Homework #4

Deep Learning for Computer Vision

NTU, Fall 2023

周奕節
r10943131

1. (15%) Please explain:

   a. the NeRF idea in your own words

   - Neural Radiance Fields is a novel view synthesis task. It solves the synthesis problem by optimizing parameters of a continuous 5D scene representation.
   - Representation: The scene is represented as a 5D function that takes location(x, y, z) and viewing direction(θ, φ) as input and outputs color(r,g,b) and volume density(σ).
   - Pipeline
     - Feed (x, d) into MLP to get (c, σ)
     - Generate predicted images by applying classical volume rendering techniques to render the color of an ray passing through every pixel.
     - Since the rendering function is differentiable, optimize the representation by minimizing the MSE loss between the predicted images and the GT images.
   - Additional techniques
     - View dependent: Design the MLP to predict σ only as a function of x but not d, so we can achieve multiview consistent.
     - Positional Encoding: Mapping the input to a higher dimensional space using high frequency functions can represent high-frequency variation in color and geometry better.
     - Hierarchical Volume Sampling: Simultaneously optimize two networks: coarse and fine, so we can sample the region that are more important.

b. which part of NeRF do you think is the most important

- Represent a scene as a 5D coordinate and allow a camera ray to go through every pixel, so we can render them to get the color and density of each pixel.

c. compare NeRF’s pros/cons w.r.t. other novel view synthesis work

NeRF vs. DVGO vs. Instant

- NeRF
    - Pros:
        - Good quality
        - Doesn’t require 3D model as ground truth
        - View-dependent
    - Cons:
        - Only fits on one specific scene
        - Time consuming to train and inference on a scene

  Many paper after NeRF are focusing on speeding up the whole process. Previous works have shown good quality with a large inference time speed up, while pre-training MLP is still required.

- DVGO: replacing the MLP with voxel grid
    - Reduce training time from 10-20 hours to 15 minutes (on their single GPU)
    - Achieved visual quality comparable to NeRF at a rendering speed about 45x faster
    - Grid resolution is about 160^3, while previous works range from 512 to 1300^3
- Instant NeRF: replace MLP with Hash

  - Trained in real time(a few seconds)
  - Render in tens of milliseconds at resolution of 1920x1080

2. (15%) Describe the implementation details of ***your NeRF model*** for the given dataset. You need to explain your ideas completely.

   Most of the implementation details are the same as the original setting, but I increased the sampling number of the fine network to 192.

   - Positional Encoding
     - xyz: L=10
     - dir: L=4
   - Sample a batch of camera rays for each iteration
     - batch_size = 1024
   - Hierarchical sampling
     - coarse network: N_samples = 64
     - fine network: N_importance = 192
   - Loss: MSE between rendered and true pixels colors for both coarse and fine renderings
   - Optimization
     - Opimizer: Adam
     - Learning rate = 5e-4, momentum =0.9
     - Scheduler = stepLR, decay_step = 20, gamma = 0.1

3. (15%) Given novel view camera pose from **metadata.json**, your model should render novel view images. Please evaluate your generated images and ground truth images with the following three metrics (mentioned in the NeRF paper). Try to use at least **three** different hyperparameter settings and discuss/analyze the results.

- **Please report the PSNR/SSIM/LPIPS on the validation set.**

Setting	Configuration	PSNR	SSIM	LPIPS
1	coarse=64, fine=192	39.4	0.986	0.167
2	coarse=64, fine=128	38.8	0.983	0.175
3	coarse=32, fine=128	37.9	0.982	0.183
4	coarse=16, fine=128	35.3	0.972	0.215

- **You also need to explain the meaning of these metrics.**
    - **PSNR:**
        - Peak Signal-to-Noise Ratio(PSNR) calculates the ratio of maximum pixel value to noise, the noise is computed by mean squared error(MSE), higer PSNR means better image quality. Though PSNR calculates pixel-wise error, it is not a good match for human vision.
    - **SSIM:**
        - Structural Similarity Index Measure(SSIM) measures the structural similarity between two images, higher SSIM represents higher image similarity. It takes into account luminance, contrast and structure, aiming to mimic human perception of image quality. This metric provides more insights in the overall structure of the two images.
    - **LPIPS:**
        - Learned Perceptual Image Patch Similarity(LPIPS)

measures perceptual differences between two images by using a neural network to learn perceptual features. It aims to capture aspects of human perception that traditional metrics can’t well-represent. Lower LPIPS means greater perceptual similarity between images.

- **Different configuration settings such as MLP and embedding size, etc.**
    - For configuration, I tried different hierarchical sampling settings. Setting1 has a larger fine sampling number(192) than the default(128), hence getting better performance. Setting2-4 I changed the coarse sampling number from 64, 32 to 16, and the result decreased as expected.

4. (15%) With your trained NeRF, please implement depth rendering in your own way and visualize your results.