TML ASSIGNMENT 4

TASK 1

Team 16

Muhammad Mubeen Siddiqui & Ahrar Bin Aslam

Network Dissection Analysis of ResNet18 Models Trained on ImageNet and Places365

Introduction:

The task requires analyzing the internal workings of two ResNet18 models, one trained on ImageNet and the other on Places365. By using Network Dissection, we aim to label the neurons in the last three layers of each model and understand what concepts or classes are learned by these neurons. This analysis will provide insights into how each model specializes in different types of learning, focusing on the neurons responsible for learning specific classes in each model.

GitHub Repo Link: https://github.com/ahrarbinaslam/TML25_A4_16

Methodology:

Data and Models:

- ResNet18 trained on ImageNet: A model trained on 1,000 object categories.
- ResNet18 trained on Places365: A model trained on 365 scene categories.

Neurons Labeling: The last three layers of each model were dissected to label neurons with learned concepts from a predefined set of 20,000 concepts.

Neuron Dissection: The **CLIP-dissect** tool was used to assign labels to the neurons based on their activations.

Comparative Evaluation: We compared the concepts learned by the two models and calculated Jaccard similarity scores for overlap in learned concepts.

ANALYSIS:

ResNet18 on Places365

1. Concept Coverage

- **Places365 Model** learned **427 unique concepts**. □ **ImageNet Model** learned **374 unique concepts**.
- While the total number of unique concepts was comparable, the types of concepts varied significantly:
  - ImageNet emphasized **objects**, materials, and textures.
  - Places365 focused more on **scenes**, environments, and background elements.

ResNet18 on Places365

The ResNet18 model trained on Places365 is clearly focused on recognizing scenes rather than individual objects. In the earlier layers, like layer2, the neurons respond to basic visual features such as colors (like red or turquoise) and patterns (like dotted, tribal, or knots). These features help the model understand the overall environment of an image, whether it's indoors, outdoors, a forest, a beach, or a room. The visualizations of neuron activations show that these neurons light up in specific areas of the image, suggesting that they are detecting textures or background elements. As we go deeper into the model (layers 3 and 4), the neurons start recognizing more complex scene features. The heatmaps also show that deeper layers detect more abstract, scene-specific patterns.

ResNet18 on ImageNet

On the other hand, the ResNet18 model trained on ImageNet focuses more on identifying objects. In layer2, many neurons are tuned to detect object textures and surface details like hoodia, lattice, dotted, and stripes. These are useful for distinguishing different kinds of objects such as animals, tools, or clothing. The activation maps show that these neurons respond to specific parts of objects in an image, which is helpful for object classification.

Just like with the Places model, the deeper layers (layer3 and layer4) in this model start recognizing more abstract object features. The heatmaps confirm this by showing stronger and more detailed patterns linked to complex object understanding

.Comparison Between Both Models

- A Venn diagram of shared concepts shows an overlap of 180 concepts, indicating some common patterns in both datasets, such as textures and basic shapes.

- The Jaccard similarity between the two models' learned concepts shows that there is a moderate overlap in concepts, particularly in layer 4, with a similarity score of 0.233. This suggests that both models capture similar high-level features in their last layers, but each has distinct specialization based on its training dataset.

  Heatmap Insights:

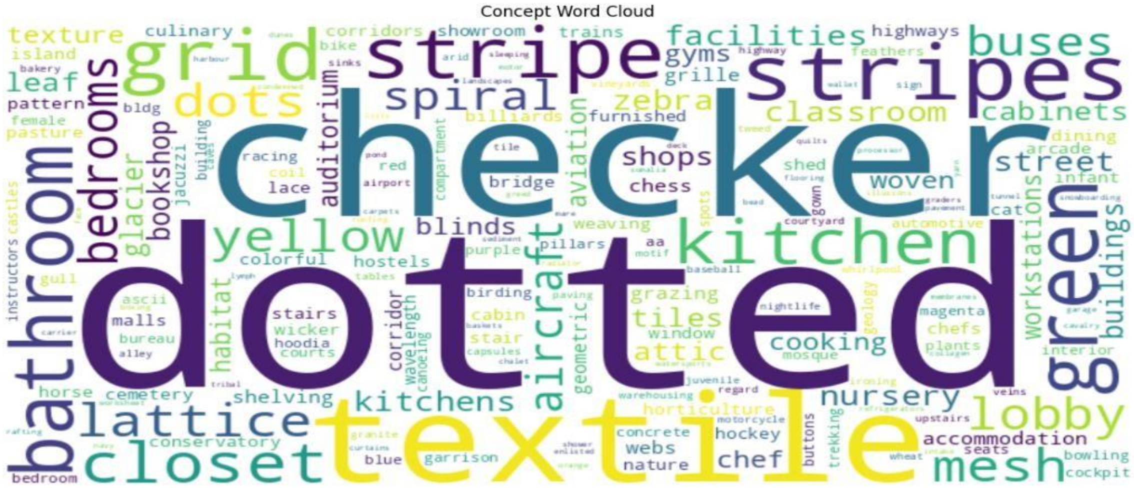
  ResNet18-Places365 : Highlights high frequencies for “dotted” (30 in layer3) and “checker” (15 in layer3), with scene-specific concepts like “kitchen” (7 in layer4) and ‘bathroom” (10 in layer4) dominating deeper layers, reflecting a shift toward scene layouts.

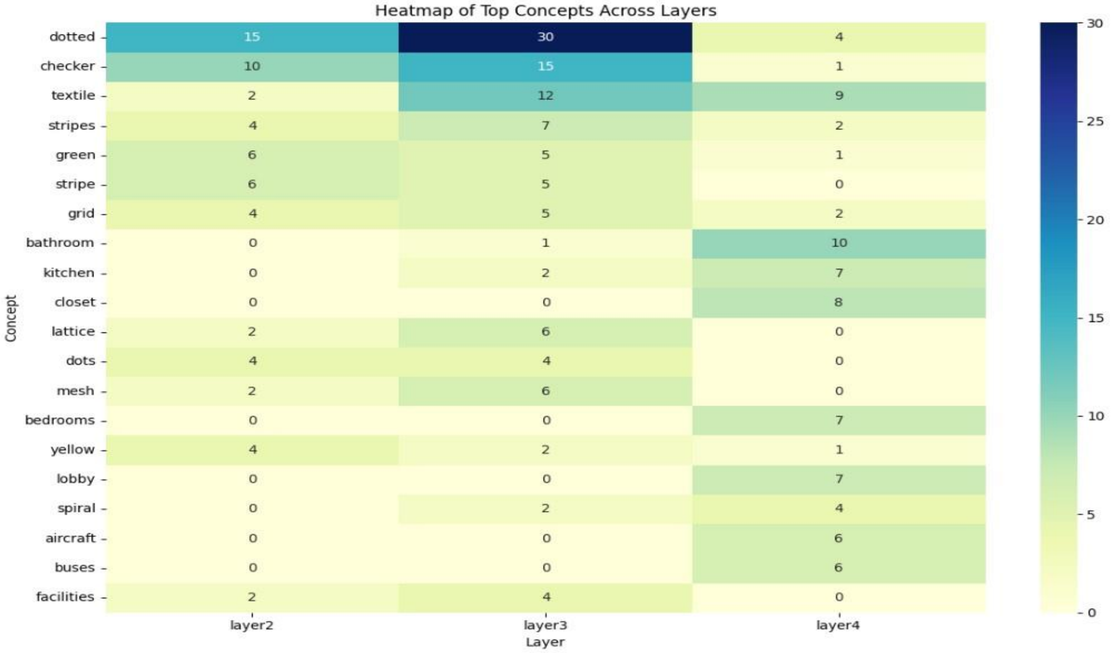
  ResNet18-ImageNet: Shows high frequencies for “dotted”(30 in layer2, 20 in layer3), “textile” (21 in layer4), and “lattice” (11 in layer4), indicating a focus on object textures and patterns that persist or intensify in deeper layers.

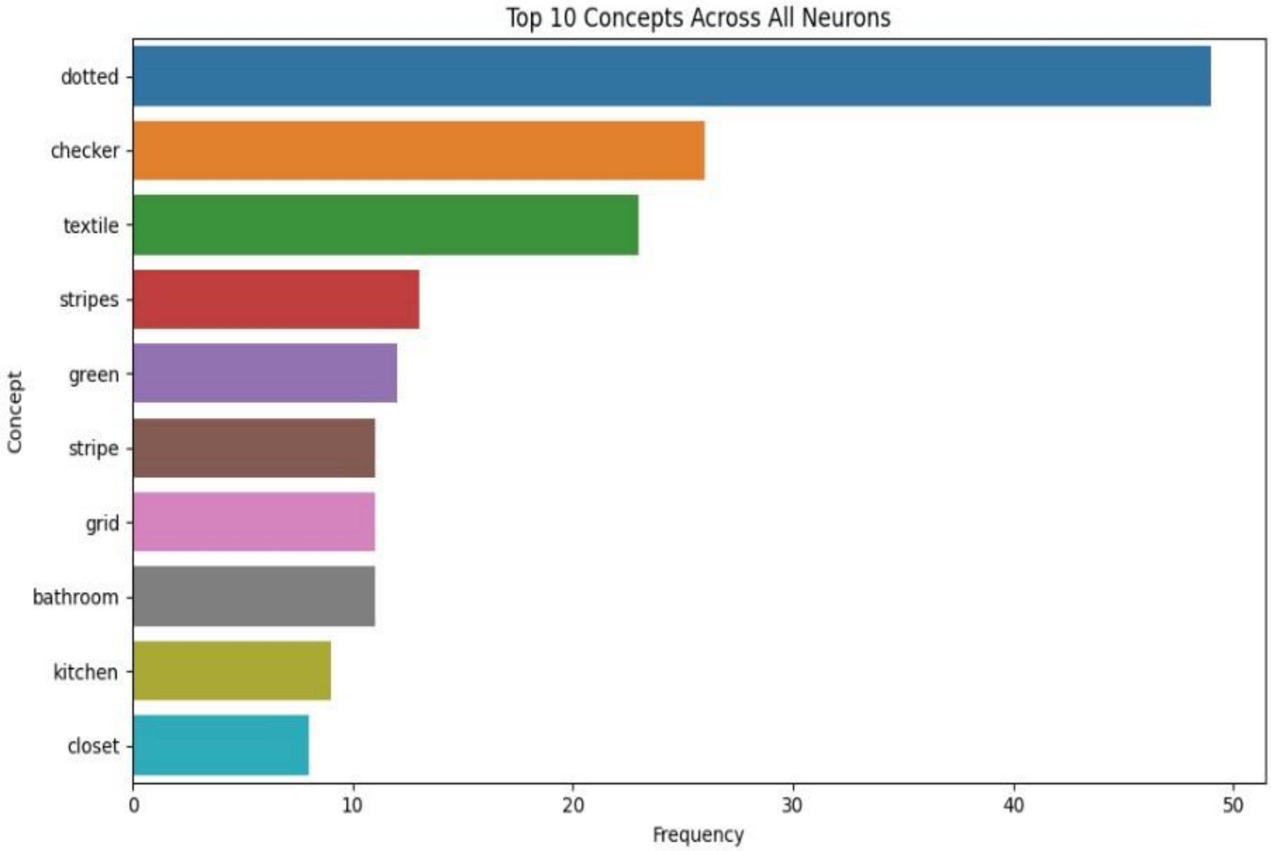
  The heatmaps reveal that while layer2 shares concepts like "dotted" and "checker," deeper layers diverge, with Places365 emphasizing scene elements like “kitchen,” “bathroom” and ImageNet focusing on object details like “textile,” “lattice”.

VISUALIZATIONS

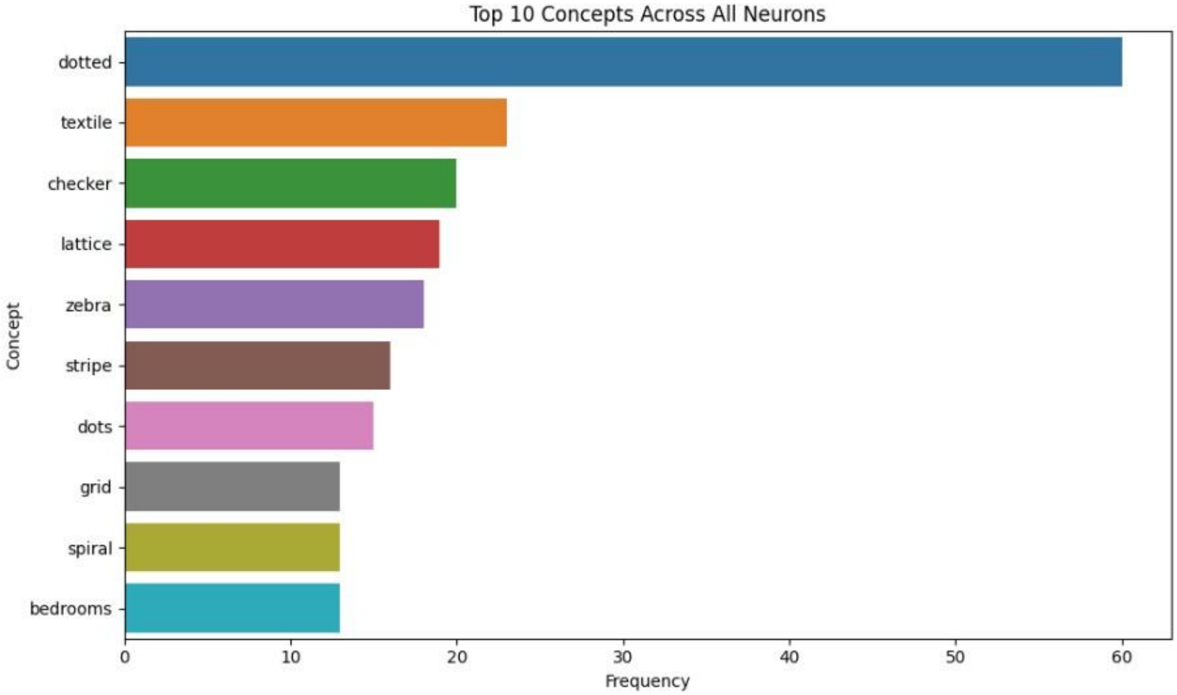
Resnet18 on Places365

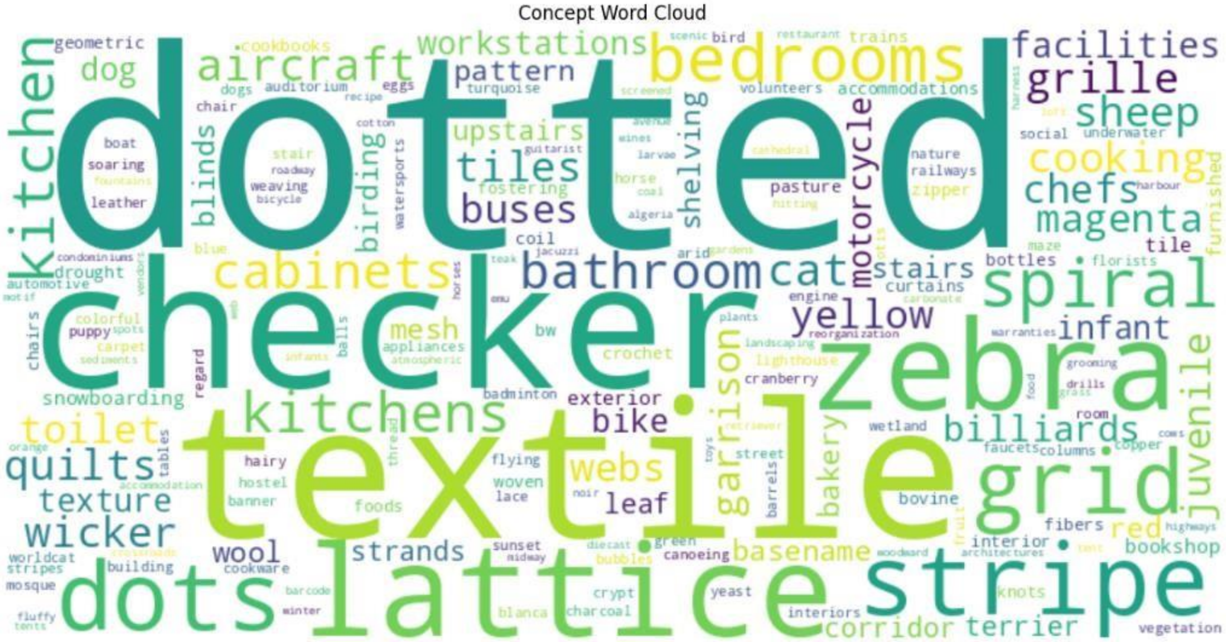


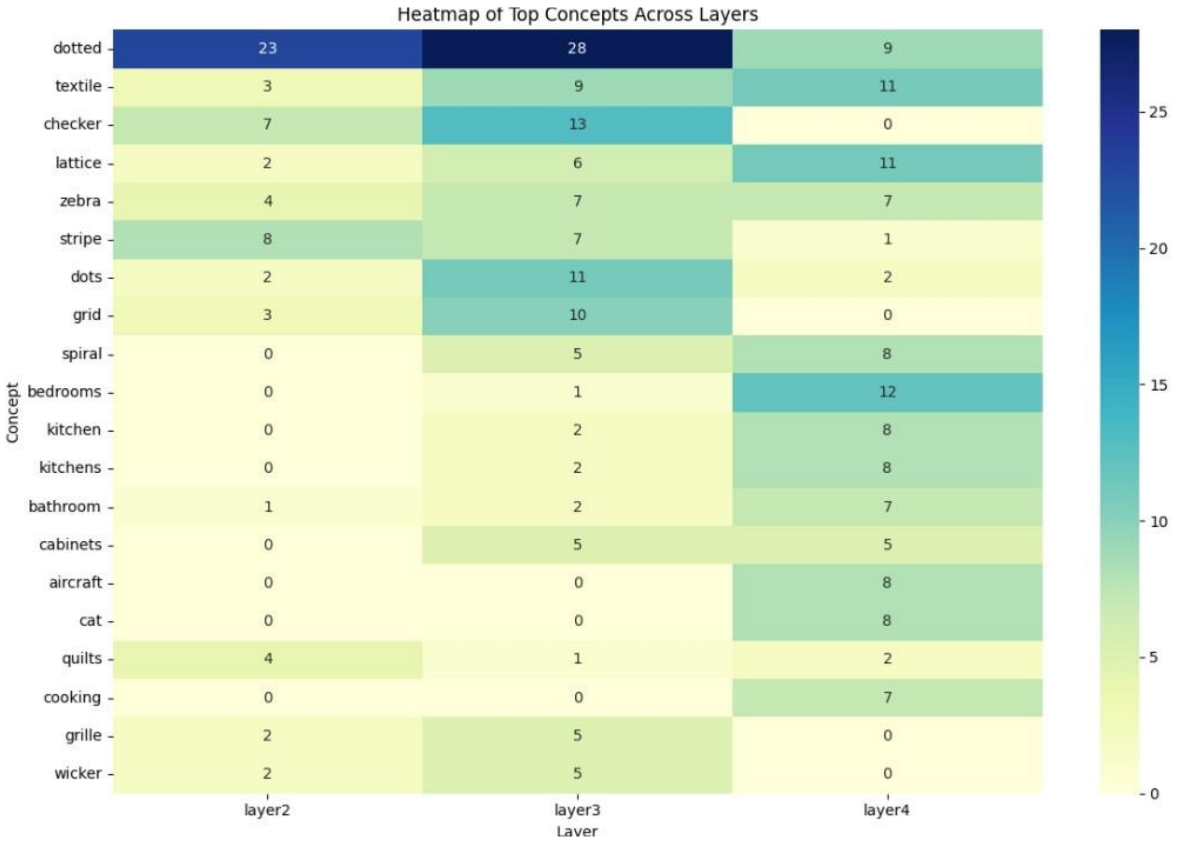



Top 10 Concepts Across All Neurons
dotted
checker
textile
stripes
green
stripe
grid
bathroom
kitchen
closet
Concept
0
10
20
30
40
50
Frequency

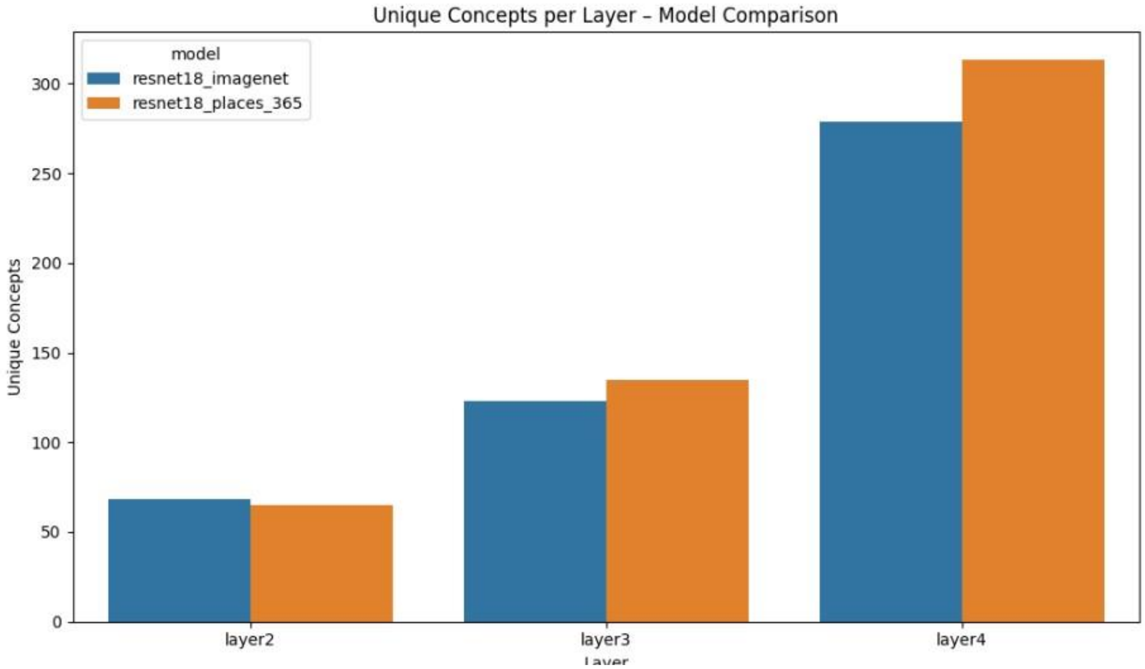
Resnet18 on Imagenet

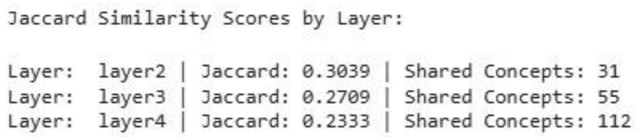






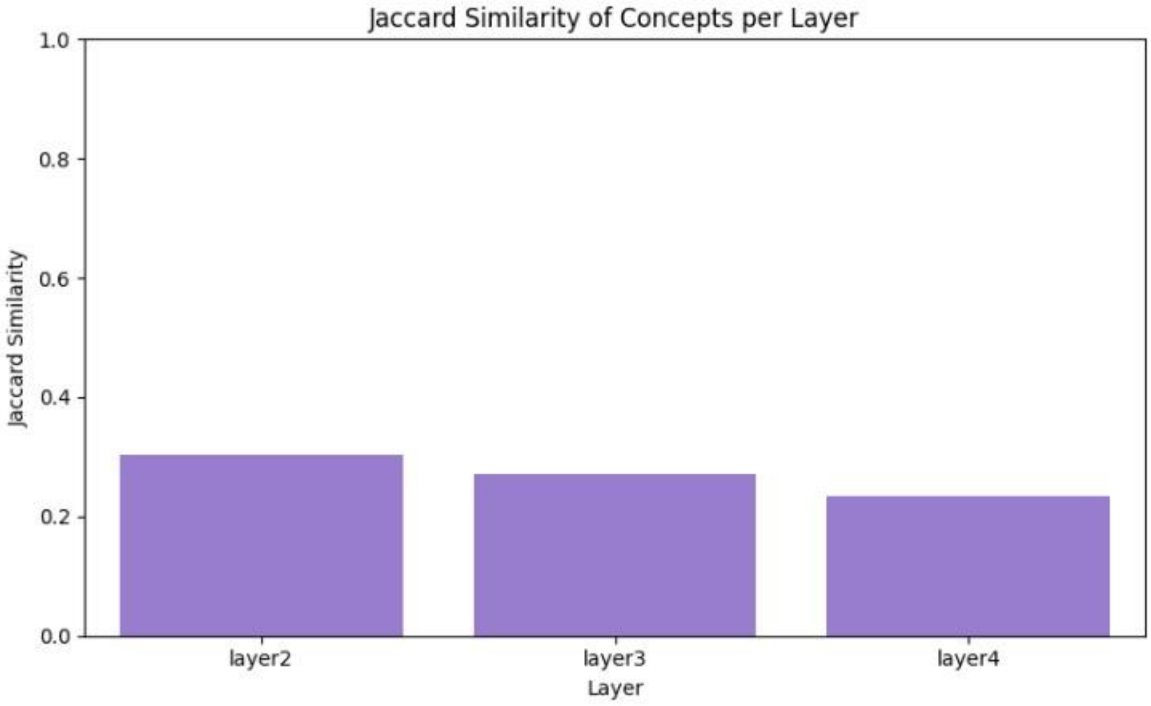
Comparision between both

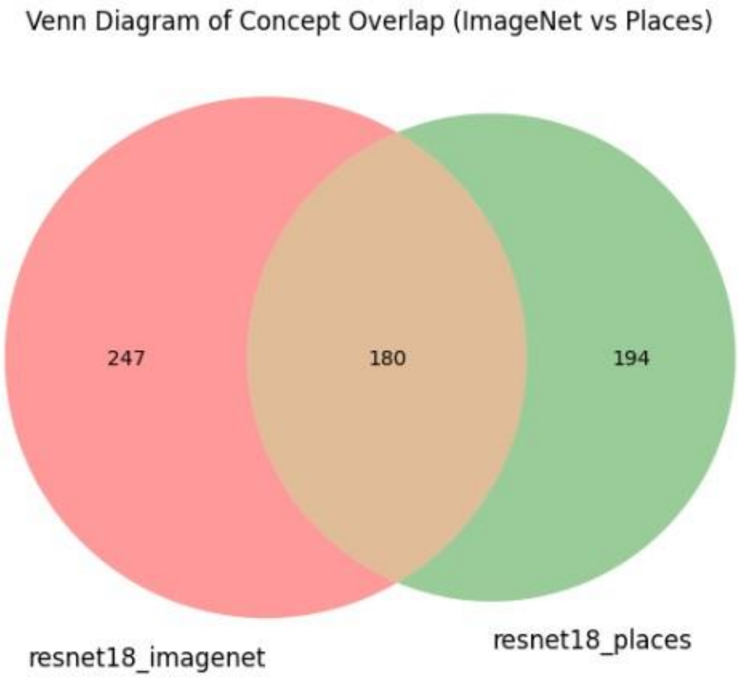




```
Jaccard Similarity Scores by Layer:

Layer:  layer2 | Jaccard: 0.3039 | Shared Concepts: 31
Layer:  layer3 | Jaccard: 0.2709 | Shared Concepts: 55
Layer:  layer4 | Jaccard: 0.2333 | Shared Concepts: 112
```


Jaccard Similarity of Concepts per Layer
Jaccard Similarity
layer2
layer3
layer4
Layer


Venn Diagram of Concept Overlap (ImageNet vs Places)
resnet18_imagenet
resnet18_places