Title: Fine-tuning Dense Prediction Vision Transformers for Image Restoration

Motivation

Ever since transformers proved their usefulness in NLP, they have been continuously adopted in a variety of fields, and computer vision is no difference. Vision transformers have been seen to quickly overtake convolutions in numerous different tasks, and are the fan favourite model nowadays. Lots of work has gone into adapting them for more complicated imaging tasks like segmentation, which are known as “Dense Prediction” due to how a prediction is required for each pixel of the image. A model trained to perform dense prediction will output a segmentation mask, a mask the same size as the input image with one channel which contains the prediction for if the pixel is a part of the object or not. To perform segmentation accurately, the models must have a good understanding of what is captured in the image.

Image restoration covers a pretty broad range of tasks, but I’ll focus on denoising and deblurring since those are broadly applicable. In denoising and deblurring, you have a corrupted version of an image, and would like to recover the original uncorrupted version, essentially predicting the three RGB channels of the image. Changing the number of output channels in a model from one to three is trivial to do. Since dense prediction vision transformers should have a good understanding of images, and are capable of creating predictions for every pixel, then I argue in theory, they should also be able to perform additional dense prediction tasks such as denoising and deblurring once finetuned to do so.

Related Works

Dense Prediction Vision Transformers

The original Vision Transformer (ViT) was created and applied on image classification, and the original independent patch structure of transformers wasn’t well-suited for dense prediction, so there have been many approaches to adapting the original transformer structure to something more suited for dense prediction. Pyramid Vision Trasformer (PVT) [9], Swin Transformer [7], and SegFormer[13] all take different approaches to creating features from different scales in the image, as being able to create and aggregate features from very small patches in the image is important for accurate segmentation along difficult edges. Vision Transformer Adapter (ViT-Adapter [3]) on the other hand, takes the original ViT and adds on some additional adapter modules so that after pretraining, the adapter modules inject additional information into it so that it is capable of performing well on dense prediction tasks. Mask2Former[4] makes use of masked attention, constraining cross-attention to the predicted mask regions. Pretraining the backbone has also taken a large focus in recent years, with Contrastive Learning[12] and Masked Image Modeling[5] both becoming competitive pretraining methods, and BEiT-3[10] achieving state of the art results on ADE20k and COCO as of November 2022.

Image Restoration

Despite how the base of all diffusion models are denoisers, the field of image denoising is still focused on transformers and convolutional networks. Previous research that focuses on image restoration often test their models on both denoising and deblurring tasks. Uformer[11] uses a hierarchical encoder-decoder network with transformer blocks, with nonoverlapping window-based self-attention and a multi-scale restoration module. Restormer[14] makes use of Multi-DConv Head ‘Transposed’ Attention (MDTA) blocks and gating mechanisms on the linear layers. Simple Baselines for

Image Restoration[2] suggests a Nonlinear Activation Free Network (NAFNet) and achieves state of the art results on both image deblurring and denoising.

Project Overview

Datasets

For image denoising, I'll be using SIDD, the Smartphone Image Denoising Dataset[1]. I will specifically be using the SIDD-Medium Dataset with sRGB images, not the full SIDD dataset, due to size constraints on my computer. For image deblurring, I'll be using GoPro[8], a dataset consisting of ground truth sharp images and realistically blurry images. I will be measuring the model performance using PSNR and SSIM for both tasks and datasets.

Models

The models I'll be exploring are ViT-Adapter, SwinTransformer, and Mask2Former, three of the best performing models on image segmentation currently. SwinTransformer in particular has already been explored on image restoration tasks[6], but I'll be additionally testing it on image deblurring here. To adapt these segmentation models to the denoising and deblurring task, I will simply be changing the output channels from one to three, and trying a variety of finetuning strategies to see what is most effective, such as:

1. Freeze the backbone encoder.
2. Mimicking the original learning rate schedule used for the model, with a learning rate 10 times smaller.

These models also all tend to overlap and borrow parts of each other. For Swin, I will not be using SwinIR as a basis, since I would like to specifically explore finetuning segmentation models for image restoration. The original Swin architecture uses Swin as its backbone, and UperNet as the decoder. For ViT-Adapter, I will be using its ViT-Adapter as the backbone pretrained using the BEiTv2 method, with Mask2Former as the decoder. For Mask2Former, I will be using Swin as a backbone and Mask2Former as the decoder. These three model choices are based on the current state of the art versions of these models on segmentation, and the trained versions of these models have been made public on their GitHub repositories.

Milestones

1. Nov 16: Project Proposal Due
2. Nov 23: Dataset and Dataloader code complete. Model code complete. Search for good hyperparameters for training. Begin training models on image restoration tasks.
3. Nov 30: Double check training progress. Run existing benchmark models to get comparison numbers.
4. Dec 7: Finish final report and poster.
5. Dec 8: Poster Presentations

References

1. Abdelrahman Abdelhamed, Lin S., Brown M. S. "A High-Quality Denoising Dataset for Smartphone Cameras", IEEE Computer Vision and Pattern Recognition (CVPR), June 2018.
2. Chen, Liangyu, et al. *Simple Baselines for Image Restoration*. Apr. 2022. *arxiv.org*, https://doi.org/10.48550/arXiv.2204.04676.
3. Chen, Zhe, et al. *Vision Transformer Adapter for Dense Predictions*. May 2022. *arxiv.org*, https://doi.org/10.48550/arXiv.2205.08534.
4. Cheng, Bowen, et al. *Masked-Attention Mask Transformer for Universal Image Segmentation*. Dec. 2021. *arxiv.org*, https://doi.org/10.48550/arXiv.2112.01527.
5. He, Kaiming, et al. *Masked Autoencoders Are Scalable Vision Learners*. Nov. 2021. *arxiv.org*, https://doi.org/10.48550/arXiv.2111.06377.
6. Liang, Jingyun, et al. *SwinIR: Image Restoration Using Swin Transformer*. Aug. 2021. *arxiv.org*, https://doi.org/10.48550/arXiv.2108.10257.
7. Liu, Ze, et al. *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*. Mar. 2021. *arxiv.org*, https://doi.org/10.48550/arXiv.2103.14030.
8. Nah, Seungjun, et al. *Deep Multi-Scale Convolutional Neural Network for Dynamic Scene Deblurring*. Dec. 2016. *arxiv.org*, https://doi.org/10.48550/arXiv.1612.02177.
9. Wang, Wenhai, et al. *Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions*. Feb. 2021. *arxiv.org*, https://doi.org/10.48550/arXiv.2102.12122.
10. Wang, Wenhui, et al. *Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks*. Aug. 2022. *arxiv.org*, https://doi.org/10.48550/arXiv.2208.10442.
11. Wang, Zhendong, et al. *Uformer: A General U-Shaped Transformer for Image Restoration*. June 2021. *arxiv.org*, https://doi.org/10.48550/arXiv.2106.03106.
12. Wei, Yixuan, et al. *Contrastive Learning Rivals Masked Image Modeling in Fine-Tuning via Feature Distillation*. May 2022. *arxiv.org*, https://doi.org/10.48550/arXiv.2205.14141.
13. Xie, Enze, et al. *SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers*. May 2021. *arxiv.org*, https://doi.org/10.48550/arXiv.2105.15203.
14. Zamir, Syed Waqas, et al. *Restormer: Efficient Transformer for High-Resolution Image Restoration*. Nov. 2021. *arxiv.org*, https://doi.org/10.48550/arXiv.2111.09881.