Image Super-Resolution Using Transformer

Abstract

Image super-resolution (SR) aims to reconstruct high-quality images. Image super-resolution is a long-standing low-level vision problem that aims to restore high-quality images from low-quality downscaled images. While state-of-the-art SR methods are based on convolutional neural networks, few attempts have been made with Transformer, which shows impressive performance on high-level vision tasks. In our project, we are going to use a fantastic model for image super-resolution tasks based on the Transformer. This model consists of two parts: shallow and deep feature extraction and high-quality image reconstruction. In particular, the deep feature extraction module is composed of several residual Swin Transformer blocks (RSTB), each of which has several Swin Transformer layers together with a residual connection. We use this model to conduct experiments on image super-resolution tasks (including classical and lightweight image super-resolution).

1. Introduction

Image SR aims to reconstruct the high-quality clean image from its low-quality degraded counterpart. Results from the image SR can largely influence the next high-level part to perform recognition and understanding of the image data. Recently, deep learning has been widely applied to solve low-level vision tasks. Although convolutional neural networks (CNNs) are primarily used in image SR tasks, these methods generally suffer from two basic problems that stem from the basic convolution layer. First, the interactions between images and convolution kernels are content-independent. Using the same convolution kernel to restore different image regions may not be the best choice. Second, convolution is ineffective for long-range dependency modeling under the principle of local processing. Thus we are going to use a transformer-based model using shift windows (Swin Transformer [2]) to improve the quality of SR tasks.

2. Related Work

CNN [1, 4] are mostly used in image super-resolution tasks. Most CNN-based methods focus on elaborate architecture designs such as residual learning [1] and dense connections [4]. Transformer and its variants have proven its success being powerful unsupervised or self-supervised pretraining frameworks in various natural language processing tasks. Due to its impressive performance, Transformer [3] has also been introduced for image super-resolution. A popular variation about Transformer is called Swin Transformer [2], whose representation is computed using shifted windows. It has shown great promise as it integrates the advantages of both CNN and Transformer. On the one hand, it can process image with large size due to the local attention mechanism. On the other hand, it can model long-range dependency with the shifted window scheme.

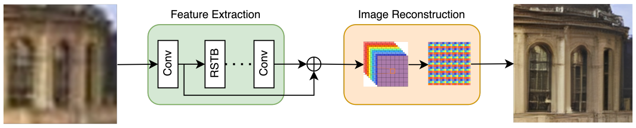

Figure 1. The architecture of our model for image super-resolution

3. Technical Overview

As shown in Figure 1, our model consists of two modules: feature extraction and image reconstruction modules.

3.1. Feature extraction

Given a low quality input(LQ), we use a 3×3 convolutional layer to extract shallow feature as:

$$F_0 = H_{SF}(I_{LQ}), \tag{1}$$

and K residual Swin Transformer blocks (RSTB) and a 3×3 convolutional layer to extract deep feature as:

$$F_{DF} = H_{DF}(F_0), \tag{2}$$

3.2. Image reconstruction

We use the sub-pixel convolution layer to upsample the features. Then the high-quality image I_RHQ is reconstructed by aggregating shallow and deep features as:

$$I_{RHQ} = H_{REC}(F_0 + F_{DF}), \tag{3}$$

where $H_REC(\cdot)$ is the reconstruction module function.

3.3. Loss function

For image SR, we optimize the parameters of our model by minimizing the L_1 pixel loss as:

$$L = \|I_{RHQ}\text{-}I_{HQ}\|_1, \tag{4}$$

4. Expected results

The proposed method will be trained on the classical and lightweight image datasets(DIV2K/Flickr2K). Then it will be evaluated on five benchmark datasets (Set5/Set14/BSD100/Urban100/Manga109) by quantitative comparison (average PSNR/SSIM/PSNR-B). The expectation is to get results similar to the state-of-the-art baselines.

References

[1] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[2] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[4] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1