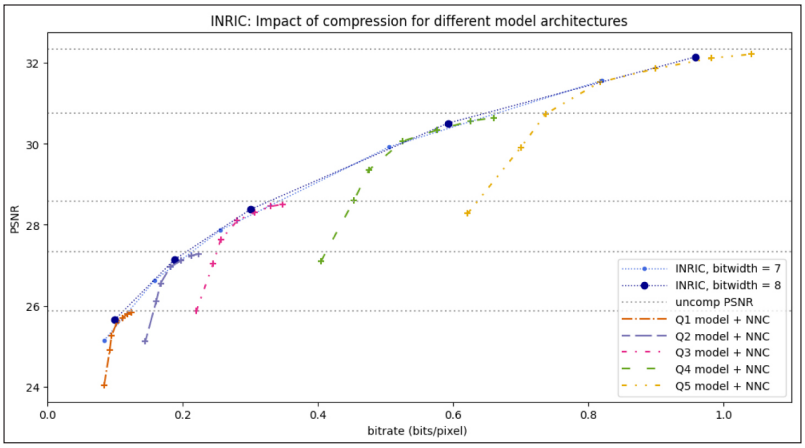
Quality Order	Number of hidden layers	Layer size
Q1	3	32
Q2	3	48
Q3	3	64
Q4	3	96
Q5	3	128

NNCodec은 QP 매개 변수에 -54에서 -42까지의 값들을 사용하여 압축되지 않은 5개의 모델에 적용하였다. 그리고 비교를 위한 QAT 방식은 영상에 대한 학습을 끝낸 후, AI Model Efficiency Toolkit(AIMET)을 사용하여 가중치를 7-8 비트 범위의 비트 너비로 양자화한다. 양자화로 인한 성능 손실을 줄이기 위해 AdaRound(가중치를 올릴지 내릴지 결정하는 2차 최적화 방법)을 사용하여 후처리 양자화 최적화를 수행한 다음, QAT를 사용하여 가중치를 세밀하게 조정하였다. 마지막으로, 가중치를 무손실 압축하기 위해 이진화된 산술 부호화 알고리즘을 사용하여 엔트로피 코딩을 수행하여 압축된 모델을 얻는 방식으로 실험을 진행하였다.

첫 번째 실험에서는 INRIC에서 학습한 각기 다른 모델들에 대해 각각 다른 양자화 매개 변수를 사용하여 NNCodec으로 압축을 수행했으며, INRIC에서 7비트와 8비트로 QAT를 사용한 5개의 다른 모델을 얻을 수 있었다. 수평으로 그려진 다섯 개의 점선들은 각기 다른 모델들로부터 얻어진 성능을 나타낸 것이다.



<그림 2> INRIC 모델별 QAT와 NNC QP별 Rate-Distortion 비교



<그림 3> NNC QP별 INRIC 모델별 Rate-Distortion 비교

실험 결과, NNC는 비트레이트와 영상 복원 품질 사이에서 어느 정도 유연성을 갖추며 좋은 성능을 보였다. 예를 들어, NNC에서 QP 매개 변수를 적절하게 선택한 경우에는 INRIC 기반의 후처리 양자화 단계보다 뛰어난 성능을 보일 수도 있다. 그러나 QAT를 사용한 8비트 양자화가 전반적으로 가장 우수한 RD(율-왜곡) 성능을 보였다.

두 번째 실험은 〈그림 3〉에서 보인 것 같이 다른 모델 학습 구조를 갖고 학습한 모델들을 가지고 실험을 수행하였다. 이 실험은 다른 학습 구조로부터 얻어진 모델들에 대해 각각 양자화 매개 변수 QP별 (-46, -47, -48, -49, -50) 압축 성능을 조사한 것이며, INRIC은 그대로 각기 다른 모델 5개에 대해 7비트와 8비트로 QAT한 결과를 그대로 사용하였다.

두 번째 실험 결과, 최고 품질 모델(Q5로 지칭)의 성능에서 보면 QP 매개 변수를 감소시키면 빠르게 영향을 받는 것을 알 수 있다(다른 모델에서도 마찬가지임). 결론적으로 NNCodec을 사용한 단일 모델 압축 시, 완만한 RD 변화를 발생시키는 것은 어렵다는 것을 알 수 있었고, 첫 번째 실험과 두 번째 실험을 종합해 보면 완만한 RD 변화를 유지하기 위해서는 NNCodec에서 단순히 양자화 매개 변수를 변화시키는 방법보다는 RD 변화 관점에서 적절히 모델을 바꾸어 가며 훈련시키는 것이 압축을 위해 필요함을 알 수 있었다.

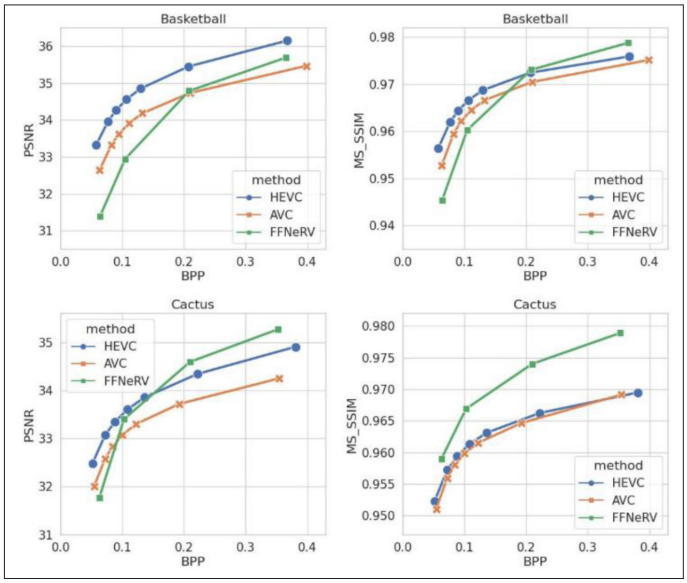
2) Inter-like mode 압축 성능 탐색 실험

Inter-like mode 압축 성능 탐색 실험에서는 전통적인 부호화 방법인 AVC, HEVC와 신경망 비디오 표현 방법인 FFNeRV 간의 압축 효율을 비교하는 탐색 실험을 진행하였다.

FFNeRV는 특징 그리드와 합성곱 신경망으로 구성된 혼합형 암시적 신경 표현 방법 중의 하나이다. 이 방법에서는 양자화를 위해 QAT를 사용하고 있다. 따라서 훈련 후 가중치 양자화와 엔트로피 부호화를 적용하여 압축을 수행할 수 있다. 실험은 JVET에서 가져온 두 개의 1080p 비디오(Cactus와 Basketball)를 대상으로 진행되었다. 이 비디오들은 50 프레임레이트로 구성되었으며, Cactus는 카메라 움직임이 없는 세 개의 회전판과 정적 물체를 포함한 특성을 가지고 있는 비디오이고 Basketball은 선수들의 빠른 움직임을 포함하고 있는 비디오이다.

〈그림 4〉를 통해 탐색 실험 결과를 다음과 같이 정리할 수 있다.

- HEVC는 카메라 움직임이 있는 장면, 즉 Basketball 시퀀스의 부호화에서 우수한 성능을 보임
- FFNeRV는 카메라 움직임이 없는 장면, 즉 Cactus 시퀀스에서 HEVC보다 우수한 성능을 보임
- FFNeRV는 MS_SSIM 측면에서 전반적으로 HEVC보다 우수한 구조적 유사성을 보임



<그림 4> HEVC, AVC, FFNeRV Rate-Distortion/MS-SSIM 비교(Basketball, Cactus)

탐색 실험 결과, 매우 완만한 움직임에 대해서는 부호화 효율이 HEVC 성능보다 좋으며, MS-SSIM 측면에서는 신경망 비디오 표현 방법이 더 좋은 결과를 보여주는 것으로 관찰되었다.

<표 2> Instant-NGP 복원 PSNR/SSIM 측정 결과("CBABasketball")

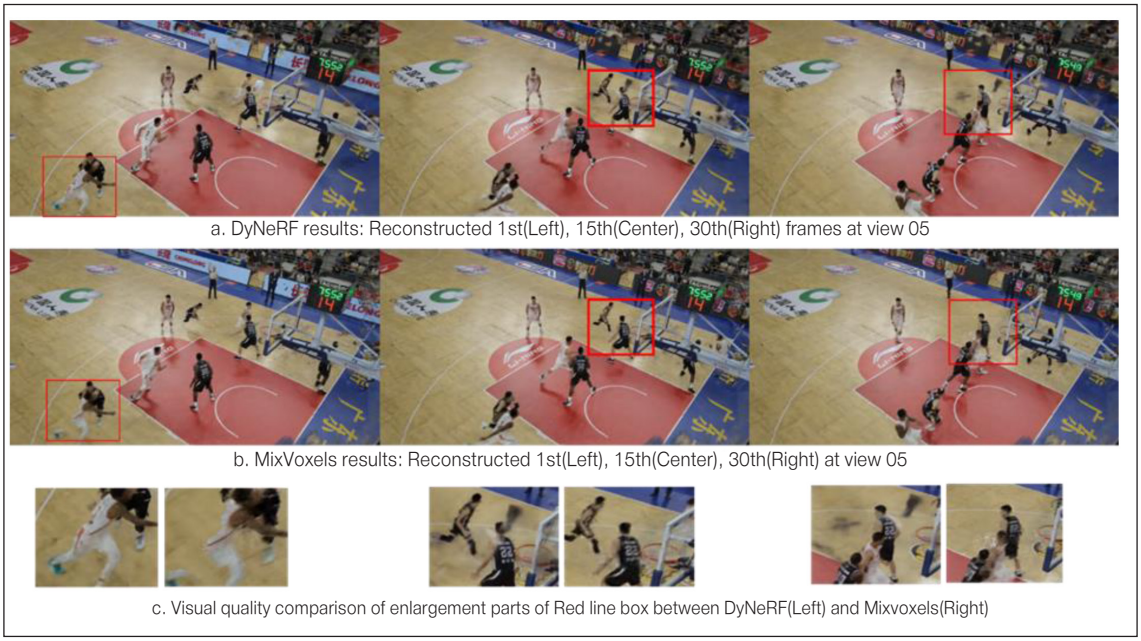
	V05	V25
Avg. PSNR	28.44	29.03
Avg. SSIM	0.9154	0.9302

2. 3D INVR 탐색 실험

1) 2.1. 3D INVR 렌더링 탐색 실험

3D INVR 탐색 실험은 Dynamic-NeRF를 사용하여 복원한 영상의 시각적 품질을 조사하고, 특히 다시점 영상에서의 시간적 일관성을 유지하는 복원 품질을 중요하게 다루었다. Dynamic-NeRF는 훈련된 모델을 사용하여 동적 장면의 고품질 3D 비디오를 재구성하는 것이 주요 목표이다. 시각적 품질의 평가는 프레임 단위 학습 방식과 시퀀스 단위 학습 방식 두 가지로 나누어서 탐색 실험이 진행되었다. Instant-NGP는 프레임 단위 학습 방식으로 선택되어, 비디오 시퀀스를 시간적 연관성 없이 프레임 단위로 훈련하였다. 시퀀스 단위 학습 방식의 경우, 동적인 장면을 훈련하기 위해 설계된 Dy-NeRF와 Mixvoxels[16]을 선택하였으며, 공통적으로 "CBABasketBall" 시퀀스를 탐색 실험에 사용하였다. 〈표 2〉와 같이 Instant-NGP의 테스트 뷰에서 프레임별 재구성 품질은 객관적으로 허용 가능한 품질로 복원되었다는 것을 알 수 있었다. 그러나 테스트 뷰에서 재구성된 프레임들 간의 시간적 일관성은 불안정하였으며, 나무 바닥의 반사면 및 바구니(Basket) 그물 주변에서 시간적으로 일관되지 않는 흐릿한 질감과 떨림 현상이 관찰되었다. 따라서 프레임 단위 학습 방식으로 비디오를 신경망으로 표현하는 것은 시간적 일관성 측면에서 좋은 선택은 아니라는 것을 알 수 있었다.

시퀀스 단위 학습 방식인 DyNeRF와 Mixvoxels의 경우, 시간적 일관성을 유지할 수 있지만 큰 볼륨 공간과 빠른 움직임에서는 여전히 단점이 관찰되었다. 〈그림 5〉는 이러한 단점을 보여주는 예인데, 경기장 바닥에 대한



a. DyNeRF results: Reconstructed 1st(Left), 15th(Center), 30th(Right) frames at view 05

b. MixVoxels results: Reconstructed 1st(Left), 15th(Center), 30th(Right) at view 05

c. Visual quality comparison of enlargement parts of Red line box between DyNeRF(Left) and Mixvoxels(Right)

<그림 5> DyNeRF, Mixvoxels의 복원 영상의 시각적 품질 비교의 예("CBABasketball")

시간적 일관성이 잘 유지되는 것을 관찰할 수 있으나, 빨간 박스로 표시된 부분에 대해서는 빠르게 움직이는 사람들이 흐릿하게 나타나거나 심지어 사라지는 것이 관찰되었다.

2) 3D INVR 압축 탐색 실험

Dynamic-NeRF의 모델 압축을 위해 DyNeRF와 Mixvoxels 모델을 Neural Network Coding(NNC)을 사용하여 압축하고 그 성능을 객관적 및 주관적인 품질 관점에서 탐색 실험을 진행하였다. 복원 영상의 품질은 NNCodec을 사용하여 5가지 다른 QP값 {-48, -43, -38 (기본값), -33, -28}으로 압축된 모델에 대해 평가되었다. PSNR 결과는 원본과 압축 전 복원 시점 영상과 비교하여 각각 평가하였다. 평가에서 PSNR 품질은 -38 이하의 QP 값에서 품질이 보존되며, 해당 QP에서 약 20%의 압축률 성능이 관찰되었다. 이 탐색 실험을 통해 NeRF 모델 압축에 대한 NNC의 압축 성능은 QP값이 어느 이상을 넘게 되면 시각적 품질이 빠르게 저하됨을 관찰할 수 있었다.

V. 결 론

본 기고는 2D/3D 암시적 신경망 비디오 표현 기술의 최신 동향을 검토하였다. 그리고 MPEG INVR 표준화에서는 2D/3D INVR에는 정적 및 동적 장면의 렌더링 품질 개선, 훈련 및 추론 가속화, 모델 압축 등 부분에서 추가적인 탐색 실험의 결과에 대해서 정리하였다. 현재까지의 탐색 실험 결과는 일부 실험에서 2D 암시적 신경망 비디오 표현 방법이 좋은 율-왜곡 결과를 보여주고 있음을 관찰할 수 있었다. 하지만 2D 암시적 신경망 비디오 표현 방법은 콘텐츠에 따라 민감하며 특정 장면에서만 잘 동작함을 알 수 있었다. 3D 암시적 신경망 비디오 표현 방법은 non-Lambertian 표면과 좁은 영역을 표현한 장면에서 좋은 성능을 보여주지만, 고주파 성분이 있는 세부 사항에는 여전히 복원 성능이 만족스럽지 못함을 보여주고 있다. Dynamic-NeRF의 시각적 품질은 시간적 일관성을 기준으로 검증되었지만, 광역 공간 및 빠른 움직임이 있는 장면에서는 흐릿한 질감이 관찰되었다. 마지막으로,

NNCodec은 암시적 신경망 비디오 표현 모델 압축을 위한 유망한 기술로 간주될 수 있으나, QP값만을 사용하여 압축률을 높이면 급격히 품질이 저하되는 단점이 있어 추가적인 탐색이 필요한 상황이다.

끝으로 암시적 신경망 비디오 표현의 장점을 파악하기 위해 MPEG INVR AhG에서는 더 많은 기술에 대해 탐색 실험이 진행될 예정이며 동시에 MPEG INVR의 표준 범위를 수립해 나아갈 계획이다.

참 고 문 헌

[1] Y. Strümpler, J. Postels, R. Yang, L. van Gool, and F. Tombari, "Implicit Neural Representations for Image Compression." arXiv, Aug. 03, 2022. Available: http://arxiv.org/abs/2112.04267, Code under MIT License: https://github.com/YannickStruempler/inr_based_compression

[2] Chen, Hao, et al. "Nerv: Neural representations for videos." Advances in Neural Information Processing Systems 34 (2021): 21557-21568.

[3] Bai, Yunpeng, Chao Dong, and Cairong Wang. "PS-NeRV: Patch-wise Stylized Neural Representations for Videos." arXiv preprint arXiv:2208.03742 (2022).

[4] S. Kim, S. Yu, J. Lee, and J. Shin, "Scalable Neural Video Representations with Learnable Positional Features," in NeurIPS, 2022.

[5] J. Lee, D. Rho, J. H. Ko, and E. Park, "FFNeRV: Flow-Guided Frame-Wise Neural Representations for Videos," arXiv preprint arXiv:2212.1229 (2022).

[6] H. Chen, M. Gwilliam, B. He, S.-N. Lim, and A. Shrivastava, "HNeRV: A Hybrid Neural Representation for Videos," in CVPR, 2023.

[7] Mildenhall, B, et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." European conference on computer vision. 2020.

[8] Zhang, Kai, et al. "Nerf++: Analyzing and improving neural radiance fields." arXiv preprint arXiv:2010.07492, 2020.

[9] Fridovich-Keil, Sara, et al. "Plenoxels: Radiance fields without neural networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[10] Müller, Thomas, et al. "Instant neural graphics primitives with a multiresolution hash encoding." ACM Transactions on Graphics (ToG) 41.4 (2022): 1-15.

[11] Chen, Anpei, et al. "Tensorf: Tensorial radiance fields." European conference on computer vision. 2022.

[12] Pumarola, Albert, et al. "D-nerf: Neural radiance fields for dynamic scenes." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[13] Li, Tianye, et al. "Neural 3d video synthesis from multi-view video." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[14] "Fraunhofer Neural Network Encoder/Decoder (NNCodec)." Fraunhofer HHI, Mar. 13, 2023. Accessed: Mar. 14, 2023. [Online]. Available: https://github.com/fraunhoferhhi/nncodec

[15] E. Dupont, A. Goliński, M. Alizadeh, Y. W. Teh, and A. Doucet, "COIN: COmpression with Implicit Neural representations," ArXiv210303123 Cs Eess, Apr. 2021, Available: http://arxiv.org/abs/2103.03123, Code under MIT License: https://github.com/EmilienDupont/coin.

[16] Wang, Feng, et al. "Mixed Neural Voxels for Fast Multi-view Video Synthesis." arXiv preprint arXiv:2212.00190 (2022).

※ 이 기고는 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2017-0-00072, 초실감 테라미디어를 위한 AV 부호화 및 LF 미디어 원천기술 개발)

저 자 소 개

방 건

- 현재 : 한국전자통신연구원 초실감메타버스연구소 표준전문위원/책임연구원
- 2014년 : 고려대학교 컴퓨터학 박사
- 2022년 ~ 현재 : MPEG INVR AhG 그룹 Co-chair
- 2008년 ~ 2013년 : MPEG FTV AhG 그룹 EE 코디네이터
- 2011년 ~ 2012년 : MIT RLE ATSP 그룹 방문 연구원
- 2002년 ~ 2006년 : ATSC T3/S2 ACAP 데이터방송 표준 에디터
- 주관심분야 : 2D/3D 비디오 부호화, 영상 처리, 컴퓨터 비전, 인공지능