Image Inpainting using a U-Net model with a fused ConvMixer Encoder

Umar Masud

Abstract

Encoder-decoder based architectures such as the U-Net model have been quite useful for various image related tasks such as segmentation, reconstruction, etc. These architectures work on the capability to condense the image into latent features and then up-sample back to reconstruct the original image from the latent space. In this work we have tried to improve the condense features using an additional fused-encoder. We used a ConvMixer model to generate additional features and then concatenated it into the U-Net model in the latent space. The model was tested on CIFAR-10 with random sized square masks and showed potential results with minimal training. Code: https://github.com/umar07/image-inpainting

1 Introduction

Image Inpainting is a task of reconstructing missing regions in an image. It is an important problem in computer vision and an essential functionality in many imaging and graphics applications, e.g. object removal, image restoration, manipulation, re-targeting, compositing, and image-based rendering.

A lot of research has been done in the field of image inpainting over the years. Both statistical methods as well as deep-learning based methods have proven quite useful for the task. Deep learning models are now used to solve for inpainting task in an end-to-end fashion. The most basic and commonly used model for image reconstruction is the U-Net [1] model. We have also made use of this model however with additional encoder representations concatenated in the latent space(see figure 1). This type of architecture is inspired by the work of Ini Oguntola's project [2] where they used this architecture with ResNet fusion for image colourisation task. We replaced the ResNet model with a recently proposed ConvMixer [3] model that has been proven to be more efficient. The results show that this type of architecure performs fairly even for image inpainting.

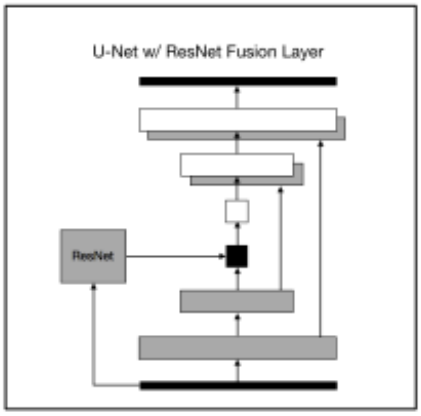

Figure 1: Model architecture (taken from [2])

2 Related Work

Various methods have been proposed over the years starting from statistical approaches such [4], [5], [6], etc. These algorithms work well on stationary textural regions but often fail on non-stationary images. More comprehensive techniques using deep learning are now used such as [7] which used a trained generative model. [8] proposed the use of partial convolutions. [9] used a contextual attention module for capturing long-range spatial dependencies. Gated convolutions were used by [10] for free-form image inpainting.

3 Approach

3.1 Proposed Model

We have used a similar architecture as used by [2]. However, we have replaced the ResNet model with the new ConvMixer [3] model. The feature embeddings from the ConvMixer [3] is concatenated in the latent space of the U-Net model. Our intuition was that the fused encoder module will increase the richness of the feature space and help in better reconstruction. The additional encoder allows us to make the U-Net module shallow and reduces the number of parameters with fair results for image inpainting.

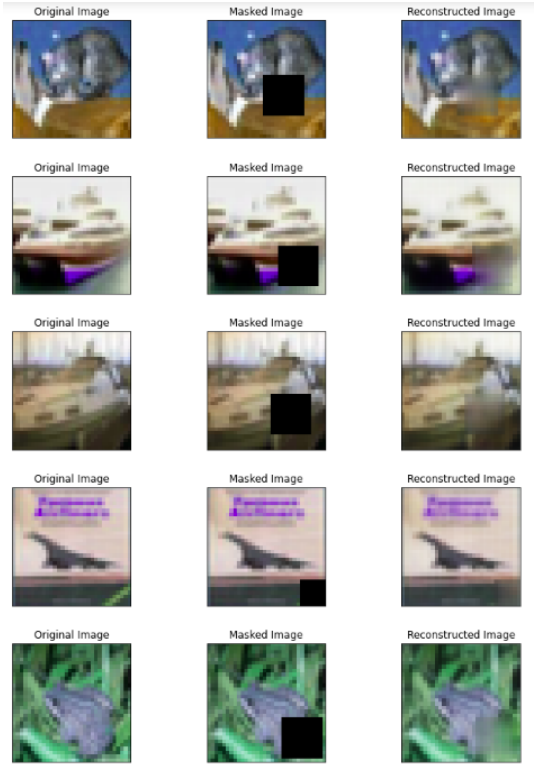

Figure 2: Results on CIFAR-10 dataset with ConvMixer fused U-Net model

3.2 Implementation Details

Due to computational constraints, we built the simplest and minimal model possible with ConvMixer of depth = 2 and twice U-Net downsampling with the overall size of image going down to just $1/4^{th}$ of the original image size. The dataset used was CIFAR-10 with 45000 training samples, 5000 validation samples and 10000 test samples. A random sized square mask used to prepare the dataset.

The loss function was Mean Sqaured Error and the evaluation metric used was Dice Coefficient. The model was trained for just 5 epochs and took around 25 minute per epoch on CPU with batch size = 2.

3.3 Results

The dice coefficient with vanilla U-Net was **0.5837** while with the fused-encoder U-Net we got **0.5915**. This proves our intuition that additional fusion layer for encoding features can improve the results. The inpainted results can be seen in figure 2.

Conclusion

In this work we tried a fused-encoder module for image inpainting task using the U-Net model. The model gave fair results with the minimal training and shallow architecture which shows potential for performing well for image inpainting task. The model can be made better by using pre-trained weights for the fusion-encoder, increasing the depth, using better loss functions meant specifically for image inpainting tasks, training on a bigger and more complex datasets.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[2] Ini Oguntola. Image colorization with conditional adversarial networks.

[3] Anonymous. Patches are all you need? In *Submitted to The Tenth International Conference on Learning Representations*, 2022. under review.

[4] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on Image Processing*, 10(8):1200–1211, 2001.

[5] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, page 417–424, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[6] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), August 2009.

[7] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with deep generative models, 2017.

[8] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions, 2018.

[9] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention, 2018.

[10] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-form image inpainting with gated convolution, 2019.