TML Assignment 4: Explainability

Task 1 Report: Network Dissection Analysis

Objective

The goal of this analysis was to investigate the internal representations of two `ResNet18` models using the `CLIP-dissect` library. The last three convolutional layers (`layer2`, `layer3`, `layer4`) were analyzed for a model trained on ImageNet and a model trained on Places365 to understand and compare the concepts learned by their neurons.

Key Findings

1. Most Commonly Learned Concepts: The Primacy of Patterns

This section answers the first question. We found that the most frequent concepts in both models are low-level patterns.

- **ResNet18 (ImageNet):** The top concept is **"dotted" (57 neurons)**, followed by **"textile" (24 neurons)**, **"lattice" (21 neurons)**, and **"checker" (19 neurons)**.
- **ResNet18 (Places365):** Similarly, the top concepts are **"dotted" (47 neurons)**, **"textile" (25 neurons)**, and **"checker" (25 neurons)**.

This overlap suggests that these patterns are essential building blocks for visual understanding in general. The charts below visualize these findings.

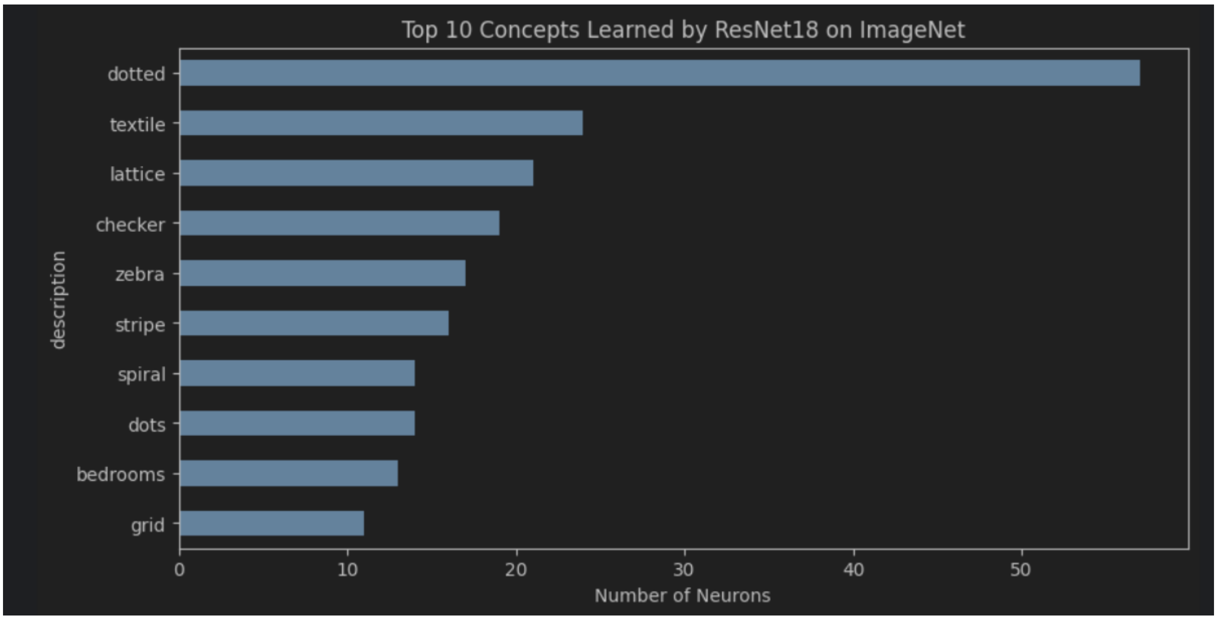

Figure 1: Top 10 concepts learned by neurons in ResNet18 trained on ImageNet.

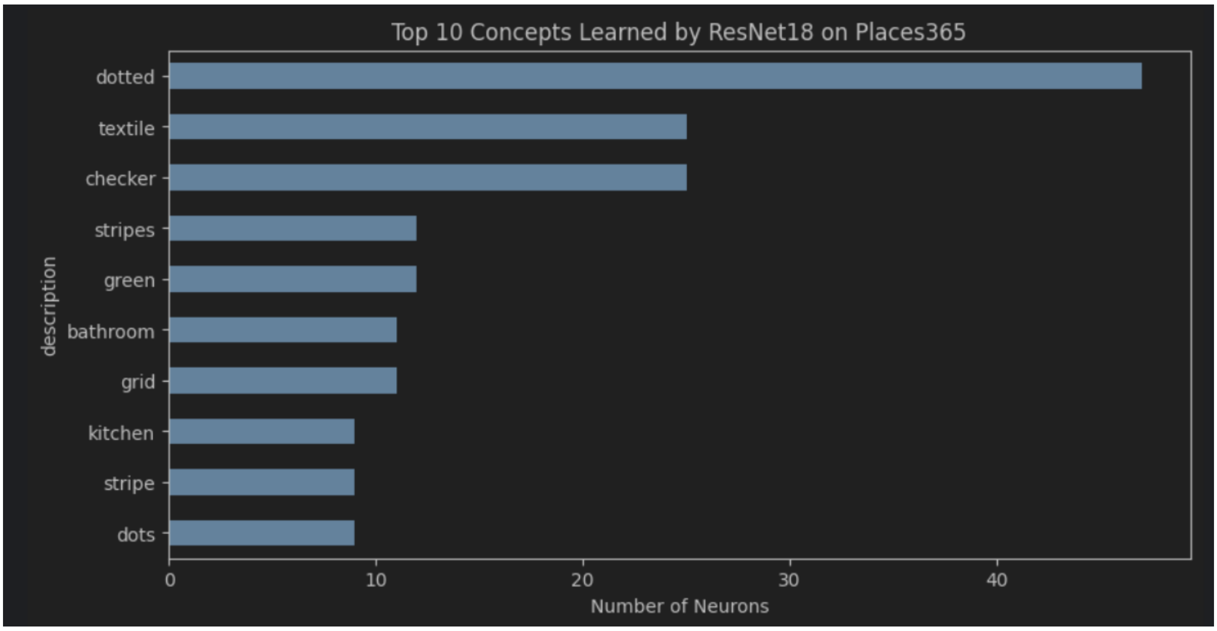

Figure 2: Top 10 concepts learned by neurons in ResNet18 trained on Places365.

2. Comparison and Vocabulary: How Training Data Shapes Specialization

This section answers the second and third questions. While both models prioritize basic patterns, the differences in their learned concepts clearly reflect their specialized training.

- **Vocabulary Size:** The Places365 model learned a slightly larger vocabulary, identifying **426 unique concepts** compared to the ImageNet model's **368 unique concepts**. This could suggest that scene classification requires a more diverse set of feature detectors.

- **Concept Specialization:** The key difference emerges in the top-ranked concepts beyond simple patterns. The **Places365** model dedicates numerous neurons to scene-specific (`bathroom`, `kitchen`) and attribute-specific (`green`) concepts. In contrast, the ImageNet model's specific object detectors are more apparent in the layer-by-layer analysis.

3. Additional Analysis: Hierarchical Learning in Action

This section addresses the fourth question by presenting an additional analysis: a layer-wise comparison within the ImageNet model.

- **Early Layers (`layer2`):** These layers almost exclusively detect simple patterns. The top concepts are **"dotted" (21 neurons)**, **"stripe" (8 neurons)**, and **"checker" (7 neurons)**. This forms the foundation of the network's vision.

- **Later Layers (`layer4`):** This deeper layer combines simple patterns to form complex, semantic concepts. Top concepts here include objects like **"cat"** and **"aircraft"**, and complete scenes like **"bedrooms"**, **"kitchen"**, and **"bathroom"**.

This demonstrates that the network actively constructs a hierarchy of knowledge, starting with simple textures and building up to meaningful, real-world concepts.

Conclusion

The Network Dissection analysis reveals three core insights:

1. A substantial portion of a neural network is dedicated to identifying fundamental, reusable patterns.

2. The model's training objective (object vs. scene recognition) dictates the *type* of high-level concepts it specializes in.

3. Neural networks build their understanding hierarchically, using simple features learned in early layers as the building blocks for complex recognition in later layers.