Efficient Transformer SRGAN

AjaySriram Muthuraman

&

Sasank Potluri

09/22/2023

___

SPTP: ADV PERCEPETION CS7180

___

Image Enhancement

Abstract:

In this project, we implemented an image enhancement neural network inspired by "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network." This paper's two-step training process, involving a ResNet-based architecture and an additional discriminator, led to a remarkable 2dB increase in PSNR. We aimed to explore if a different neural network could achieve similar enhancements.

Our novel neural network design closely resembled the "Transformer for Single Image Super-Resolution" architecture. By subjecting it to SRGAN's training techniques, we sought to evaluate its performance gains. We utilized the SRGAN repository for training and testing pipelines and integrated ESRT's modules to construct our network.

The core of our approach lies in a generator module that combines convolution and transformer paths, optimizing complex feature reconstruction and image detail enhancement. High-frequency pattern blending (HPB) modules, including high-frequency feature mapping (HFM) and adaptive residual feature blending (ARFB), were employed to separate image components and enhance synthesis.

Our training method mirrored SRGAN, initially training with an L2 loss function and then adversarial training with an unaltered discriminator. This project underscores the potential of diverse neural network combinations to advance image enhancement, aiming to achieve a similar leap in PSNR as SRGAN while exploring alternative architectures.

Introduction & Prior Work:

We have implemented a neural network for image enhancement. Our primary inspiration for this project came from a research paper titled "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network." In the aforementioned paper, the authors employed a ResNet-based architecture to generate enhanced images. During the training process, they initially trained the model conventionally, similar to other models. However, after training it with an L2 Loss, they further refined the model by incorporating another discriminator. This additional step resulted in a performance improvement of 2dB in PSNR (Peak Signal-to-Noise Ratio). This motivated us to explore whether training a different neural network using a similar approach would yield similar performance gains.

Our second neural network closely resembles the one discussed in the paper titled "Transformer for Single Image Super-Resolution." We intentionally chose this similarity to assess whether our new neural network, when trained using the SRGAN's training technique, would also achieve comparable performance advantages. To conduct this project, we utilized the training and testing pipelines from the SRGAN repository, and we leveraged ESRT's repository to incorporate individual modules into building our neural network.

Methods:

Our generator module draws significant inspiration from the ESRT architecture. It primarily comprises two pathways following a convolutional layer. Path1 consists of convolutional blocks followed by a transformer block, while Path2 exclusively contains convolutional blocks. Our primary rationale behind employing this structure is that the transformer pathway can effectively reconstruct complex features within the image, while the convolutional pathway enhances the detail of each feature during image construction. Both pathways incorporate HPB (High-Pass Block) modules. However, in Path1, the HPB modules are followed by transformers, whereas in Path2, they are succeeded by upsampling convolutions.

The HPB module consists of two sub-modules, namely the HFM (High-Frequency Module) and ARFB (Adaptive Residual Feature Block) modules. The HFM module primarily serves to segregate high-frequency image features from low-frequency ones. After passing through the HFM module, the high-frequency data is separated and passed through a distinct set of ARFB modules. This high-frequency information is subsequently added back to the low-frequency data. The low-frequency data also goes through a separate set of ARFB modules. The architecture of the ARFB modules closely resembles cascading blocks, where blocks from previous layers are concatenated, and the number of channels is adjusted in subsequent layers.

The Efficient Transformer block we employed closely resembles the Swin Transformer, focusing on attending to the surrounding pixels. When combining the outputs of both Path1 and Path2, we introduce a scaling factor that serves as a learnable parameter. This scaling factor assigns weightage to each pathway.

Our training methodology is directly adopted from the SRGAN paper, and it remains unchanged. We have not altered the discriminator architecture either. As discussed in the SRGAN paper, we initially train the generator model using the L2 loss function. Subsequently, we retrain the model using the discriminator to evaluate if our network exhibits a similar increase in PSNR (Peak Signal-to-Noise Ratio).

Results:

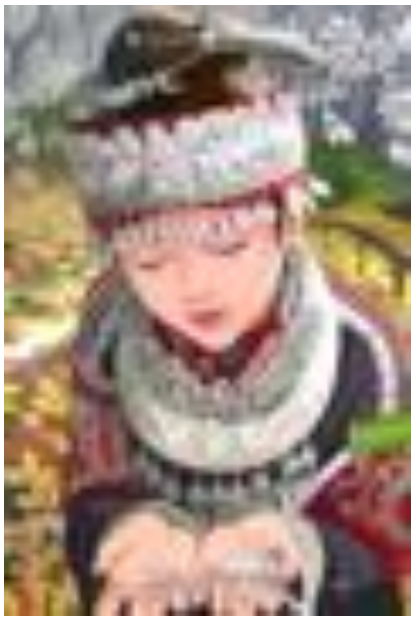

Figure 1: *Input Image*

Figure 2: *Output Image*

The result image appears to be a bit more vivid with the minute features being sharpened and displayed properly.

Figure 3: *Input Image*

Figure 4: *Output Image*

The output image is slightly sharper than the input low resolution image, it is evident that the model did not perform well with this specific image.

Figure 5: *Input Image*

Figure 6: *Output Image*

The model performs relatively better with this bird image, with the edges and the textures being highlighted in a better manner.

Set5	Scale	EffTeNet	ETGAN
PSNR	4	8.940(9.340)	8.39(8.62)
SSIM	4	0.0182(0.02120)	0.0138(0.0152)

Reflection:

During this project, we encountered certain challenges that influenced the performance of our model. We opted to work with the Set5 dataset for training, primarily due to computational constraints that necessitated a reduction in both the dataset size and the number of training epochs. Consequently, the model's training process faced limitations, and we observed a relatively high training loss, starting at approximately 0.03 for the SSRNET and increasing to 0.45 during SRGAN training. These limitations stemmed from a combination of inadequate dataset size and the constraints of our CPU resources.

Despite these challenges, our model demonstrated commendable performance on images characterized by high contrast between foreground and background elements. This success can be attributed to the efficient transformer integration we introduced, which enabled the model

to excel in such scenarios. However, it is essential to acknowledge that there remains substantial room for improvement, particularly with access to more extensive datasets and more powerful computing resources.

Acknowledgments:

We would like to extend our sincere gratitude to our professor and the dedicated teaching assistants for affording us the opportunity to embark on this innovative journey and explore new horizons within the field. Their guidance, support, and encouragement were invaluable throughout the project, enabling us to push our boundaries and delve into uncharted territories.

Furthermore, we extend our appreciation to the authors of the two influential papers that served as the foundation for our work. Their pioneering research provided the necessary inspiration and insights that guided our project, reinforcing the collaborative and evolving nature of scientific exploration.