Exercise Sheet 6: Image Captioning & ViT

Due on 27.06.2025, 10:00

Olga Grebenkova, Felix Krause (o.grebenkova@lmu.de, felix.krause@lmu.de)

Important Notes:

1. **Submission:** Your submission should consist of one single ZIP file which includes a PDF file and the corresponding codes. Both the PDF file and ZIP file should contain your surname and your matriculation number (Surname-MatriculationNumber.zip) for grading purposes. You may use Jupyter for exporting your python notebooks as PDF files, but you still have to hand in your .ipynb or .py files for us to test your code. For this exercise, please export the .ipynb as a PDF file and include that in the ZIP file. Submissions that fail to follow the naming convention or missed PDF/code files will not be graded.
2. **Deadline:** The due date for this exercise is the 26th of June.

Task 1: Image Captioning (30P)

Image captioning is vital to the field of Computer Vision. In this exercise we will utilize a pretrained VGG19 model to build feature vectors for positions of images. For this we will train an LSTM and draw its attention maps. We will provide you with already vectorized MSCOCO17 images to save you time.
For this exercise we have detailed instruction in the notebook. Please put your code solutions in the predefined areas in the notebook (# YOUR CODE).

Task 2: Vision Transformer (ViT) (10P)

In this smaller part of the exercise you will be asked to implement all critical parts of the wildly used Vision Transformer (ViT).
For this exercise we have detailed instruction in the notebook. Please put your code solutions in the predefined areas in the notebook (# YOUR CODE).