NeRF 계열 모델의 경량화를 위한 양자화 고려 훈련 기법

하유진 [1], 방건 [2], 류은석 [3]
성균관대학교 컴퓨터교육과 [1], 한국전자통신연구원(ETRI)[2], 성균관대학교 실감미디어공학과 [3]
uj0110@g.skku.edu, gbang@etri.re.kr, esryu@skku.edu

Applying Quantization-Aware Training (QAT) for Lightweighting NeRF-Based Models

Yoojin Ha[1], Gun Bang[2], Eun-Seok Ryu[3]
Department of Computer Education, Sungkyunkwan University,
Electronics and Telecommunications Research Institute(ETRI),
Department of Immersive Media Engineering, Sungkyunkwan University

요 약

본 논문은 최근 주목 받고 있는 3 차원 재구성 기술의 증가하는 수요와 관련하여, NeRF(Neural Radiance Fields)를 경량화하고 화질을 유지하는 방법을 제안한다. NeRF 는 3 차원 재구성을 위한 심층 신경망 기술로, 렌더링 시간이 길어지는 문제를 가진다. 이에 양자화를 고려한 훈련(Quantization-Aware Training, QAT) 방법을 NeRF 에 적용하여 경량화하고 렌더링 화질을 유지한다. 특히 NeRF 의 특성에 따라, ReLU 와 선형 레이어를 융합 후 양자화 함으로써 렌더링 성능을 향상 시킬 수 있었다. 따라서 제안하는 방법을 통해 NeRF 의 양자화를 수행하면 렌더링 화질을 유지하면서 모델을 효율적으로 경량화 할 수 있다.

1. 서론

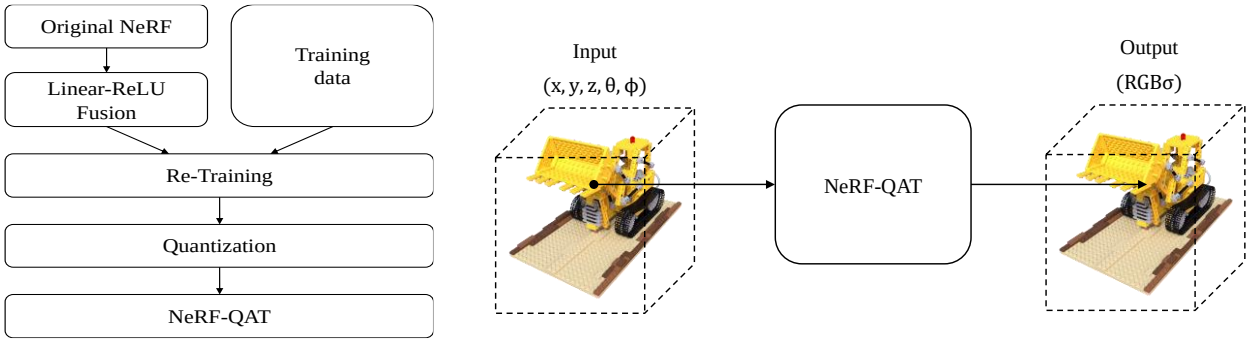
최근, 헬스케어, 제조, 미디어 등 다양한 분야에서 3 차원 재구성(3D reconstruction) 기술의 수요가 급증하고 있다. 이러한 분야에서 3 차원 재구성은 더 높은 화질과 효율성을 요구하며, 그에 따라 심층 신경망 기술인 NeRF(Neural Radiance Fields)[1]가 주목받고 있다.

NeRF 는 제한된 시점의 입력 이미지를 바탕으로 3 차원 장면의 새로운 시점 이미지를 만들어 내는 것을 목표로 한다. 이는 완전 연결 심층 네트워크(fully-connected deep network)를 이용하며, 입력으로 공간 위치와 시점 방향을 받고, 해당 위치에서의 색상과 부피 밀도를 출력한다. 출력된 색상과 밀도를 이미지로 투영하기 위해 고전적인 부피 렌더링 기술을 사용한다. 그러나 NeRF 는 높은 복잡도로 인해 렌더링 시간이 오래걸리는 문제가 있다. 이에 따라 렌더링 속도와 모델의 경량화에 대한 연구가 활발히 진행 중인 상황이다.

양자화는 모델의 실행 성능과 효율성을 향상하기 위한 기술이다. 가중치와 활성화 함수 출력을 더 작은 비트 수로 표현하도록 변환하여 모델의 크기를 축소하고 계산 속도를 높이며 메모리 사용량을 줄여 모델의 효율적인 배포와 실행을 가능하게 한다. 양자화 방법 중에는 양자화를 고려한 훈련(Quantization-Aware Training, QAT)[2] 이라는 기법이 있다. 이는 훈련 중에 가중치 양자화를 고려하여 모델을 조정하며, 이를 통해 원본 모델을 양자화에 강건하게 만들고, 양자화로 인한 기존의 정확도 손실 문제를 개선할 수 있다.

같은 양자화 방법을 사용하더라도 모든 딥러닝 모델에서 동일한 효과가 나타나는 것은 아니며, 레이어 특성에 따라 성능이 달라진다. 따라서 실험과 조정을 통해 최적의 설정을 찾는 과정이 필요하다.

본 논문은 양자화를 고려한 훈련 방법을 NeRF 에 적용하여 이를 경량화하고, 동시에 렌더링 화질을 유지하는 방법을 제안한다. 이를 통해 3 차원 재구성 분야에서 양자화 기술을 효과적으로 적용할 수 있는 방안을 제시하고 있다.



(a) NeRF 에서 양자화를 고려한 훈련 과정 (b) 제안된 모델을 통한 장면 표현의 구조도

그림 1. 제안된 모델의 구조

2. 관련 연구

NeRF 경량화와 관련한 최신 연구들로는 Voxel Pruning 을 사용하는 방법[3], Tensor Decomposition 을 사용하는 방법[4], 그리고 해싱 인덱스를 사용하는 방법[5]이 있다.

Plenoxel[3]은 NeRF 모델의 메모리 효율을 향상하기 위해 Voxel Pruning 기술을 사용한다. Voxel Pruning 은 3D 공간을 격자로 나누고, 불필요한 빈 격자를 제거하여 메모리를 절약한다. 모델 크기와 메모리 사용량을 감소시킴으로써 학습 속도와 렌더링 성능을 향상시켰다. 그러나 이 방법은 효율성과 성능 사이의 트레이드오프가 있다.

TensoRF[4]은 Tensor Decomposition 을 활용하여 NeRF 모델을 경량화한다. Tensor Decomposition 은 복잡한 텐서를 여러 부분 텐서로 분해하여 모델 크기를 줄이면서도 원래 모델과 유사한 렌더링 성능을 얻을 수 있다.

Instant-ngp[5]는 해싱 인덱스를 활용하는 연구이다. 해싱을 사용함으로써 렌더링 된 이미지와 관련된 3D 포인트를 효율적으로 검색할 수 있어 학습 및 추론 시간을 단축한다.

이러한 방법들은 높은 렌더링 성능과 학습 속도를 가지지만 메모리 면에서 효율성이 떨어진다. 이에 본 연구는 메모리 문제를 극복하기 위해 양자화를 고려한 훈련 방법을 제안한다. 양자화를 통해 모델 크기와 메모리 사용량을 줄이면서도 렌더링 성능을 유지할 수 있다.

3. NeRF 경량화를 위한 양자화 고려 훈련 적용 방안

제안하는 기법은 그림 1 과 같이 양자화 고려 훈련 방법을 NeRF 에 적용한다. 전반적인 프레임워크는 (1) 양자화 적용 범위 설정 후, (2) 훈련을 진행하고, (3) 훈련 완료한 모델의 가중치를 양자화 하여 저장한다. 이후 (4) 렌더링 시에 저장한 모델을 불러와 사용한다. 그렇게 렌더링 된 이미지를 성능 평가에 사용하였다.

NeRF 는 완전 연결 레이어와 그에 따르는 ReLU (Rectified Linear Unit) 함수로 이루어져 있다.[1] 일반적으로 양자화 과정에서 모든 레이어가 양자화 되는 것이 반드시 효율적인 것은 아니며, ReLU 의 가중치는 선형 레이어와 함께 레이어 융합(Layer Fusion)되어 양자화에 적용될 수 있다.[6] 그러나 레이어 융합의 효과는 모델에 따라 다르게 나타나기 때문에, 해당 모델에서의 성능을 실험을 통해 확인하는 것이 필요하다.

실험을 통해 선형 레이어와 ReLU 의 융합에 따른 양자화 효과를 확인했으며, 그 결과 레이어 융합을 했을 때 융합하지 않았을 때보다 성능이 좋았음을 확인할 수 있었다. 자세한 실험 결과는 4.2 절에 정리하였다. 이러한 실험 결과에 따라 ReLU 를 각 선형 레이어에 융합하고, 전체 레이어에 대해서 양자화를 적용하여 진행하였다.

4. 실험 결과

4.1 데이터셋

데이터셋은 실제 세계에 대한 전방 촬영 이미지로, Local Light Field Fusion[7] 논문에서 가져온 fern 이미지를 활용하였다. fern 은 20 개의 촬영된 이미지로 이루어져 있으며, 8 프레임 당 1 개를 테스트 데이터로 활용하였다.

4.2 레이어 융합의 효과 및 렌더링 성능 확인

선형 레이어 및 ReLU 의 레이어 융합의 효과를 확인하기 위해 양자화 상황을 가정하고, 레이어 융합하지 않은 모델과 레이어 융합했을 때의 모델을 비교하였다. 평가 지표로는 PSNR(Peak Signal-to-noise Ratio)[8]을 활용하였다.

기준 이미지 레이어 융합 있음 레이어 융합 없음

그림 2. 레이어 융합에 따른 이미지 렌더링 결과

표 1. 레이어 융합 유무에 따른 이미지의 PSNR 값 비교

	레이어 융합 있음	레이어 융합 없음
PSNR ↑	**30.793**	24.587

표 2. 양자화 고려 훈련 적용 유무에 따른 렌더링 성능 비교

	PSNR ↑	SSIM ↑	LPIPS ↓	Size ↓
NeRF[1]	31.49	0.899	0.050	4.78MB
Plenoxel[3]	31.71	0.958	0.049	778MB
TensoRF[4]	33.14	0.963	0.047	71.8MB
Instant-ngp[5]	**33.18**	**0.963**	**0.045**	64.1MB
NeRF-QAT	30.56	0.819	0.114	**1.33 MB**

실험 결과, 레이어 융합이 있는 경우 PSNR 이 더 높게 나왔으며(표 1), 레이어 융합하지 않은 경우 Blur 현상이 심하게 일어나는 것이 확인되었다.(그림 2).

모델의 렌더링 성능 확인을 위해 PSNR, SSIM(Structural Similarity Index Measure)[8] 그리고, LPIPS(Learned Perceptual Image Patch)[9] 을 측정한다. 렌더링 성능은 기존 NeRF 와 선행 연구인 Plenoxel, TensoRF, Instant-ngp, 그리고 제안된 방법인 양자화 고려 훈련을 적용한 모델(NeRF-QAT)을 비교하였다.

양자화 고려 훈련을 적용한 경우 기존 NeRF 에 비해 세 가지 성능 평가 지표가 소폭 감소하였으나(표 2), 모델 크기 면에서 약 4 배의 이득이 있음을 확인할 수 있었다. 또한 다른 세 가지 모델에 비해서도 모델 크기가 많이 감소하였다.

5. 결론

본 논문은 양자화를 고려한 훈련 방법을 NeRF 계열 모델에 적합한 방법으로 적용하여 모델의 성능을 최대한 유지하면서 모델 크기를 줄여 그 효과를 확인한다.

실험 결과, NeRF 의 ReLU 를 선형 레이어에 융합했을 때 양자화 시의 손실을 최소화할 수 있었다. 또한 렌더링 성능 평가 결과, 양자화로 인해 성능이 소폭 감소함에 비해 모델 크기를 약 4 배 줄였고, 이를 통해 효과적으로 NeRF 의 렌더링 효율을 향상할 수 있었다.

논문 사사

이 논문은 2023 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00981-002, 전배경 정합 3D 객체 스트리밍 기술개발) 및 (No. 2017-0-00072, 초실감 테라미디어를 위한 AV 부호화 및 LF 미디어 원천기술 개발)

참고 문헌

[1] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM*65.1 (2021): 99-106.

[2] Gholami, Amir, et al. "A survey of quantization methods for efficient neural network inference." *Low-Power Computer Vision*. Chapman and Hall/CRC, 2022. 291-326.

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

[4] Chen, Anpei, et al. "Tensorf: Tensorial radiance fields." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

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

[6] Krishnamoorthi, Raghuraman. "Quantizing deep convolutional networks for efficient inference: A whitepaper." *arXiv preprint arXiv:1806.08342* (2018).

[7] Mildenhall, Ben, et al. "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines." *ACM Transactions on Graphics (TOG)* 38.4 (2019): 1-14

[8] Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity." *IEEE transactions on image processing* 13.4 (2004): 600-612.

[9] Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.