TDT 4195 - Lab Assignment 5

Kari Lovise Ness (MTING)
Håvard Kjellmo Arnestad (MTFYMA)

October 2019

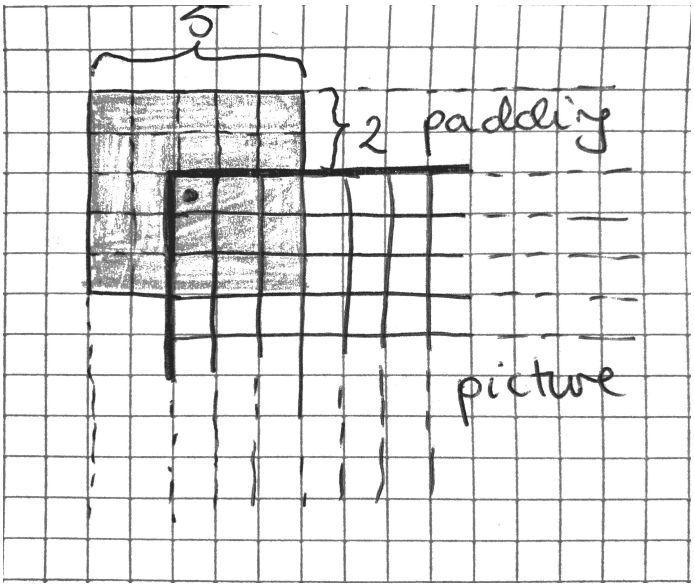
Task 1: Theory

a) Given a single convolutional layer with a stride of 1, kernel size of 5 x 5, and 6 filters. If I want the output shape (Height x Width) of the convolutional layer to be equal to the input image, how much padding should I use on each side?

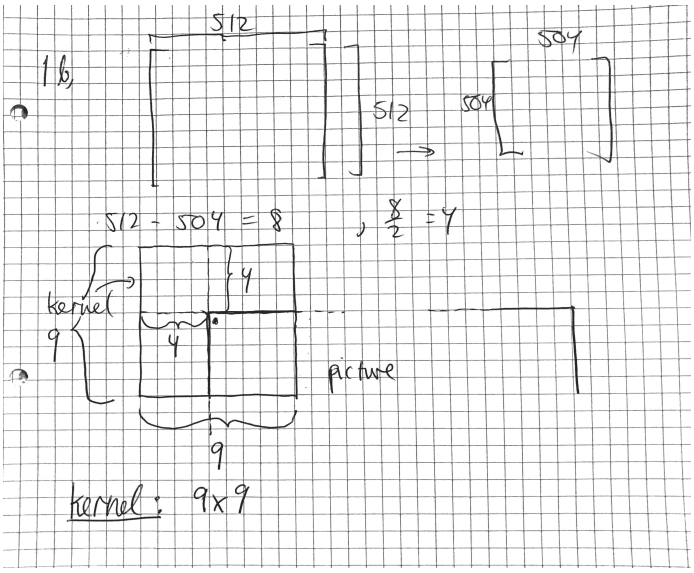
When using a 5x5 kernel, one needs a padding of thickness 2 in order to have an output shape equal to the input shape. As seen in figure 1, the padding makes it possible to make the convolution all the way to the borders (in the figure, the corner pixel marked with a black dot).

b) You are told that the spatial dimensions of the feature maps in the first layer are 504 x 504, and that there are 12 feature maps in the first layer. Assuming that no padding is used, the stride is 1, and the kernel used are square, and of an odd size, what are the spatial dimensions of these kernels? Give the answer as (Height) x (Width).

A feature map is an output channel of the convolutions. As the feature maps are 8 elements smaller than the images, and the kernel is quadratic, the kernel will "cut away"4 elements in both x- and y-direction from each border of the original image. This can be seen in figure 2, where the pixel marked with a black dot will be the outermost pixel of the output feature map. 4+1+4 = 9, so a 9 x 9 kernel.



Figur 1: 5x5 kernel over corner pixel.



Figur 2: 5x5 kernel over corner pixel.

c) If subsampling is done using neighborhoods of size 2 x 2, with a stride of 2, what are the spatial dimensions of the pooled feature maps in the first layer? Give the answer as (Height) x (Width).

With feature maps of size 504x504, neighborhoods of size 2x2 and a stride of 2, the spatial dimensions of the pooled feature maps in the first layer will be

252x252.

d)What is the depth (number) of the pooled feature maps in the first layer?

The depth of the pooled feature maps in the first layer is 12 (the number of feature maps in the first layer).

e)The spatial dimensions of the convolution kernels in the second layer are 3 x 3. Assuming no padding and a stride of 1, what are the sizes of the feature maps in the second layer? Give the answer as (Height) x (Width).

The input image in layer 2 has the size 512x512. As there is no padding and a kernel with the size 3x3, the feature maps in the second layer will have the size 510x510.

f) Table 1 shows a simple CNN. How many parameters is in the network? In this network, the number of parameters is the number of weights + the number of biases. Assume the network takes in an RGB image as input and the image is square with a width of 32.

See figure 3, 4, 5 and 6 for calculations. The calculations have been carried out with the following equations:

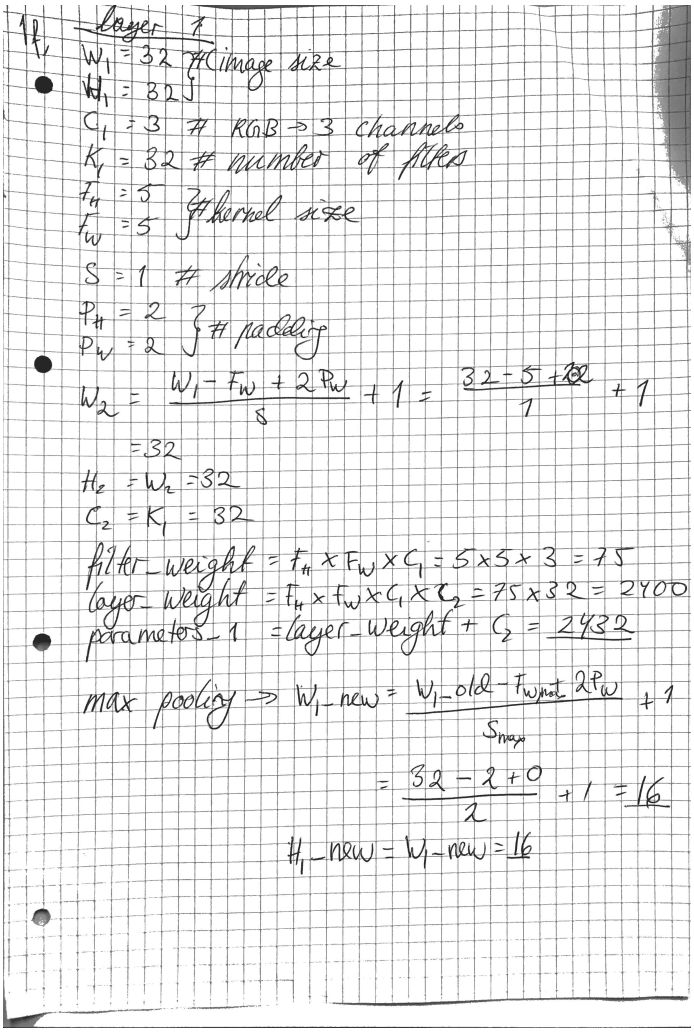
$$W_2 = (W_1 - F_W + 2P_W)/S_W + 1 \tag{1}$$

$$H_2 = (H_1 - H_W + 2P_H)/S_H + 1 \tag{2}$$

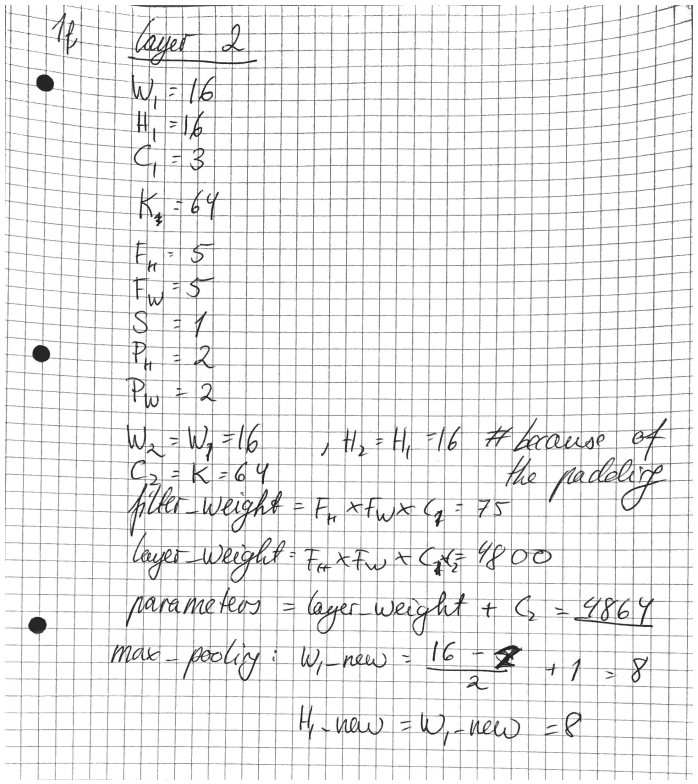
$$filter\_weight = F_W \times F_H \times C_1 \tag{3}$$

$$layer\_weight = filter\_weight \times C_2 \tag{4}$$

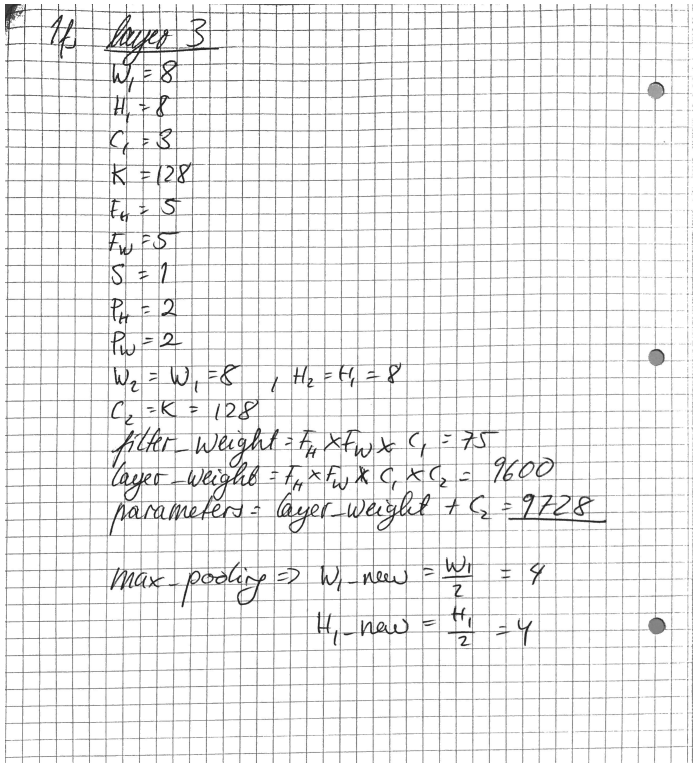
$$parameters = layer\_weight + C_2 \tag{5}$$



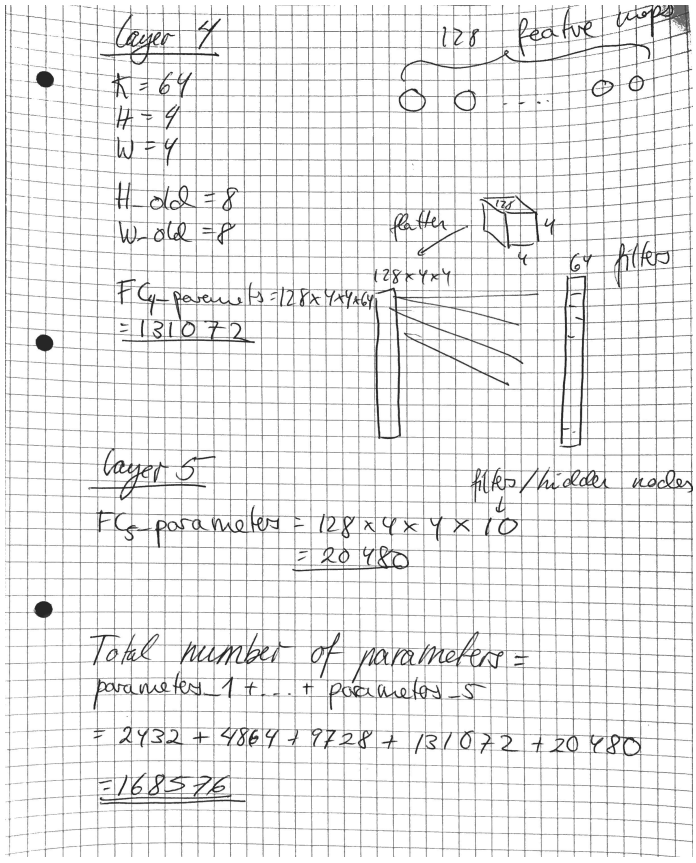
Figur 3: Task 1



Figur 4: Task 1f, layer 2



Figur 5: Task 1f, layer 3



Figur 6: Task 1f, layer 4 and 5

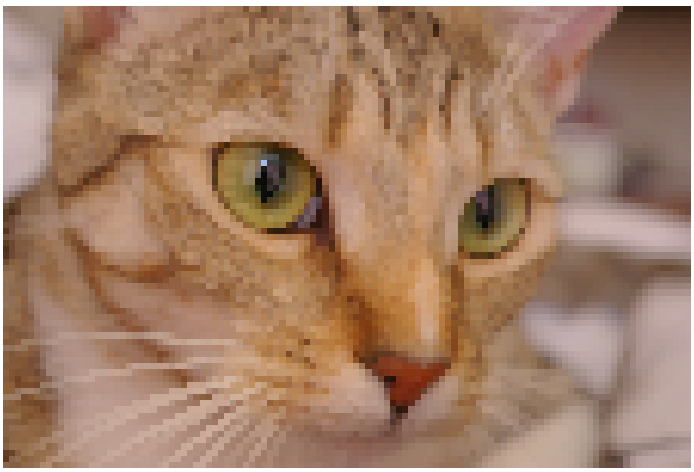
The final number of parameters is 168576.
Each of the layers has the following amount of parameters:

1. 2432
2. 4864
3. 9728
4. 131072
5. 168576

Task 2: Basic Image Processing

See figures further down.

a)

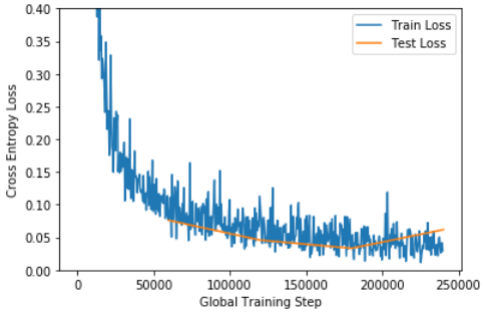


Figur 7: The result of applying the maxpool-function to the chelsea-image. It is slightly brighter than the original image, but the resolution is smaller.

b)

Final Validation loss: 0.06171464130403266. Final Validation accuracy: 0.9793. It appears that we see overfitting since the testloss begins to increase; the network starts to overspecialize on the training dataset and loses generality which is seen as worse performance on the unseen images.

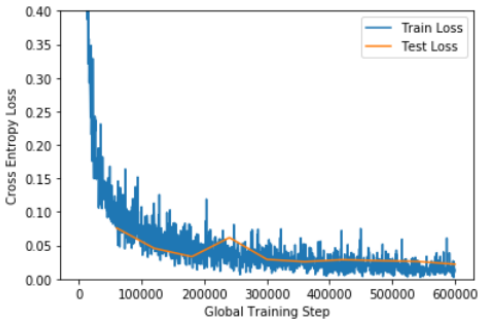
Figur 8: The result of applying the maxpool-function to the checkerboard-image. Sadly"(for vizualisation purposes) it is supposed to be white, since the maxpool takes the white pixel value (since it is maximum).



Figur 9: The result of training the network as specified. It appears that we see overfitting near the very end since the testloss begins to increase; the network starts to overspecialize on the training dataset and loses generality which is seen as worse performance on the unseen images.

c)

Final Validation loss: 0.022048618289413 84. Final Validation accuracy: 0.9926. It seems that the apparent overtraining was not really a problem since the test loss continued downwards shortly after. It is slightly suspicious that the test loss increases at the very same time (given that there is some randomization).



Figur 10: The result of training the network as specified. Overfitting is not seen anymore, since the peak is temporary.

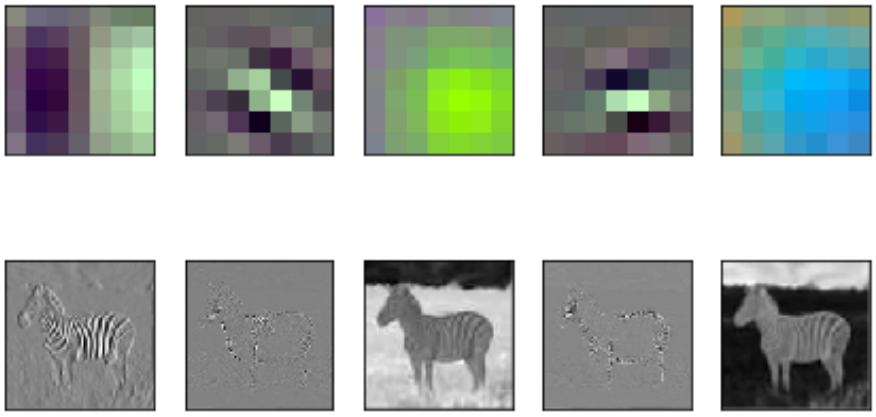
d)

e)

The visualization of filters and activations on figure 11 can be seen in figure 11. From left:

Figur 11: The original zebra image.

- A vertical sobel-filter which detects vertical edges in the image. One can see that it detects vertical edges as there is a vertical split in colours in the filter.
- The filter detects diagonal edges. This can be seen as the filter has a diagonal colour tendency, from the upper left corner towards the right lower corner.
- The filter provides a measurement of green components in the image. The more green in the pixel, the whiter it is in the activation.
- The filter highlights the outline of the zebra to a certain extent, and activates in areas which have a high density of stripes. The could be a second order edge detection. However, the activation is noisy, and the filter does not show any clear trends - mostly a diagonal edge filter, with a different direction from nr. 2.
- The filter provides a measurement of blue components in the image. The more blue in the pixel, the whiter it is in the activation.



Figur 12: Visualization of filters and activations in ResNet50 with indices [5,8,19,22,34]

Task 3: Theory (filtering)

a)

1) Pad the kernel with zeroes so that it matches the image dimensions. If you don't want a circular convolution, pad both with zeroes again to twice the dimensions. 2) Take the `FFT` of the image and the kernel. 3) Multiply both in the frequency domain. 4) Perform the inverse transform $\texttt{FFT}^{-1}$. The resulting image will be the original image filtered with the kernel.

b)

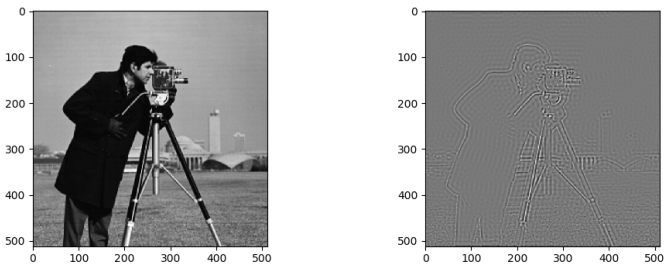
As with audio signals, images can be decomposed into sine-waves. These waves have high or low frequency, corresponding to their spatial variation (for images). A low pass filter conserves the slow variations in the image, but removes any fine details that are built up from higher frequencies. A high-pass filter is opposite, it removes any slow variations in the image.

c)

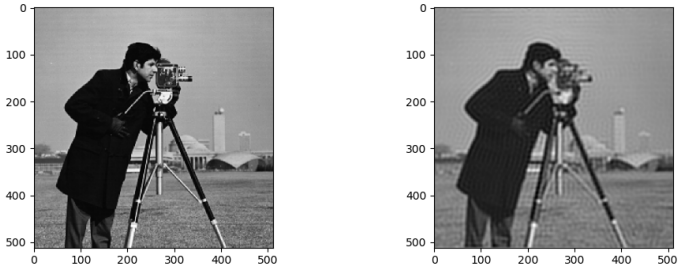
Typically zero is taken to be black, and 255 is taken to be white. In that case (a) is a high-pass filter in the y-direction, since it zeroes out frequencies that are low (near center) and in the x-direction. (b) Is a high-pass filter, since the low frequencies near the center is zeroed. (c) is a low-pass, since higher frequencies are removed.

Task 4: Programming (filtering)

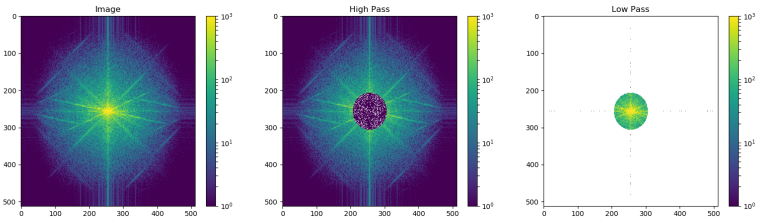
a)



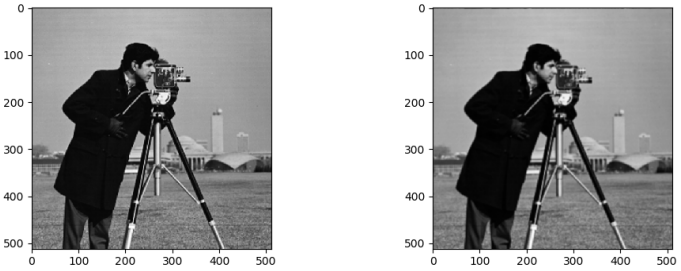
Figur 13: Original (left), and high-pass, 4a).



Figur 14: Low-pass, 4a). Original to the left.



Figur 15: Original and filtered images in 4a). White regions are below the logarithmic cutoff.

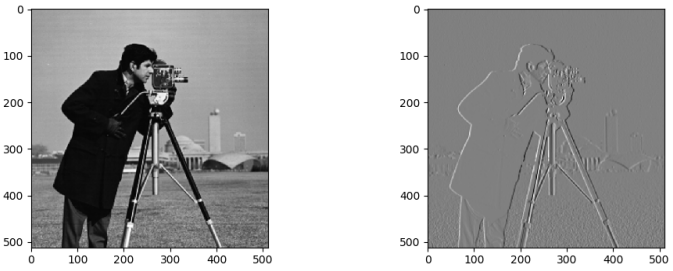


Figur 16: Gaussian, 4b). Original to the left.

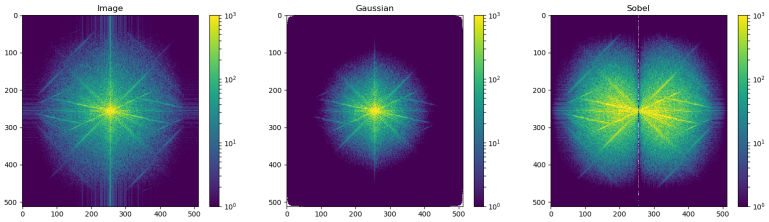
b)

c)

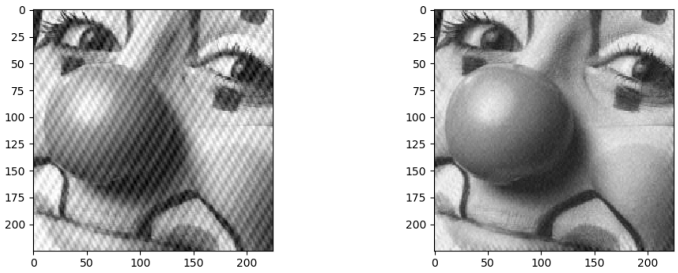
d)



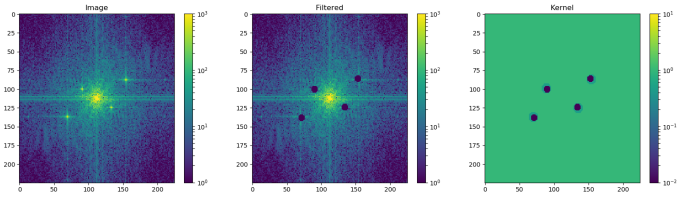
Figur 17: Sobel, 4b). Original to the left.



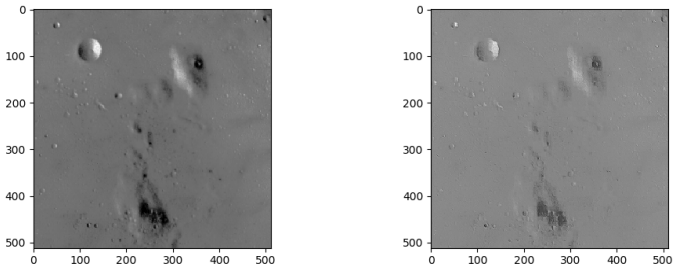
Figur 18: Original and filtered images, 4b).



Figur 19: Clown image convolved with unknown kernel, 4c). Original to the left.



Figur 20: We see that the unknown kernel exactly targets the diagonal lines in the image. Such a filter is called a notch filter, since it targets a very specific and narrow frequency band.



Figur 21: Left: original, Right: sharpened moon image. The contrast is modified because of a few pixels getting extreme values. This could be fixed, but either way details are sharpened a bit.