Can Diffusion Model Generalize Well in Image Super Resolution Task with Limited Fine-Tuning?

Project Proposal

Kejia Yin

1 Motivation

Image Super Resolution(SR) is a classic low-level computer vision task, which aims to generate High Resolution image from Low Resolution image[1]. There are numerous existing works on this task, and many of them are Generative Adversarial Networks(GAN) based methods[2]. Very recently, researchers start to use Denoising Diffusion Probabilistic Model to deal with generation tasks including SR[3]. Even though these deep learning models achieved very good SR performance on various datasets, there is still a question: whether these models can generalize well beyond the training and testing dataset? Unfortunately, related research is almost absent and only one recent research tried to set up a benchmark for this issue[4].

Generalization ability is a very important part of machine learning methods and there are studies points out existing SR method may not be able to generalize well to new data[5]. Thus, for this project, I want to evaluate how well can the diffusion model generalize to new domains.

2 Project Overview

For this project, I mainly want to answer the following questions:

First, can the existing diffusion based SR model generalize to new domain? To be specific, I will use SR3[3] model as the representative as it's the first diffusion model based SR method and also achieved very good performance compared to SOTA methods. As for the datasets, the SR3 is pretrained on FFHQ[6] dataset, and I will evaluate its performance on FFHQ, CelebA-HQ[7] and an animation faces dataset released by Prof. Huang-yi Lee in his Machine Learning courses at National Taiwan University. All these datasets are faces data, where FFHQ are real human faces dataset which covers a wider variation than CelebA-HQ in terms of age, ethnicity, image background, and accessories such as eyeglasses, sunglasses, hats, etc. The animation faces dataset consists of animation faces from the internet. My first thing to do is to directly apply the pretrained model to these datasets and see the result. The metrics will be both qualitive and quantitative, as for quantitative evaluation I will use PSNR, SSIM and LPIPS[8]. LPIPS is a recently proposed full-reference image quality metric, which correlates much better with human perception than PSNR and SSIM.

Secondly, if it cannot directly generalize to the new domain, can we fine tune the pretrained model with few data or few updates? I have already done some evaluation and found that directly apply the pretrained model to the animation faces dataset will not have satisfying results , because there exists a large domain gap between real human faces and animation faces. As for this project, I may not be able to have enough computational resources and time to train the model from scratch, so I'm wondering if the most naïve fine-tuning could help the pretrained model perform better. I'll try to limit the total amount of new domain data and the total steps of gradient updates, and also try to find the minimum amount of data and steps needed.

Thirdly, do we need to fine-tune all the time steps in diffusion model? Diffusion model is based on Markov Chain diffusion and reverse process, which will gradually transform image to pure Gaussian noise and vice versa. Thus, it's reasonable to think that the domain gap between images will gradually diminish along the diffusion process. According to this assumption, we may achieve good performance only fine-tuning the former part of the Markov Chain.

Lastly, will the fine-tuning impair the model's performance on original training dataset? Even though we may find better performance through fine-tuning with the new domain data, could its performance on original dataset be impaired? Ideally, we want our model could deal with all the tasks well at the same time, and it may not be good if fine-tuning will sacrifice its original performance.

Note that SR3 is a diffusion model based method and it takes quite long time to inference one HR image. Thus, I will not evaluate on the entire datasets, but use some samples as the representative.

3 Millstones and Timeline

Week1-first half: Set up environment, Download pretrained model and datasets, Evaluate on all the datasets with pretrained model.

Week1-second half ~ Week2: Fine-tune the pretrained model and find the minimum amount of data and steps, also the proper range of time steps, needed to perform well on the new domain, Evaluate on all datasets with the model fine-tuned with the found setting.

Week3: Write the report, Make the poster, Prepare presentation.

References

[1] Chen, H., He, X., Qing, L., Wu, Y., Ren, C., Sheriff, R. E., & Zhu, C. (2022). Real-world single image super-resolution: A brief review. Information Fusion, 79, 124-145.

[2] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... & Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European conference on computer vision (ECCV) workshops (pp. 0-0).

[3] Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., & Norouzi, M. (2022). Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[4] Liu, Y., Zhao, H., Gu, J., Qiao, Y., & Dong, C. (2022). Evaluating the Generalization Ability of Super-Resolution Networks. arXiv preprint arXiv:2205.07019.

[5] Menon, S., Damian, A., Hu, S., Ravi, N., & Rudin, C. (2020). Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 2437-2445).

[6] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

[7] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

[8] Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.