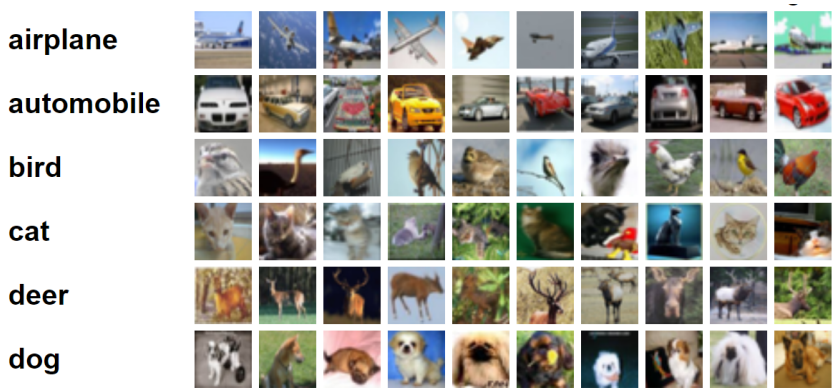
Vision Transformers

Exercise 1.2

Advanced Deep Learning in Computer Vision

February 2026

In this exercise, you are asked to build a vision transformer model for image classification of small images (CIFAR).



Your task is to first complete the implementation of the given vision transformer model, train the model on the given dataset, and evaluate your results illustrating how different choices impact the performance of your model at your given task.

Your tasks are as follows:

1. Complete the implementation of the `image to patches` and the `patch embedding`. (See files `"playground.py"` and `"vit.py"`).
2. Train and evaluate your model on the CIFAR dataset using the `imageclassification.py`.
3. Illustrate how different choices for your model affect the performance of your model at your given task.
4. Visualize the attention map of your trained ViT (see Fig. 6 in the ViT paper).
5. **OPTIONAL**: Visualize the positional encoding and the similarity of learned positional embedding (see Fig. 7 in the ViT paper).
6. Your model, data, process, performance evaluation, and results should be documented and discussed in a PDF (up to 4 pages) to be uploaded on DTU Learn together with exercises 1.1 and 1.3. More details are in the given template.