Accelerating NeRFs: Optimizing Neural Radiance Fields with Specialized Hardware Architectures

6.5930 Project

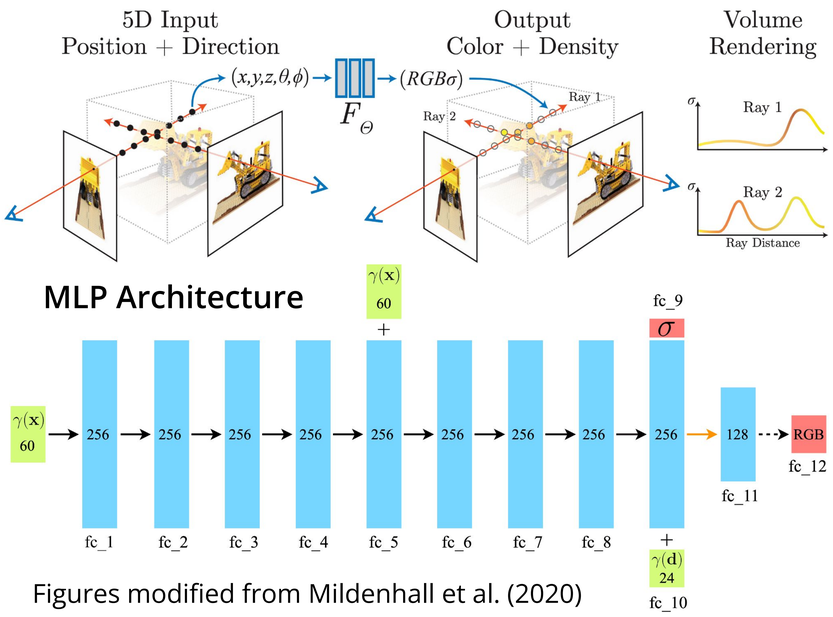
William Shen*, Willie McClinton*

Website

Motivation

- Neural Radiance Fields (NeRFs) are slow 🐌 and require modern GPUs with substantial VRAM
- Limited work on hardware acceleration potential (only 5 papers, 2 within the last month)
- NeRFs model 3D scenes from just 2D images, acceleration enables rendering on AR/VR 🥽, digital assets in game engines, and use in visual effects 🎬

NeRF Background



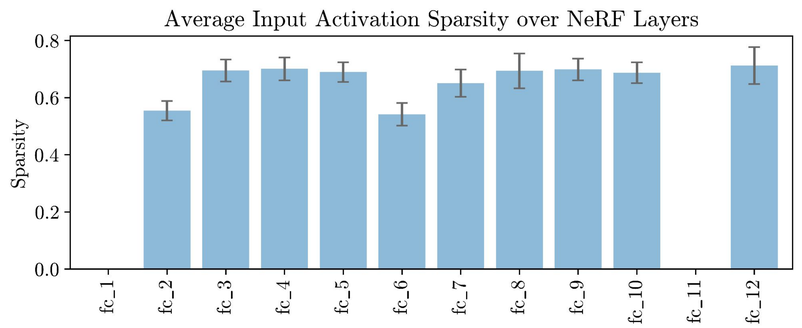
Figures modified from Mildenhall et al. (2020)

Technical Contributions

- **Challenge**: rendering a single image requires over 400x400x128 ≈ 20 million ray samples
- 🔎 In-depth profile of NeRFs to identify bottlenecks and areas for hardware acceleration
  - Novel workload, profile over 8 scenes in benchmark
- Existing work focus on algorithmic advances at software level. Compared to hardware advances, we target MLP NeRFs and exploit sparsity

Exploiting Activation Sparsity

- **65.8%** input activation sparsity across FC layers due to ReLU activation function, 0% weight sparsity

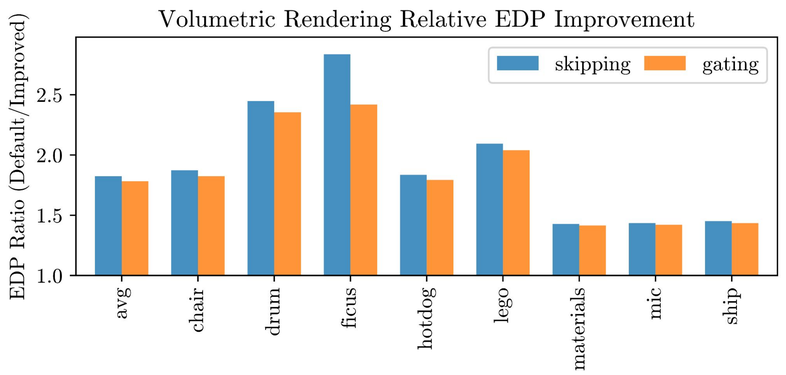


- Exploit using **Eyeriss** with no (baseline) and multiple ablations of sparse optimizations
- Map FC layers to 1x1 convolutions

	Energy (μJ)	Cycles	EDP (J $*$ cycle)	Area (mm^2)
Eyeriss	650.93	568192	3.70e8	16.51
w/ Compression only	416.01	568192	2.36e8	10.89
w/ Skipping only	386.28	307616	1.19e8	10.89
w/ Gating only	336.59	307616	1.04e8	10.25
w/ Skipping	316.71	227813	7.22e7	10.89
w/ Gating	313.05	307611	9.63e7	10.25

Accelerating Volumetric Rendering

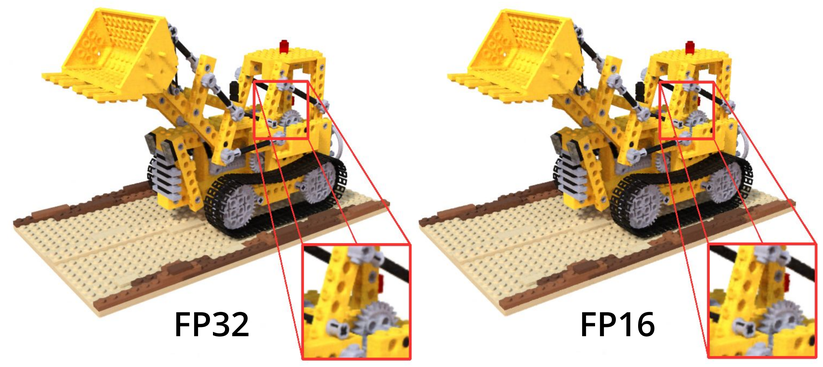
- **33.8%** average sparsity of density-based weights
- **Volumetric Rendering** = weighted sum = dot product = 1x1 convolution with vector length as filter size
- Use Eyeriss with sparse optimizations



Quantization to FP16

- Post-training FP16 quantization with no fine tuning on a NVIDIA RTX 3090 GPU
- 🏎️💨 **2.7x** improvement in rendering speed (sec)
- 🔋 **3.1x** improvement in energy consumption (kJ)
- 📉 **2.7%** decrease in Peak signal-to-noise ratio

Can you tell the difference?



Key Insights and Takeaways

- Understanding your workload is key to deciding which components to accelerate.
- Exploiting activation sparsity leads to >50% improvements in energy and cycles
- Results from accelerating volumetric rendering are correlated with the scene-dependent sparsity
- Quantization can significantly improve speed and energy with minimal loss in quality.
- **Future work:** custom architecture for NeRF MLPs, accelerate positional encoding, explore different NeRF models (e.g., grid-based), INT8 quantization

https://williamshen-nz.github.io/accelerating-nerfs