엔트로피 손실 함수를 통한 암시적 신경 표현에서의 이미지 압축 기법

이순빈, 정종범, 류은석

성균관대학교 컴퓨터교육학과

soonbinlee@skku.edu, uof4949@skku.edu, esryu@skku.edu

요 약

최근 딥러닝 분야에서는 암시적 신경 표현(Implicit Neural Representation)을 활용한 연구가 활발히 진행되고 있다. 암시적 신경 표현은 입력 좌표로부터 직접적인 신호 값으로의 파라미터화를 통해 신호의 복원을 진행한다. 본 논문에서는 암시적 신경 표현을 기반으로, 네트워크 가중치 도메인에 대한 엔트로피 손실 함수를 추가한 이미지 압축 기법을 제안한다. 원본 이미지를 네트워크에 과적합시킨 뒤, 이 네트워크 가중치 자체를 압축하여 전송하고 이를 복원함으로써 효율적인 이미지 전송이 가능함을 보인다. 제안하는 방법에서 최종 학습된 네트워크 가중치는 산술 부호화를 통해 최적화된 압축 형태로 전송되며, 암시적 신경 표현 모델을 이미지 압축 분야에 활용하여 기존 이미지 코덱인 JPEG2000 및 컨볼루션 기반 이미지 압축 모델과 비교될 수 있는 성능을 보였다.

1. 서론

이미지 압축은 멀티미디어와 신호처리 분야에서 핵심적인 연구 주제로 자리잡고 있다. 기본적인 이미지 압축 알고리즘은 공간적 중복성을 제거하기 위해 주파수 변환, 양자화, 엔트로피 부호화 과정 등을 거친다. 예를 들어 대표적인 이미지 코덱인 JPEG 에서는 이산 코사인 변환(DCT, discrete cosine transform)을 사용하여 신호 영역에서 주파수 영역으로 변환하여 이미지를 압축하게 된다.

기존 이미지 및 동영상 코덱 기술에 대한 표준화와 최적화가 이루어짐과 동시에, 최근 몇 년 동안 신경망의 표현 능력을 통해 이미지 압축 기술을 발전시키기 위한 연구가 진행중이다. 기존의 방식과 다르게, 데이터 중심 학습 모델을 통해 효율적인 압축 코덱을 설계하려는 연구 방향이 점점 더 주목받게 되었다.

대표적인 심층 이미지 압축 모델은 오토인코더(auto-encoder)를 기반으로 하는 모델로, 이미지를 잠재 벡터(latent vector)로 변환하는 인코더와 이 잠재 벡터를 다시 원래 이미지로 복원하는 디코더로 구성된다. 단순히 이미지의 정보를 축약하는 것이 아니라 전송해야 할 이미지의 크기를 최소화하기 위해, 인코딩된 잠재 벡터에 대한 엔트로피 코딩을 수행하는 심층 이미지 압축 모델이 제안되었다 [1].

잠재 벡터에 정보 손실 함수를 추가하여 산술 부호화(arithmetic coding)시 최소의 크기를 갖는 잠재 벡터의 형태가 되도록 학습이 진행되며, 신경망 기반 압축 기법에서 높은 성능을 보이고 있다.

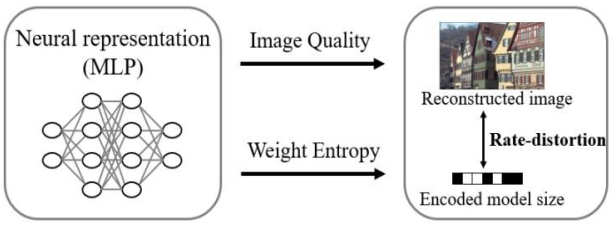

그림 1. 본 논문에서 제안하는
이미지 압축 프레임워크

그와 더불어 최근 암시적 신경 표현(INR, Implicit Neural Representation)을 통한 연구 역시 발전하고 있다. 암시적 신경 표현은 다층 퍼셉트론(MLP, MultiLayer Perceptron)에 입력 좌표와 출력 신호 값을 과적합하는 것만으로 높은 품질의 데이터를 표현할 수 있음을 보여주어 주목받고 있다 [2, 3, 4]. 암시적 신경 표현은 입력 좌표와 해당 데이터 값 사이의 매핑(예: RGB 값)을 학습하는 것으로 좌표와 관련 데이터 신호 사이의 매핑 함수를 학습하여 해당 데이터를 표현하는 것으로도 해석할 수 있다. 따라서 복원 시에 별도의 커다란 모델이 필요하지 않으며, 과적합된 MLP 가중치를 전송하는 것만으로도 복원이 가능한 장점을 갖고 있다.

본 논문에서는 이러한 점에 착안하여 암시적 신경 표현을 기반으로 한 이미지 압축 모델을 제안한다. 제안하는 모델은 암시적 신경 표현 모델에서 신경망 가중치의 엔트로피를 직접적으로 손실 함수에 포함한다. 그림 1 은 제안하는 이미지 압축 프레임워크를 나타낸다.

즉, 학습이 진행될수록 모델의 가중치의 엔트로피가 최소화되도록 학습이 진행되며, 이를 통해 산술 부호화 등의 부호화 과정을 거쳤을 때 최대한 작은 크기를 갖도록 모델을 유도할 수 있다.

암시적 신경 표현 기술을 활용한 많은 연구들이 이루어지고 있지만 해당 기술을 통해 데이터 신호를 압축하려는 방향은 아직 제한적인 범위에 머물고 있다 [5]. 본 논문에서는 기존 암시적 신경 표현을 기반으로 한 압축 기법의 성능을 크게 개선할 수 있으며, 단순한 모델 압축 방법으로 기존의 다른 어떠한 암시적 신경 표현 모델에도 별다른 처리 없이 적용가능한 장점을 가진다.

2. 관련 연구

학습 기반 이미지 압축 모델과 제안하는 방법의 기반이 되는 암시적 신경 표현 모델, 마지막으로 네트워크 가중치 압축에 대한 연구들을 서술한다.

2.1 학습 기반 심층 이미지 압축 모델

이미지 압축 모델의 목적은 잠재 벡터의 엔트로피를 최소화하는 것으로, 이는 부호화된 비트스트림의 크기와 원본에 대해 재구성된 이미지의 왜곡 간의 트레이드오프를 다음과 같이 표현할 수 있다.

$$\mathcal{L} = \lambda D + R$$

여기서 λ는 트레이드오프 파라미터로, D와 R은 각각 왜곡과 비트레이트를 나타낸다. 입력 이미지로부터 네트워크를 통해 출력된 잠재 벡터 y와, 성능 향상을 위해 일반적으로 압축 모델에서 사용되는 하이퍼인코더 모듈을 통해 출력된 추가 잠재 벡터 z가 전송 대상이 된다. 두 잠재 벡터는 산술 부호화 등 부호화를 거쳐 전송되는 것을 가정하므로, 압축 모델에서의 비트레이트와 왜곡은 다음과 같은 구성이 된다.

$$\mathcal{L} = \lambda D + R = \underbrace{\lambda d(x, \hat{x})}_{\text{Distortion}} + \underbrace{H(\hat{y}) + H(\hat{z})}_{\text{Rate}}$$

여기서 d는 PSNR 및 지각 손실(Perceptual loss)과 같은 왜곡 메트릭이고, H는 부호화된 잠재 벡터의 엔트로피를 나타낸다 [6]. [7]에서는 해당 식을 컨볼루션 네트워크 기반 종단간 모델을 통해 효과적으로 이미지를 최적화할 수 있음을 입증했다. [8]에서는 보다 발전된 형태로 가우스 혼합 모델(GMM, Gaussian mixture model)을 사용한 오토인코더 모델을 제안하였다. 다른 연구에서는 엔트로피 최적화 과정에서 이산 값에 대한 영 기울기(zero gradient) 문제를 해결하기 위한 연속화 기법을 제안하였다 [9]. [7]에 의해 제시된 모델은 공간적으로 인접한 픽셀의 정보를 보다 정확히 파악하도록 하이퍼 모듈을 포함하는 연구들로 확장되었으며 압축 성능을 개선하기 위한 다양한 연구들이 진행되고 있다 [10, 11, 12,13,14].

2.2 암시적 신경 표현 모델

암시적 신경 표현 모델에서 이미지를 예로 들면, I는 인코딩할 이미지를 나타내고 $I[x,y]$는 지정된 픽셀 위치 (x,y)에서의 RGB 값을 가정한다. 이 때 모델은 좌표를 입력받아 신호값의 데이터를 매핑하는 함수로써 이미지, 비디오, 3 차원 객체와 같은 좌표 기반 데이터를 표현할 수 있으며 다음과 같이 표현된다.

$$I : (x, y) \rightarrow (R, G, B)$$

이 매핑 함수는 일반적으로 다층 퍼셉트론(MLP)의 가중치와 함께 f(x,y)=(R,G,B)의 형태로 근사화할 수 있다. 일반적으로 암시적 신경 표현 모델의 손실 함수로 원본 이미지와 복원된 출력 이미지 사이의 평균 제곱 오차(MSE, Mean Squared Error)가 사용된다.

$$\mathcal{L}_{img} = \min_{\theta} \sum_{x,y} \| f_\theta(x, y) - I(x, y) \|_2^2$$

기존의 오토인코더 기반 이미지 압축 모델과 달리 암시적 신경 표현 모델은 네트워크 가중치 θ에 모든 이미지 정보를 인코딩한다. 따라서 암시적 신경 표현 압축 모델에서는 이미지를 재구성하기 위해 잠재 벡터가 아닌, 네트워크 가중치 자체를 전송한다. 암시적 신경 표현 모델에서 고주파 정보를 표현하는 것에 대한 어려움이 있었지만, 최근 이와 관련한 여러 연구가 제안되었다 [15,16]. 관련 연구에 따르면 이러한 어려움은 주파수(frequency) 인코딩을 통해 완화될 수 있으며, 본 논문에서는 [15]에서 제시된 sine activation 기반 위치 인코딩을 적용하였다.

2.3 네트워크 가중치 압축

본 논문에서 제안하는 모델은 암시적 신경 표현의 네트워크 모델 압축 연구로도 간주할 수 있다. 네트워크 모델 압축 연구의 대표격으로 [17]은 이미지 압축 프로세스와 유사하게 가지치기, 양자화 및 엔트로피 코딩을 거치며 각 단계 사이에 재훈련하는 기법을 제안했다. 본 논문의 모델은 신경망 가중치를 매개변수화하여 엔트로피를 최적화하는 연구에 착안하였다 [18]. 해당 연구에서는 모델 가중치의 엔트로피를 손실 함수로 제시하여 최적의 엔트로피를 갖도록 모델의 가중치를 조정한다. 따라서 학습이 진행될수록 가중치는 산술 부호화 등의 부호화를 거쳤을 때, 보다 작은 크기를 갖게 된다.

3. 제안 모델

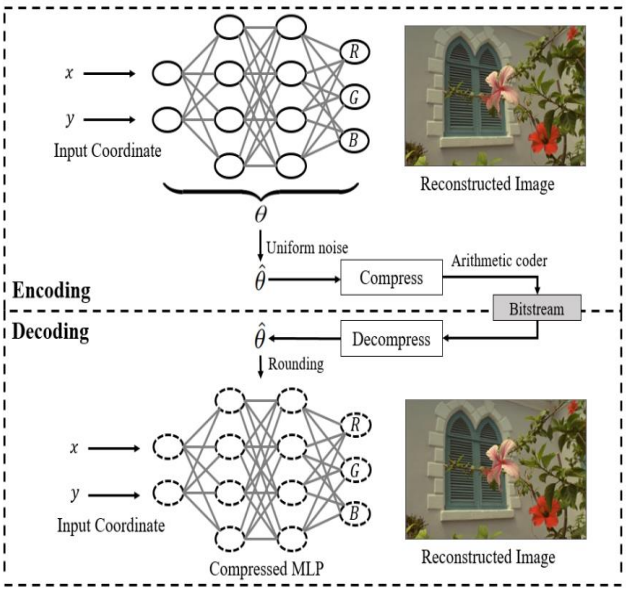

그림 2. 제안하는 모델의 압축 및 전송 구조도

그림 2 는 제안하는 모델의 구조도를 나타낸다. 학습이 완료된 후, 산술 부호화기를 이용하여 네트워크 가중치를 압축하고, 이진 비트스트림으로 전송하여 이를 다시 복호화하여 MLP 의 가중치로 사용함으로써 압축된 이미지 전송과 복원이 이루어진다. 기존의 컨볼루션 기반 압축 모델과 달리, 각 개별의 이미지에 과적합된 네트워크의 가중치를 압축하여 전송하는 형태가 된다.

네트워크 가중치는 이산적인(discrete) 데이터로 미분 값을 도출하기가 어려우므로, 학습 과정에서 확률적 경사 강하(SGD, Stochastic Gradient Descent) 방법을 이용해 가중치의 엔트로피를 최적화하기 위해서는 근사치로 대체하는 방법이 필요하다. 미분 가능한 근사치를 통한 가중치의 최적화를 위해, 우선 네트워크 모델의 엔트로피 I 는 다음과 같이 표현될 수 있다.

$$I(\theta) = -\log_2 q(\theta)$$

이산적인 데이터에서의 미분 값을 추정하고 엔트로피를 최소화하는 방향으로 학습하기 위해 [7]에서는 균등분포를 추가한(additive uniform noise) 대체함수 Φ 를 사용함으로써, 해당 대체함수가 원래의 이산 데이터 분포의 근사치로 사용될 수 있다는 것을 관찰하였다. 따라서 다음과 같이 네트워크 가중치에서 균등분포가 더해진 대체 함수 Φ 를 엔트로피 최적화에 이용한다.

$$\Phi(\theta) = q(\theta + u), u \sim \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right)$$

$\mathcal{U}$는 독립항등분포(i.i.d, independent and identically distribution)를 나타낸다.

이처럼 이산적 데이터에서 나타나는 미분하기 어려운 데이터 분포를 근사치로 대체하여 학습하고자 하는 여러 연구가 진행되었으며, 대표적으로 STE(straight through estimator) 방법이 있다 [1]. 그 중에서도 균등분포 잡음을 추가하는 방법은 딥 러닝 기반 이미지 압축에서 제안되어 널리 사용되고 있지만, 네트워크 가중치 영역이 아닌 인코딩된 잠재 벡터에 대하여 사용되었다. 본 논문의 모델은 균등분포 잡음을 네트워크 가중치 영역에 사용하여 엔트로피 손실 함수를 최적화한다. 이미지 분류 모델 등에서 해당 방법으로 모델을 압축하려는 시도를 제안했지만, 본 논문에서는 암시적 신경 표현을 직접적인 이미지 압축 모델로 활용될 수 있다는 점에 착안한다.

$$\mathcal{L}_{\text{entropy}} = \sum_{\hat{\theta} \in \Phi} I(\hat{\theta})$$

따라서 네트워크 가중치 θ의 네트워크 엔트로피 손실 함수는 위와 같이 나타낼 수 있다. 학습이 완료된 후에는 이산 데이터를 보존하고 복호화 시에 라운딩(rounding)을 통해 정량화된 가중치를 구하게 다 [7]. 따라서, 최종 목적 함수는 앞 장에서 제시한 이미지 간 왜곡 손실과 네트워크 가중치의 엔트로피로 구성할 수 있다.

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{img}}_{\text{Distortion}} + \underbrace{\lambda \mathcal{L}_{\text{entropy}}}_{\text{Rate}}$$

파라미터 λ를 조절하여 압축된 모델 크기와 모델 성능의 트레이드오프를 관찰할 수 있다. 따라서 최종 목적 함수를 통해 암시적 신경 표현 모델 에서 기존의 이미지 압축 모델과 같은 식을 최적화하는 형태로 나타낼 수 있다

4. 실험 결과 및 분석

제안하는 모델의 검증을 위해, 768×512 해상도의 24 개 이미지로 구성된 Kodak 이미지 데이터 세트에 대한 실험을 수행하였다. 제안 모델을 오토인코더 기반 압축 모델과 비교하였으며, 각각 Factorized prior [7], Hyperprior [11]로 서술한다.

기존 이미지 코덱은 JPEG, JPEG2000 및 BPG(Better Portable Graphics) 과 비교를 진행하였으며, 암시적 신경 표현 기반 이미지 압축 초기 모델인 COIN(Compression with implicit neural representations) 또한 비교 대상으로 포함하였다 [5]. 제안하는 모델은 PyTorch 에서 구현되었으며 RTX2080Ti GPU 에서 수행되었다.

$$\text{bitrate} = \frac{\text{total number of bits}}{W \times H} \quad [\text{bpp}]$$

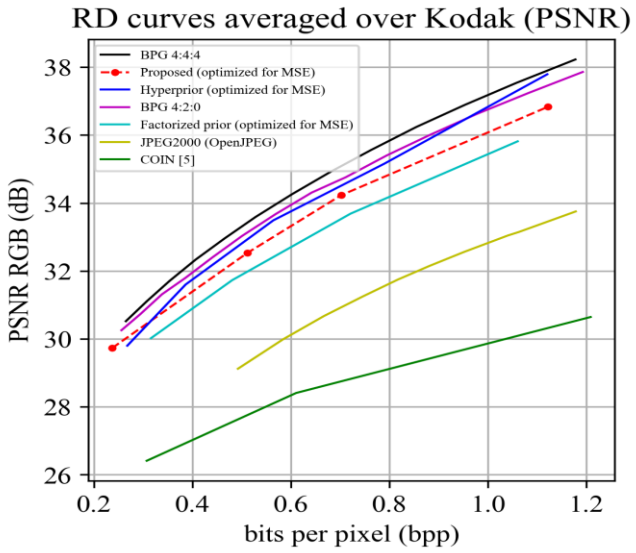

그림 3. Kodak 데이터셋에 대한 모델별 Rate-distortion(율-왜곡) 비교 (**PSNR**)

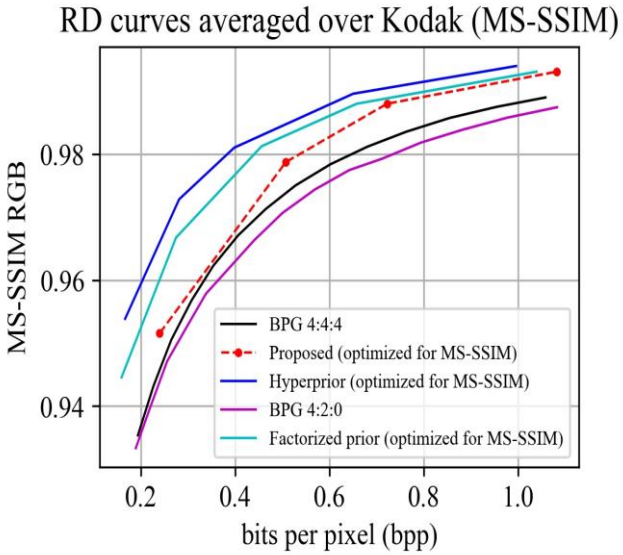

그림 4. Kodak 데이터셋에 대한 모델별 Rate-distortion(율-왜곡) 비교 (**MS-SSIM**)

표 1. 제안된 모델의 파라미터 별 성능 분석

M (hidden units)	Lambda (λ)	PSNR (dB)	Relative Entropy (%)
32	0.1	29.62	-92.6%
	0.001	30.37	-85.3%
64	0.1	32.56	-93.5%
	0.001	33.05	-86.7%
96	0.1	34.25	-94.0%
	0.001	34.72	-87.5%
128	0.1	36.12	-93.7%
	0.001	36.65	-87.2%

전체 비트레이트는 이미지 코덱 압축 성능 비교를 위하여 전체 해상도로 나눈 값인 bpp (bits per pixel)를 기준을 따라 성능 평가를 진행하였다. 암시적 신경 표현 모델과의 정확한 비교를 위해, COIN 과 동일한 학습 전략을 사용하였으며, 하이퍼파라미터 역시 동일한 값을 사용하였다 [5].

제안하는 모델은 CompressAI 라이브러리에서 제공된 *EntropyBottleneck* 클래스와 산술 부호화기를 통해 구현되었다 [19]. 압축 모델에 대한 율-왜곡 평가를 위해 다양한 지점에서 성능 평가를 진행하였으며 은닉 계층 M 의 개수, 즉 MLP 폭에 따라 해당하는 지점의 성능을 평가하였다. Kodak 이미지 데이터셋에 대하여, M = {32,64,96,128}을 사용하였다.

파라미터 λ의 트레이드오프에 따라 모델의 성능이 좌우되므로, 초기화로 시작하여 [0.1, 0.001] 범위 내에서 λ를 훈련하였다. 표 1 은 제안된 모델의 파라미터 별 성능 분석을 나타낸다. λ의 값이 높을수록, 네트워크 엔트로피가 감소하여 모델 사이즈 또한 줄어들게 되지만, 복원 성능이 크게 떨어지는 것을 관찰할 수 있다. 0.1 의 λ 파라미터 값은 엔트로피 총량을 10% 이하로 낮추어 필요 비트레이트를 크게 절약할 수 있지만 모델의 복원 성능 역시 크게 감소하는 것을 관찰할 수 있다.

따라서 λ 파라미터의 튜닝(tuning)을 통해 보다 최적화된 율-왜곡을 구성할 수 있음을 알 수 있다. 제안하는 모델에서는 낮은 비트레이트 대역인 M={32,64}에서는 0.001 의 값을, 높은 비트레이트 대역인 M={96,128}에서는 0.1 의 값을 사용하여 성능 비교를 진행하였다. 표 1 에서 나타나듯이 비트레이트 대역에 관계없이 높은 λ 값은 복원 화질을 크게 감소시키는 것을 확인할 수 있다.

그림 3 과 4 는 화질 평가 메트릭인 PSNR(Peak Signal-to-noise ratio)과 MS-SSIM(Multi-scale Structural Similarity)에 대해 각각 제안된 방법의 성능 비교를 나타낸다. 그림 3 에서 나타나듯이, PSNR 로 최적화하여 학습된 압축 모델의 경우 Kodak 이미지에 대해 오토인코더 기반 압축 모델인 Factorzed prior 보다 동일한 비트레이트에서 더 높은 복원 성능을 나타낸다. 또한 제안하는 모델은 암시적 신경 표현 모델인 COIN 보다 큰 폭으로 성능이 향상되었음을 볼 수 있다. 하지만 전통적인 이미지 코덱인 BPG 보다는 낮은 성능을 보인다.

그림 4 에서 MS-SSIM 으로 최적화된 제안 모델은 기존 오토인코더 기반 압축 모델들보다 낮은 성능을 보여준다. MS-SSIM 최적화의 경우, 학습 기반 이미지 압축 모델이 보다 우수한 성능을 나타내는 경향이 있으며 제안하는 모델 또한 BPG 보다 높은 성능을 보여주었다. 기존 오토인코더 기반 압축 모델들은 인코딩된 잠재 벡터를 이미지로 복원하기 위해서는 10MB~40MB 에 달하는 전체 모델을 먼저 전송해야 하는 단점이 있는 반면, 제안하는 모델은 학습된 이미지에 대한 가중치만을 전송하는 것만으로 독립적인 복원이 가능하다는 장점을 가지고 있다. 하지만 암시적 신경 표현 모델에서는 입력 이미지의 일반화를 고려하지 않기 때문에, 각 개별의 이미지를 과적합시켜 학습해야하는 과정이 필요하다 [5].

(a) Original (b) JPEG2000 (c) Factorized prior (d) Proposed method

그림 5. Kodim24 이미지에 대한 PSNR 최적화 모델별 주관적 화질 비교 (**0.5bpp 대비**)

5. 결론

본 논문에서는 최근 부상하고 있는 암시적 신경 표현 모델을 이미지 압축 방법론에 적용하고, 실험 결과를 논의하였다. 기존의 오토인코더 기반 모델과 달리, 제안하는 방법은 단일 네트워크의 가중치만을 필요로 하므로 모델 경량화 등 다양한 시나리오에서 적용될 수 있는 잠재력을 가지고 있다. 본 논문에서 도입한 엔트로피 손실 함수를 통해 모델의 가중치를 직접적으로 압축할 수 있으며, 제안하는 방법이 기존 학습 기반 압축 모델과 비슷한 성능을 나타냄을 보였다. 특히, 제안하는 방법은 다른 암시적 신경 표현 모델에 추가적인 처리 없이 손쉽게 적용될 수 있다는 장점을 가지며 이를 통해 암시적 신경 표현 모델을 활용한 압축 분야에 대한 논의가 더욱 활발해질 것으로 기대된다.

감사의 글

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (No.2020-0-00920-002, (세부 2)중대형 공간용 초고해상도 비정형 플렌옵틱 영상 저장/압축/전송 기술 개발)

참고문헌

[1] Johannes Ballé, Philip A Chou, David Minnen, Saurabh Singh, Nick Johnston, Eirikur Agustsson, Sung Jin Hwang, and George Toderici, “Nonlinear transform coding,” IEEE Journal of Selected Topics in Signal Processing, vol. 15, no. 2, pp. 339–353, 2020.

[2] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P. Srinivasan, Jonathan T. Barron, and Ren Ng, “Learned initializations for optimizing coordinate-based neural representations,” 2021.

[3] Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov, “Image generators with conditionally-independent pixel synthesis,” 2020.

[4] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” 2020.

[5] Emilien Dupont, Adam Golinski, Milad Alizadeh, Yee Whye Teh, and Arnaud Doucet, “Coin: Compression with implicit neural representations,” 2021.

[6] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in European conference on computer vision. Springer, 2016, pp. 694–711.

[7] Johannes Ballé, Valero Laparra, and Eero P Simoncelli, “End-to-end optimized image compression,” arXiv preprint arXiv:1611.01704, 2016.

[8] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár, “Lossy image compression with compressive autoencoders,” arXiv preprint arXiv:1703.00395, 2017.

[9] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc Van Gool, “Soft-to-hard vector quantization for end-to-end learned compression of images and neural networks,” arXiv preprint arXiv:1704.00648, vol. 3, 2017.

[10] David Minnen, Johannes Ballé, and George D Toderici, “Joint autoregressive and hierarchical priors for learned image compression,” Advances in Neural Information Processing Systems, vol. 31, pp. 10771– 10780, 2018

[11] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, “Variational image compression with a scale hyperprior,” arXiv preprint arXiv:1802.01436, 2018.

[12] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack, “Contextadaptive entropy model for end-to-end optimized image compression,” arXiv preprint arXiv:1809.10452, 2018.

[13] Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson, “High-fidelity generative image compression,” 2020.

[14] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool, “Generative adversarial networks for extreme learned image compression,” 2019.

[15] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein, “Implicit neural representations with periodic activation functions,” 2020.

[16] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng, “Fourier features let networks learn high frequency functions in low dimensional domains,” 2020.

[17] Song Han, Huizi Mao, and William J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding,” 2016.

[18] Deniz Oktay, Johannes Ballé, Saurabh Singh, and Abhinav Shrivastava, “Scalable model compression by entropy penalized reparameterization,” 2020.

[19] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja, “Compressai: a pytorch library and evaluation platform for end-to-end compression research,” arXiv preprint arXiv:2011.03029, 2020.