DLCV HW2

Céline Nauer

October 2019

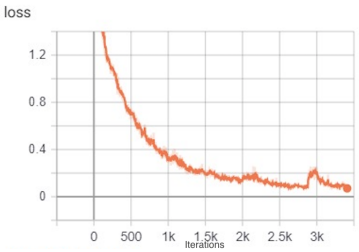
1 Baseline Model

1.1 Data Preprocessing

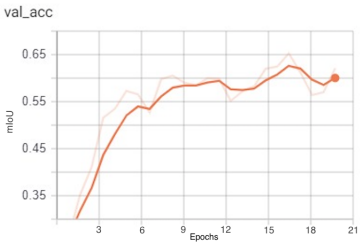
The training as well as the testing images are transformed to tensors and normalized with the following values:

- MEAN = $[0.485, 0.456, 0.406]$
- STD = $[0.229, 0.224, 225]$

1.2 Results

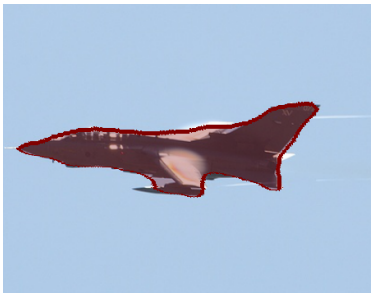


(a) Loss as a function of iterations



(b) miou score as a function of epochs

1.3 Visualisations



(a) Aeroplane

(b) Person

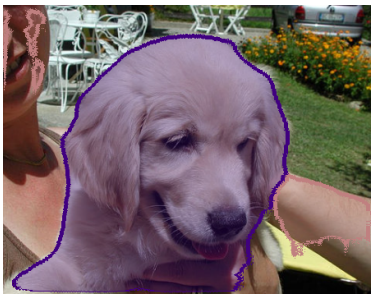
(c) Car

(d) Cat

(e) Horse

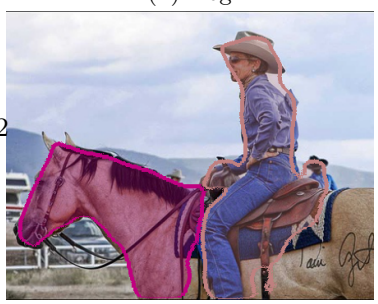
(f) Bus

(g) TV/Monitor



(h) Dog

(i) Multiple categories: Bus and Person



(j) Multiple categories: Horse and Person

1.4 Evaluation

```
class #0 : 0.89826
class #1 : 0.72292
class #2 : 0.66150
class #3 : 0.69556
class #4 : 0.39669
class #5 : 0.58645
class #6 : 0.55187
class #7 : 0.68930
class #8 : 0.67065

mean_iou: 0.652578

Testing Accuracy: 0.6525782590139693
```

The achieved mioU score is approximately 65%. Looking at the evaluation, the class 4 "Monitors" is performing worse than the other categories. One explanation for this behaviour could be that there is a high variability within this class, since monitors can either be turned on or off and thus, may look very different.

2 Improved Model

2.1 Model Architecture

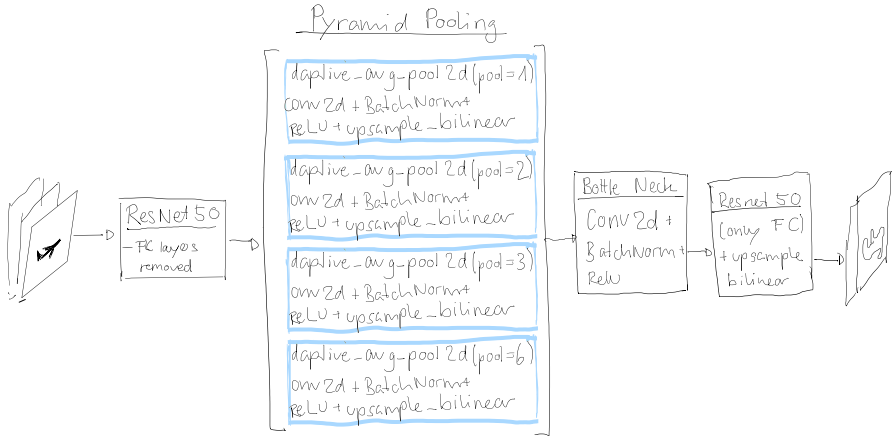
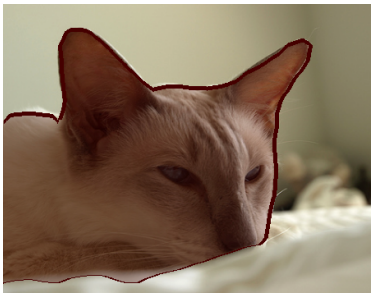


Figure 3: Network Architecture: Resnet layers, Pyramidal Scene Parsing to get sub-region representations, upsampling and concatenation layers to form the final feature representation, convolution layer to get the final per-pixel prediction.

2.2 Comparison to Baseline model

Several sources argue that models such as CNN or ResNet have difficulties to integrate local and global context information. One way to improve the integration of global context is to enlarge the receptive fields and thus, to add pooling layers. In 2016, Zhao et al. [1] suggested to use a Pyramid Scene Parsing Network where different pooling layers are applied to get different sub-region representations. Via a bottle neck, these representations are integrated in order to get global context information. This network has proven to achieve an mioU score of 85.4% when performing segmentation on the PASCAL VOC data set when pre-trained on MS-COCO. Thus it is a suitable network for the task in this homework.

2.3 Results



(a) Cat improved model

(b) Cat baseline model

(c) Bus improved model

(d) Bus baseline model

(e) Monitor improved model

(f) Monitor baseline model

(g) Person on horse improved model

(h) Person on horse baseline model

The above images show a few cases where the improved model clearly performs superior to the baseline model. The evaluation mean IoU is as follows:

```
class #0 : 0.91122
class #1 : 0.76740
class #2 : 0.70270
class #3 : 0.78456
class #4 : 0.40486
class #5 : 0.75102
class #6 : 0.72546
class #7 : 0.77003
class #8 : 0.77661

mean_iou: 0.732650

Testing Accuracy: 0.7326497365173155
```

3 Image Filtering

3.1 Proof that Convolution with Gaussian Filters is separable

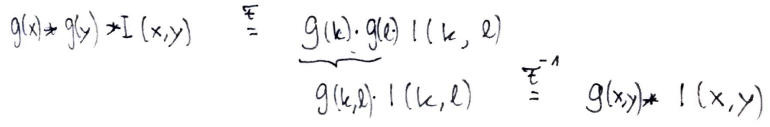


Figure 5: I use the the fact that the Fourier transform F of a Gaussian G applied to image I is a Gaussian and the property that the Fourier transform of a convolution is a multiplication.

3.2 Gaussian Blur

(a) Original image

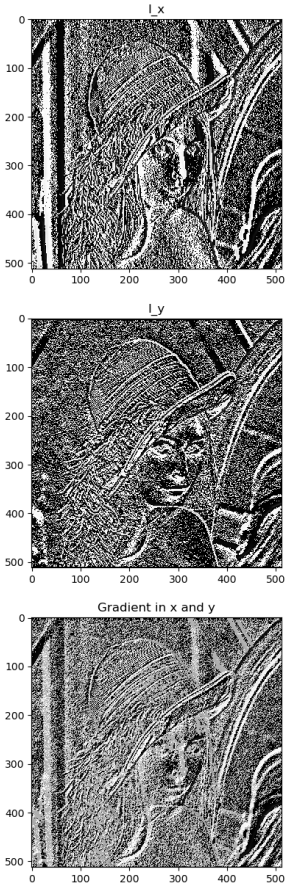
(b) Blurred image

The effect of the shown Gaussian filter is that the image gets blurred. This is an effective technique to reduce high-frequency components in the image and to get less noisy edges and corners when doing feature extraction. This becomes visible in the following example.

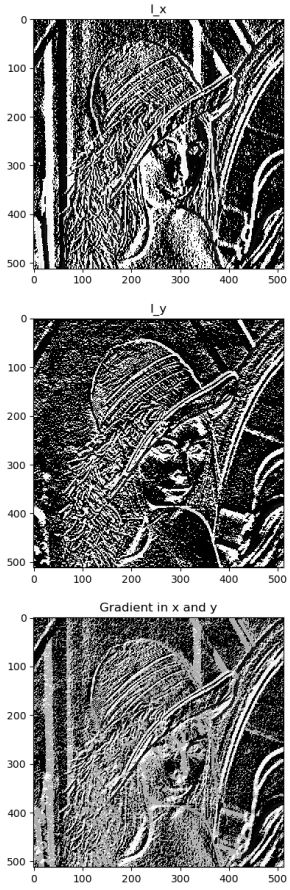
3.3 Derivative and Gradient of the Image

The kernels computed with the given formulae are:

- $k_x = [-1/2, 0, 1/2]^T$
- $k_y = [-1/2, 0, 1/2]$



(a) The derivative in x, in y and the gradient of the original image



(b) The derivative in x, in y and the gradient of the blurred image

The plots above show that the effect of the blurring is the reduction of high-frequency noise and thus, more consistency in the detected edges. However, the trade-off of this method is the loss of information, such as blurred edges and the vanishing of very small objects.

Disclosure

Besides my own contributions, I have used the provided example code from the Pytorch tutorial as well as code from online, which I have specified with links in comments. In addition, I have discussed ideas and approaches to the exercises with Julia Maricalva (A08922117), Carlos Marcal (A08922106), Ricardo Manzanedo (R08942139) and Javier Sanguino (T08901105).

References

[1] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya. Jia. Pyramid scene parsing network. 2016.