CS 231n: NeRF course notes

Yu-Ann Wang Madan
email yuann@stanford.edu

June 2021

Contents

1 Tl;dr What is NeRF and what does it do

NeRF stands for Neural Radiance Fields. It solves for view interpolation, which is taking a set of input views (in this case a sparse set) and synthesizing novel views of the same scene. Current RGB volume rendering models are great for optimization, but require extensive storage space (1-10GB). One side benefit of NeRF is the weights generated from the neural network are ∼6000 less in size than the original images.

2 Helpful Terminology

Rasterization: Computer graphics use this technique to display a 3D object on a 2D screen. Objects on the screen are created from virtual triangles/polygons to create 3D models of the objects. Computers convert these triangles into pixels, which are assigned a color. Overall, this is a computationally intensive process.
Ray Tracing: In the real world, the 3D objects we see are illuminated by light. Light may be blocked, reflected, or refracted. Ray tracing captures those effects. It is also computationally intensive, but creates more realistic effects. **Ray**: A

ray is a line connected from the camera center, determined by camera position parameters, in a particular direction determined by the camera angle.
NeRF uses ray tracing rather than rasterization for its models.
Neural Rendering As of 2020/2021, this terminology is used when a neural network is a black box that models the geometry of the world and a graphics engine renders it. Other terms commonly used are *scene representations*, and less frequently, *implicit representations*. In this case, the neural network is just a flexible function approximator and the rendering machine does not learn at all.

3 Approach

A continuous scene is represented as a 3D location $x = (x, y, z)$ and 2D viewing direction (θ, ϕ) whose output is an emitted color c = (r, g, b) and volume density σ. The density at each point acts like a differential opacity controlling how much radiance is accumulated in a ray passing through point x. In other words, an opaque surface will have a density of ∞ while a transparent surface would have $\sigma = 0$. In layman terms, the neural network is a black box that will repeatedly ask what is the color and what is the density at this point, and it will provide responses such as "red, dense."
This neural network is wrapped into volumetric ray tracing where you start with the back of the ray (furthest from you) and walk closer to you, querying the color and density. The equation for expected color $C(r)$ of a camera ray $r(t) = o+td$ with near and far bounds t_n and t_f is calculated using the following:

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t), d)\, dt$$

where

$$T(t) = exp(-\int_{t_n}^{t} \sigma(r(s))\, ds)$$

(1)
To actually calculate this, the authors used a stratified sampling approach where they partition $[t_n, t_f]$ into N evenly spaced bins and then drew one sample uniformly from each bin:

$$\hat{C}(r) = \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i\delta_i))c_i, where T_i = exp(-\sum_{j=1}^{i-1} \sigma_j\delta_j) \tag{2}$$

Where $\delta_i = t_{i+1} - t_i$ is the distance between adjacent samples. The volume rendering is differentiable. You can then train the model by minimizing rendering loss.

$$min_\theta \sum_i \|render_i(F_\Theta - I_i\|^2 \tag{3}$$

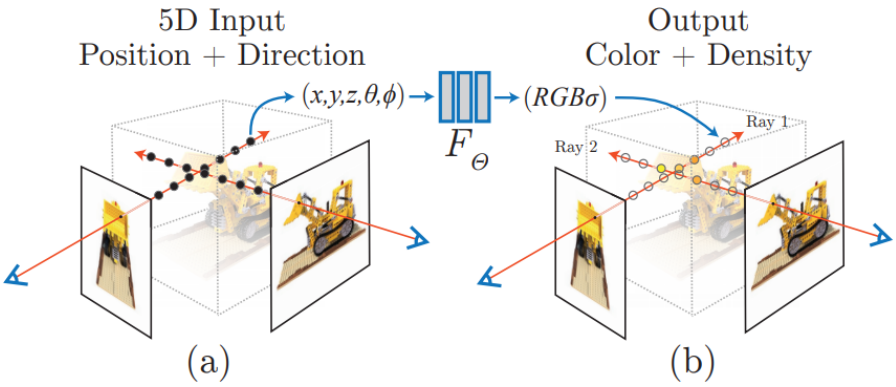

Figure 1: In this illustration taken from the paper, the five variables are fed into the MLP to produce color and volume density. F_Θ has 9 layers, 256 channels

In practice, the Cartesian coordinates are expressed as vector d. You can approximate this representation through MLP with $F_\Theta = (x, d) \rightarrow (c, \sigma)$.
Why does NeRF use MLP rather than CNN? Multilayer perceptron (MLP) is a feed forward neural network. The model doesn't need to conserve every feature, therefore a CNN is not necessary.

4 Common issues and mitigation

The naive implementation of a neural radiance field creates blurry results. To fix this, the 5D coordinates are transformed into positional encoding (terminology borrowed from transformer literature). F_Θ is a composition of two formulas: $F_\Theta = F'_\Theta \cdot \gamma$ which significantly improves performance.

$$\gamma(p) = (sin(2^0\pi p), cos(2^0\pi p), ..., sin(2^{L-1}\pi p), cos(2^{L-1}\pi p)) \tag{4}$$

L determines how many levels there are in the positional encoding and it is used for regularizing NeRF (low L = smooth). This is also known as a Fourier feature, and it turns your MLP into an interpolation tool. Another way of looking at this is your Fourier feature based neural network is just a tiny look up table with extremely high resolution. Here is an example of applying Fourier feature to your code:

```
B = SCALE * np.random.normal(shape = (input_dims, NUM_FEATURES))
x = np.concatenate([np.sin(x @ B), np.cos(x @ B)], axis = -1)
x = nn.Dense(x, features = 256)
```

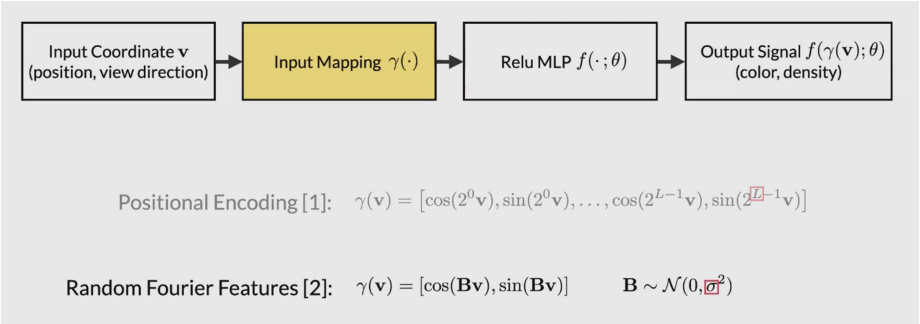

Figure 2: Mapping how Fourier features are related to NeRF's positional encoding. Taken from Jon Barron's CS 231n talk in Spring 2021

NeRF also uses hierarchical volume sampling: coarse sampling and the fine network. This allows NeRF to more efficiently run their model and deprioritize areas of the camera ray where there is free space and occlusion. The coarse network uses N_c sample points to evaluate the expected color of the ray with the stratified sampling. Based on these results, they bias the samples towards more relevant parts of the volume.

$$\hat{C}_c(r) = \sum_{i=1}^{N_c} w_i c_i, w_i = T_i(1 - exp(-\sigma_i \delta_i)) \tag{5}$$

A second set of N_f locations are sampled from this distribution using inverse transform sampling. This method allocates more samples to regions where we expect visual content.

5 Results

The paper goes in depth on quantitative measures of the results, which NeRF outperforms existing models. A visual assessment is shared below:

6 Additional references

What's the difference between ray tracing and rasterization? Self explanatory title, excellent write-up helping reader differentiate between two concepts.
Matthew Tancik NeRF ECCV 2020 Oral Videos showcasing NeRF produced images.
NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis Simple and alternative explanation for NeRF.
NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis arxiv paper
CS 231n Spring 2021 Jon Barron Guest Lecture

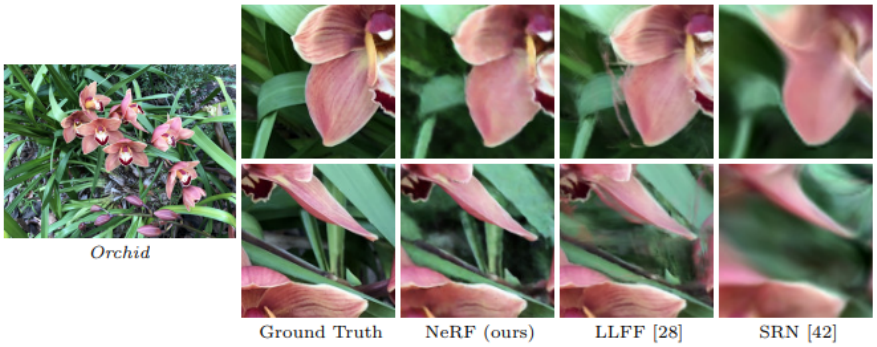

Figure 3: Example of NeRF results versus existing SOTA results