Deep Learning for Computer Vision HW4

Pin-Han, Huang: R10946003

December 14, 2022

1 Problem1

1. **NERF**

   - The goal of NERF is to learn a 3D representation and perform novel view synthesis. NERF uses a fully connected network whose input is a 5D coordinate, including 3D spatial location image views and two viewing directions of camera poses. Volume rendering is applied in NERF to project output densities and colors. The idea is resulted from the amount of light blocked along the camera ray. NERF also applies positional encoding, enabling MLP to represent higher frequency functions to alleviate low-resolution representation and inefficiency.
   - The most important part in NERF is the differentiable rendering procedure; however, positional encoding is indispensable since training network directly on the 5D-input coordinates results in

renderings performing poorly at representing high-frequency functions.

- In comparison with previews methods, the advantages of NERF is the ability to recover fine details in both geometry and appearance across rendered views with merely images as MLP input. However, NERF takes a long time to train and converge, which training time lasts from hours to days.

2. **Implementation of DVGO**

- The main problem DVGO aims to solve from NERF is the lengthy training time ranges from hours to days. DVGO uses a density voxel grid for scene geometry and a feature voxel grid with a shallow MLP for complex view dependent appearance. The key to speed up the training process is to optimize the volume density modeled in a voxel grid. There are two vital features in Direct Voxel Grid Optimization: Post-activation interpolation is applied for all non-linear activation after trilinear interpolation, which is capable of producing sharp decision boundary. Another important features is setting priors to robustify the optimization process. The first prior is the low-density initialization; that is, setting the initial grid values to zero and bias to a log formula to alleviate getting trapped into suboptimal geometry. The second prior is setting different learning rates for different grid points. Each grid point is counted by the number of training views, the base learning rate is then scaled. The second prior is used to mitigate the

condition where some voxels are visible to too few training views in capturing.

3. **Experiment with DVGO**

- The performance of experiment setting with number of iterations=20000, learning rate density=0.1, number of voxels=1024000 is shown as follows:

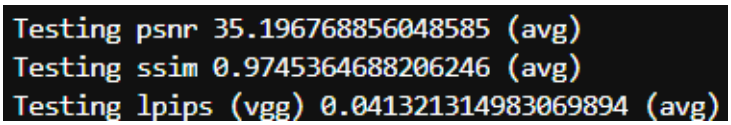

Figure 1: PSNR=35.19, SSIM=0.974, LPIPS(VGG)=0.041

- The performance of experiment setting with number of iterations=20000, learning rate density=0.01, number of voxels=1024000 is shown as follows:

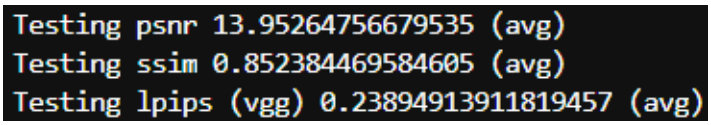

Figure 2: PSNR=13.952, SSIM=0.852, LPIPS(VGG)=0.239

- PSNR: (Peak Signal-to-Noise Ratio)-An engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. PSNR is mostly easily defined via MSE.

- SSIM:(Structural Similarity)-SSIM is a method for predicting the perceived quality of digital images and videos and is used for measuring the similarity between two images.
- LPIPS: (Learned Perceptual Image Patch Similarity)-LPIPS is used to judge the perceptual similarity between two images. LPIPS essentially computes the similarity between the activations of two image patches for some pre-defined network. This measure has been shown to match human perception well. A low LPIPS score means that image patches are perceptual similar.

2 Problem2

1. **Implementation of BYOL method**
    - BYOL is a self-supervised learning method which does not require contrastive learning and negative pairs. Data augmentation is applied to attract augmented images from the same original image to be close to each other. Without negative pairs, BYOL introduces a predictor to avoid mode collapse. In addition, exponential moving average is used to gradually update the momentum encoder.
    - Resnet50 is pretrained on mini imagenet dataset with BYOL. In terms of experiment setting, epochs=500 (with early stop patience 20 epochs), learning rate is $3e-4$, batch size is 256, image size is 128.
2. **Report 3 other settings of experiment.**

- The results of different experiment settings are as follows.

	Experiment Setting		Validation Accuracy
A	-	Train full model (backbone + classifier)	29.064
B	w/ label (TAs have provided this backbone)	Train full model (backbone + classifier)	25.616
C	w/o label (Your SSL pre-trained backbone)	Train full model (backbone + classifier)	43.842
D	w/ label (TAs have provided this backbone)	Fix the backbone. Train classifier only	25.369
E	w/o label (Your SSL pre-trained backbone)	Fix the backbone. Train classifier only	30.296

Figure 3: Experiment results of different settings. Accuracy is used as our metric.

- Setting C has the highest accuracy on office dataset, 43.842. In comparison with setting A where no pretrained backbone is used, the accuracy increases 14%. In terms of setting D and E, the pretrained backbone using BYOL with resnet50 backbone fixed surpasses TAs pretrained backbone.