Nupur Dhananjay Kulkarni
Léonie Muller

Assignment 4
Task 1 Report

Analyzing Network-Dissect

This report gives an overview of the insights obtained on implementing Task 1 from Assignment 4. The main idea of the report is to interpret the Resnet-18 predictions on ImageNet and Places 365 and get insights into the concepts learned using the clip-dissect library. The implementation can be found here. The other visualizations obtained during the experiment like images corresponding to the neurons, comparison plot of concepts learned per layer, and description files generated for explaining both models can be accessed here.

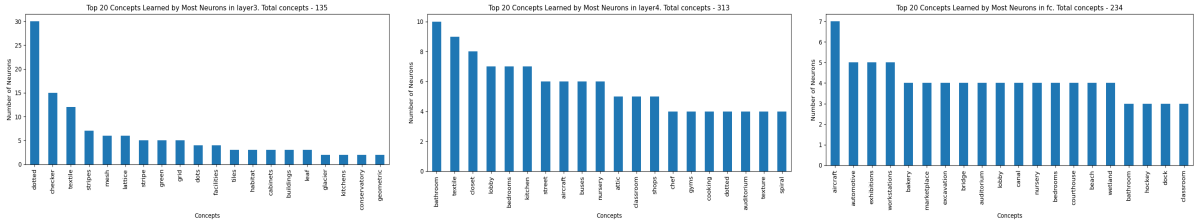

Figure 1: Most concepts learnt - Resnet18 on places 365

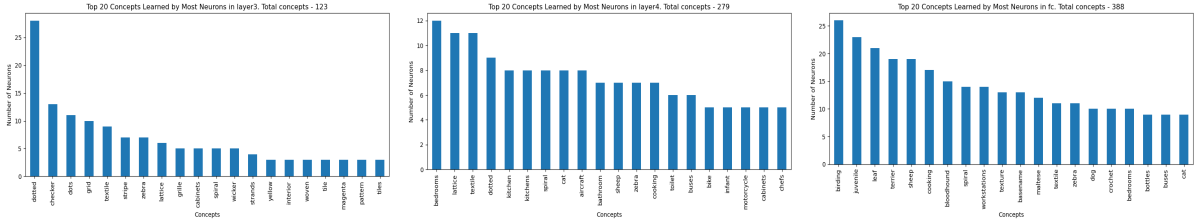

Figure 2: Most concepts learnt - Resnet18 on ImageNet

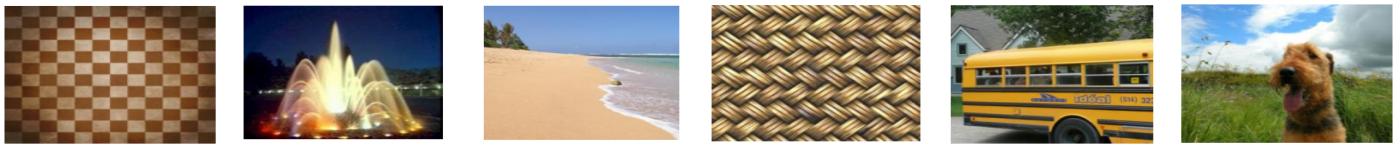

Figure 3: Layer 3, Layer 4, fc - Resnet18 (Places 365) and Resnet18 (ImageNet)

Figures 1 and 2 illustrate the concepts that are learned by most of the neurons. We conducted network dissection only for the last 3 layers of the Resnet-18 model trained on Places 365 and ImageNet - layer 3, layer 4, and the fully connected layer. It is quite intuitive that, neurons in layer 3 of Resnet 18 trained on places 365 capture the fundamental visual features like dotted, checker, stripes, grids, tiles, etc, eventually forming the base to recognize relatively more complex concepts. A similar pattern is observed for Resnet18 trained on ImageNet where we see dotted, checker, spiral, grids, etc. For Layer 4, on the contrary, the focus is more on capturing the high-level concepts such as the bathroom, closet, buses, etc for places 365 and bottles, cat, dog, sheep, etc for ImageNet, thus bringing in some context to the previously learned patterns. The fully connected layer finally adds up the patterns and objects learned in all previous layers and makes an informed classification like aircraft, beach, auditorium, etc for places 365 and object categories like leaf, zebra, dog, bottles, cat, etc for ImageNet. The stark difference in the maximum number of neurons learning one concept (7 for places 365 and 26 for ImageNet) reflects the greater specialization required in object recognition for the ImageNet dataset. This is probably because the model needs to make fine-grained distinctions between many object classes. For Places365, fewer neurons may be sufficient to capture the broader scenes. It is evident that neurons in layer 3 (for both the models) capture quite similar low-level features, however, the subsequent layers start focusing on concrete objects in the case of the ImageNet dataset and scenes in the case of the Places 365 dataset. The first three images in Figure 3 are obtained by plotting the neurons in layers 3, 4, and fc in the Resnet18 trained on places 365). The first image suggests that the neurons in layer 3 capture basic patterns like a checkerboard. The neurons in layer 4 recognize more complex concepts like a fountain and finally, a fully connected layer can capture and summarize the information the model learned so far and identify more comprehensible concepts like beaches, which correspond to one of the categories in the Places 365 dataset. Similarly, for the last three images that are generated by plotting neurons in layers 3, 4, and fc in the Resnet18 trained

on ImageNet, we see layer 3 recognizing concepts like wicker, layer 4 recognizing more complex concepts like a bus, and fc layer capturing all this information and recognizing concepts like a dog. Thus, to summarize, Resnet18 models trained on Places 365 and ImageNet serve a specific purpose of classifying scene-based and object-based images respectively, and can capture features ranging from low-level to complex concepts.