Electrical Engineering, Mathematics and Computer Science, TU Delft

CS4240 Deep Learning

Reproducibility project:

Rethinking Image Inpainting via a Mutual Encoder-Decoder with Feature Equalization.

by Liu H, Jiang1 B., Song Y., Huang W. and Yan C.

April 11, 2021

Group Number:	121	
Group Members:	Marie Humblet Vertongen	4580729
	Aymar De Brouchoven de Bergeyck	5391369

1 Introduction

Image inpainting has been a very popular subject for the last few years. There are several techniques to do inpainting. One of them is by trying uncorrupted images over hole regions via patch-based image matching. The other one is by using CNN in different layers and layers sizes, which produces more global and consistent results, the Encoder-Decoder.

There are different ways to use an Encoder-Decoder. First, there is the Simplies' way, by just using the simple Encoder-Decoder. The problem with this one is that the results are limited. Another way is by using two sequential Encoder-Decoder, one that focuses on the structure and the second one that focuses on the texture. The problem with this is that it may produce inconsistent structures and textures in the hidden part of the image. This paper tries another technique, that globally works better than the two other mentioned methods.

Link to paper: https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123470715.pdf
Link to Code: https://github.com/KumapowerLIU/Rethinking-Inpainting-MEDFE

2 Architecture

The reproduced paper proposes a mutual encoder-decoder network for image inpainting. The shallow layers of the CNN represent the textures and the deep layers represent the structure of the image. These two features pass through two separate branches, with both a multi-scale filling block for hole filling. Then these two features pass through a feature equaliser to ensure that the structure and the texture are consistent. The feature equaliser consists of channel reweighing and bilateral propagation. The equalized features are added to the decoder features in all the feature levels via the encoder-decoder skip connections. This method has shown to be efficient to remove blur and artefacts around the hole regions. It has been shown that this method performs better than the usual methods.

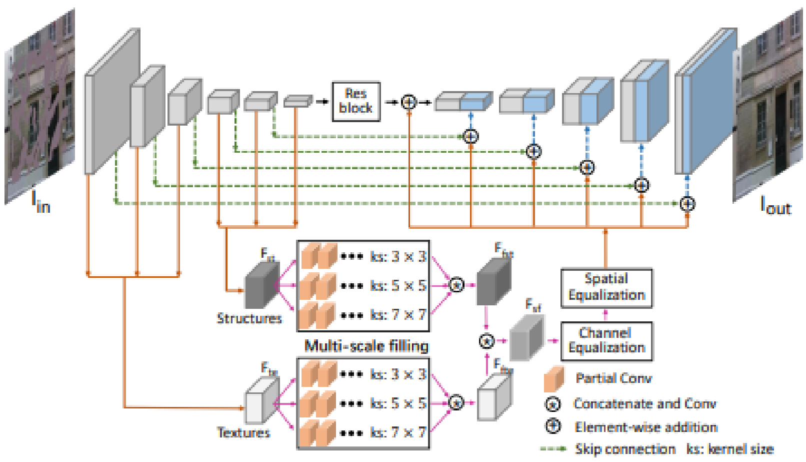

Figure 1: Structure of the Mutual Encoder-Decoder with Feature Equalization

3 Experiments

In this reproducibility project, several criteria have been reproduced and/or analysed. The first criteria were to try to get the same results as in the paper itself. Another aspect was to use another data set to train the model. Finally, the importance of the number of Res-Blocks is analysed and discussed.

3.1 Equivalent results

The results that are being replicated are the one of Fig. 5 in the original paper. They represent a face, that has been covered with a square mask, the goal of the model is to recreate the covered face.
The author used the CelebA data-set to train and test his model. In the paper, the author used the entire CelebA data-set, or 202,599 images, and did 6 epochs. For reproducing it, we used a data-set of 10 000 images of the CelebA data-set, 14 epochs and the same square mask they used. The size of the data-set was limited because of the computation time to transform the images into smooth images. The number of epochs was limited by the training time, on "Google Colab Pro" an epoch of 10000 images took around 1 hour to process, the total training time of the model was around 14 hours.

Figure 2: Good output image

Figure 3: Bad output image

The results were sometimes satisfying, sometimes not. By looking at the total loss of the training model, the results look satisfying because for the lasts epochs the error flattens a bit. But the model performs less good during a subjective and visual analysis. Some output images look very good, but others look pretty bad. By comparing the subjective output quality of the last training epoch and the test set it is concluded that the model didn't overfit.

3.2 New data-set

The new data sets were different types of masks. The two different masks types are, one that looks like a chess-board, the other covers places in the form of "COPYRIGHT".

For these new data sets, the model is trained twice for both masks, on 10000 different images, from the CelebA data-set, for 8 epochs. The training took around 8 hours each and the results are very good. Firstly, because the total error for the lasts epochs was becoming more and more constant, but still with a lot of noise. The visual and subjective results were very good, all the output images looked in both cases very convincing.
These good results are probably due to the easy task the model has to do. The program had to reinvent much fewer structures and textures than when the square mask was used. The surface of the chessboard mask covers two times more surface than the middle square mask, but with less training, it obtains better results. This is because, with the chessboard mask, the model doesn't need to "reinvent" an entirely new face, it just needs to

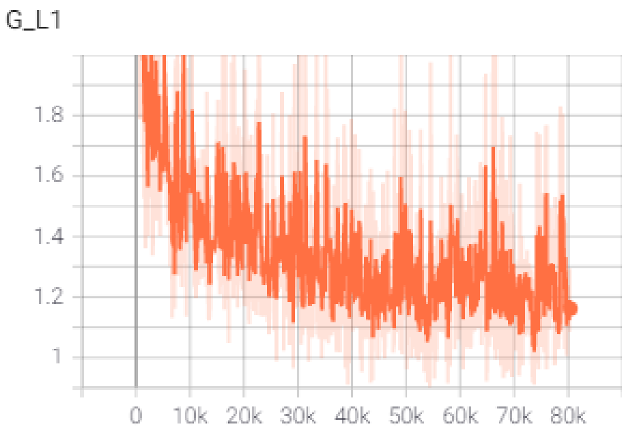

Figure 4: Total loss: "chess board" mask

reinvent much more but much smaller parcels of the image. For each parcel there are it's surrounding that helps it, recreating this covered part. The same explanations go for the "COPYRIGHT" mask.

Figure 5: "Chess board" mask, ouput

Figure 6: "COPYRIGHT" mask, ouput

3.3 Ablation Study: effect of the Residual Blocks on the PSNR value

In this paper, the effect of the number of Residual blocks, ResBlocks, on the visual quality of the resulting image after the impaintment by the encoder-decoder method has been analysed. Since judging the outcome purely on the digital image can be subjective, the Peak signal-to-noise ratio, the PSNR value, has been used as the metric of comparison.
The PSNR value is calculated from the Mean Square Error, MSE. The MSE allows us to compare the 'true' pixel value of the original image to the resulting image of impaintment. The MSE value represents the average of errors between both images. The error is the amount by which the values of the original image differ from the degraded image. This means that the higher the PSNR, the better the resulting image.[1]
In this paper the PSNR value has been calculated as follows:

$$MSE = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - y_{ij})^2 \tag{1}$$

$$PSNR(x, y) = \frac{10\log_{10}[\max(\max(x), \max(y))]^2}{x - y^2} \tag{2}$$

m	number of rows in cover image	n	number of columns in cover image
x_{ij}	pixel value from original image	y_{ij}	pixel value from resulting image

In this experiment, the effect of one, four, eight and sixteen res-blocks on an altered database has been analysed. The image illustrates the average PSNR value against the number of ResBlocks. This resulted in very minimal differences in the PSNR values between the amount of ResBlocks. The resulting PSNR values are quite normal for this kind of experiment.
The results in Figure 7 show that the average PSNR value over the test data sets increase until a maximum at 4 ResBlocks and at 8 ResBlocks the PSNR value stays constant. Considering another paper that also looked into the influence of ResBlocks on the PSNR values, Super-resolution using lightweight detailnet network by Barzegar, S., Sharifi, A. Manthouri, M., concluded that the addition of ResBlocks improved the PSNR value.[2] However, in their research they only considered 1 to 4 ResBlocks on the contrary of our research where we went to 16 ResBlocks. So considering the range from one to four ResBlocks, our research would approve this statement considering the range from one to four ResBlocks.

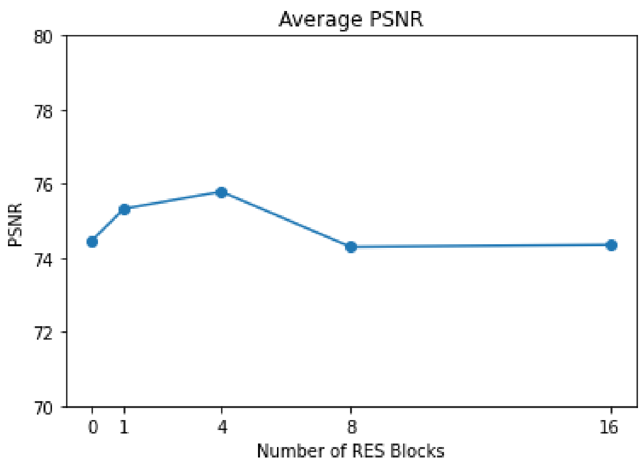

Figure 7: Graph with results average PSNR

Looking into more specific images, a darker image and a lighter image, the results of this paper concluded that the average tendency also applied for both types of images. However, a higher PSNR value for a darker image occurred compared to the PSNR value of a lighter[3]. These values can be found in Figure 8 for the darker image and in Figure 9 for the lighter image. This could be the cause of a variety of reasons such as the database or the colours represented in the image since the PSNR value is also more variable in grey on the contrary of coloured images since they are more susceptible to luminosity. Confirming this assumption is out of the scope of this research.
The high PSNR value can be due to the limited area of the image that has been reconstructed. Looking into this is also out of the scope of this paper.

[1] https://www.ni.com/fr-be/innovations/white-papers/11/peak-signal-to-noise-ratio-as-an-image-quality-metric.html
[2] https://link.springer.com/article/10.1007%2Fs11042-019-08218-4
[3] The images with the occurring number of ResBlocks can be found **??**

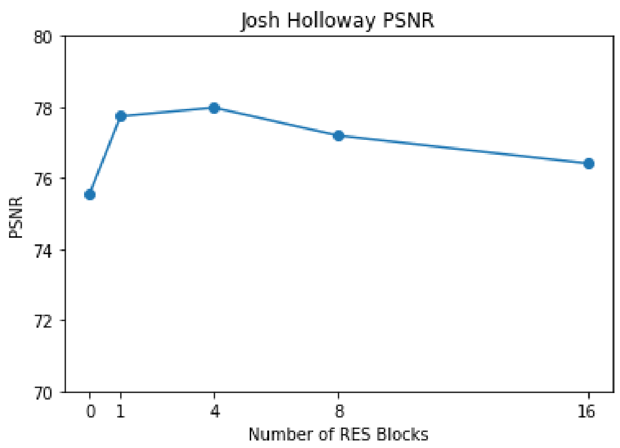

Figure 8: PSNR values Josh Holloway

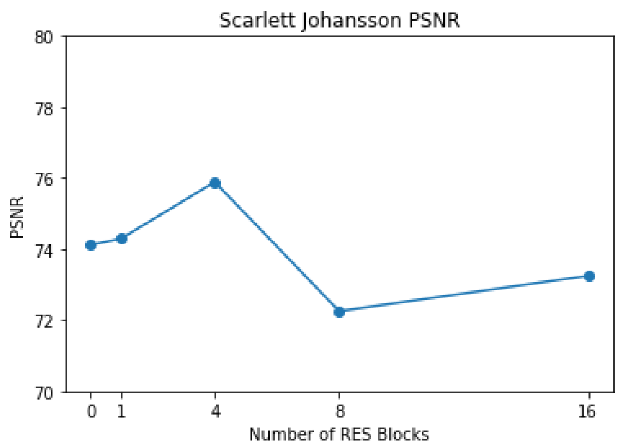

Figure 9: PSNR values Scarlett Johansson

4 Work distribution

This reproducibility project has been realised by two people: Aymar De Brouchoven de Bergeyck and Marie Humblet Vertongen.

Aymar mostly worked on debugging the code, making the test and train code workable. He also experimented with the code and the several datasets and mask.

Marie created the different masks and worked on the ablation study, the effect of the number of ResBlocks on the quality of the results.

5 Conclusion

This project looked into the reproducibility of the paper; Rethinking Image Inpainting via a Mutual Encoder-Decoder with Feature Equalization by Liu H, Jiang1 B., Song Y., Huang W., and Yan C., the stability of the inpaintment when provided to new masks on the data and performed an ablation study concerning the Residual Blocks.

Concerning the reproducibility, the paper was relatively easy to understand and well documented, some code was provided. However, the provided code was not well explained; it was not commented and the classes were not well structured and while running the code, several codes popped up. Considering these aspects, we would give this paper a 6/10.

The code reacted well when showed new data. The quality of the inpaintment of the result is quite high. The algorithm showed a similar result on the different datasets. Another observation was that the prediction of the dark cell region was not very accurate, but overall it manages to successfully remove cell nuclei.

The ablation study on the residual blocks suggested that the amount of resblocks does not have a major influence on the quality of the results. This study suggested that the fact of implementing a ResBlock will on average shift the quality of the PSNR value by a maximum of 2 points.

Figure 10: Image of Josh Holloway using 0 ResBlock

Figure 11: Image of Scarlett Johansson using 0 ResBlocks

Figure 12: Image of Josh Holloway using 1 ResBlock

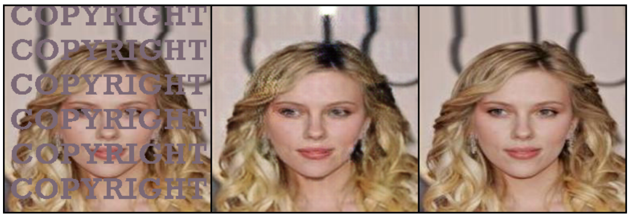

Figure 13: Image of Scarlett Johansson using 1 ResBlock

Figure 14: Image of Josh Holloway using 4 ResBlocks

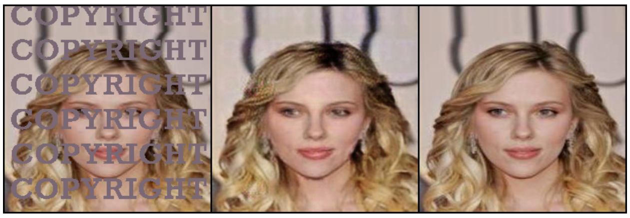

Figure 15: Image of Scarlett Johansson using 4 ResBlocks

Figure 16: Image of Josh Holloway using 8 ResBlocks

Figure 17: Image of Scarlett Johansson using 8 ResBlocks

Figure 18: Image of Josh Holloway using 16 ResBlocks

Figure 19: Image of Scarlett Johansson using 16 ResBlocks

Figure 20