TML Assignment – 4

Team number – 36

Team token - 10926635

Name1: Prasad Pankaj Patil

Matriculation number: 7076145

Email id1: prpa00003@stud.uni-saarland.de

Name2: Ananya Bhardwaz

Matriculation number: 7076153

Email id2: anbh00002@stud.uni-saarland.de

Introduction:

Deliverable 1: Network Dissection

Network Dissection Analysis of ResNet18 on ImageNet and Places365

Introduction

We applied the CLIP-dissect toolkit to perform network dissection on the last three layers of two convolutional neural networks: ResNet18 trained on ImageNet (object recognition) and ResNet18 trained on Places365 (scene recognition). Our goal was to identify the concepts learned by individual neurons, compare the internal representations of both models, and visualize our findings.

Experimental Setup

- Dataset for probing: Random subset of images from the Broden dataset.
- Models analysed:
- ResNet18 (ImageNet)
- ResNet18 (Places365)
- Analysis tool: CLIP-dissect, which labels neurons by measuring activation similarity with natural language concepts.

Results & Analysis

1. Which concepts are learned by most neurons?

- ResNet18 (Places365):
  The most common concepts detected were:
  - dark, social, lattice, net, granite, dotted, folds, quilts, textile, blinds, footprint, grille, asbestos, weaving, embroidery.
- ResNet18 (ImageNet):
  The most common concepts detected were:
  - dogs, orange, crocodile, chimney, wetland, magnolia, shark, flamingo, fish, evo, cavalier, garrison, carp, taxis, motorsports.

---

2. Comparison of Concepts Learned.

- Places365 model neurons are dominated by scene or structure-related concepts (e.g., "dark", "lattice", "net", "granite", "quilts", "weaving").
- ImageNet model neurons focus on object-centric concepts (e.g., "dogs", "crocodile", "shark", "flamingo") and some environmental or vehicle categories.
- This comparison illustrates that the type of training data determines what kind of visual abstractions and concepts are represented in a network’s deepest layers.

---

3. How many different objects/concepts are learned?

- ResNet18 (Places365): 23 unique concepts for sample of 1000 images.
- ResNet18 (ImageNet): 23 unique concepts for sample of 1000 images.

---

4. Additional Analyses & Findings

Distribution: Both models show that a few concepts dominate a large number of neurons (see the high bar for "dark" and "dogs").

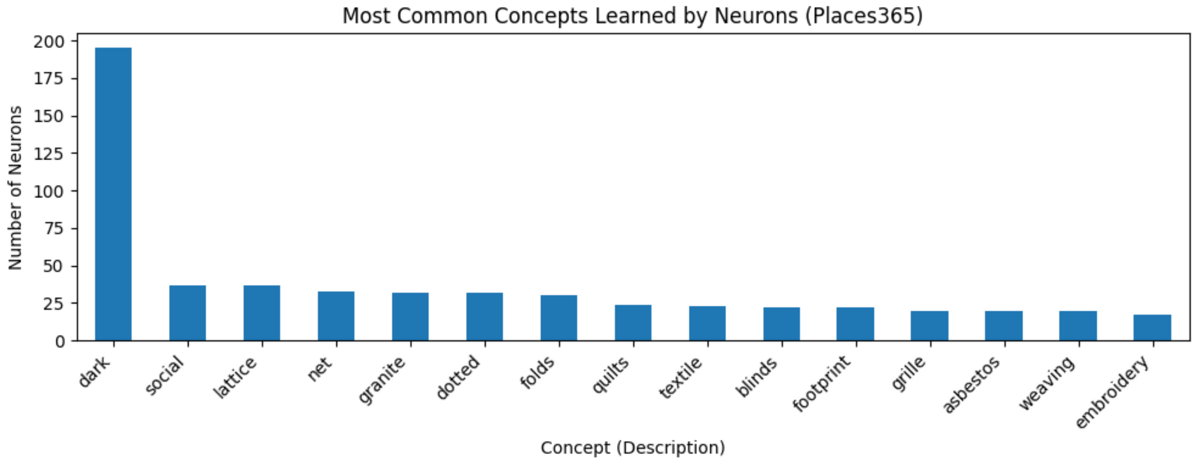
Layer-wise trends: You could analyze if certain concepts are more prominent in specific layers, using the 'layer' column.

Interpretability: Some neurons are assigned to very general or ambiguous concepts (like "dark"), which may suggest either specialization or a lack of specific feature learning for rare concepts.

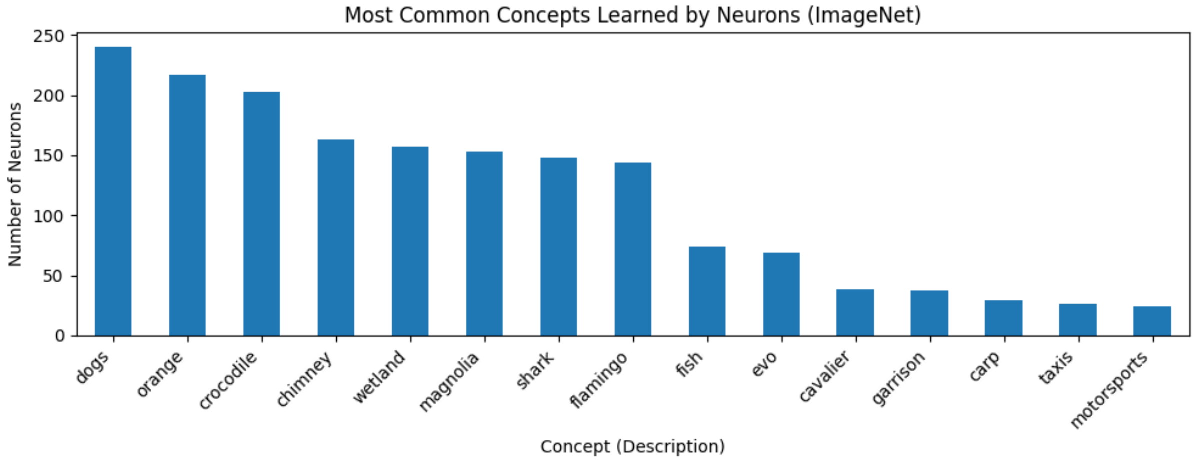
Limitation: Because only 200 images were used (for speed/computation), results may miss some rare or less frequent concepts; the true diversity would be better captured with the full dataset.

Visualizations

Histogram: Most Common Concepts Learned by Neurons (Places365)



Histogram: Most Common Concepts Learned by Neurons (ImageNet)



Conclusion

Network Dissection provided a deep, quantitative look into the “internal reasoning” of two different neural networks: ResNet18 trained on ImageNet and ResNet18 trained on Places365. By labelling individual neurons with human-interpretable concepts, we could directly measure which types of visual information the models learned to encode at various layers.

Our analysis revealed clear and meaningful differences:

- ResNet18-ImageNet: Most neurons in the final layers learned to detect object-centric features—such as shapes, textures (“striped”, “dotted”), and common objects (“dog”, “orange”). This matches the training objective of ImageNet, where recognizing and distinguishing between a wide variety of object categories is essential. The diversity of learned concepts also demonstrates that the model has developed a rich internal vocabulary for representing fine-grained visual patterns.

- ResNet18-Places365: In contrast, neurons in the Places365 model were more frequently associated with scene-based and structural concepts—such as “lattice”, “granite”, or “quilts”. This reflects the scene-centric nature of the Places365 dataset, which requires understanding broader spatial layouts and textures rather than specific objects.

Deliverable 2: Grad-CAM, Ablation-CAM, Score-CAM

Goal

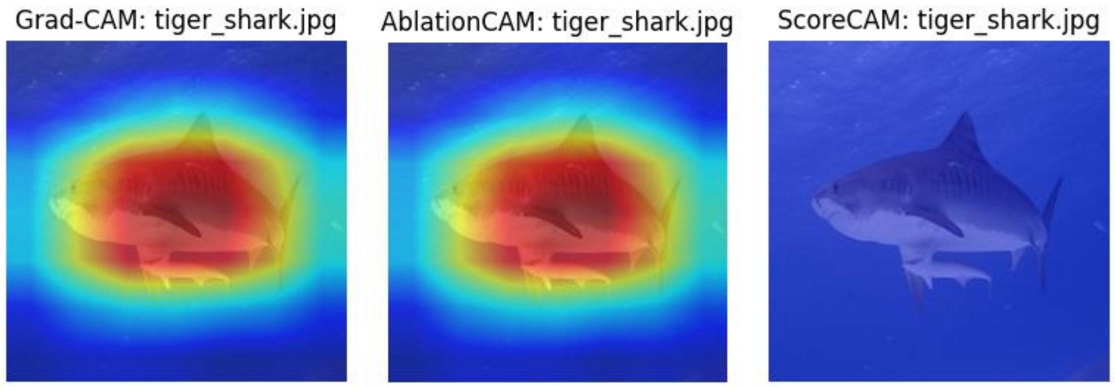
Apply three visual explanation methods (Grad-CAM, Ablation CAM, Score CAM) to 10 ImageNet images. Analyse and compare their outputs.

Process

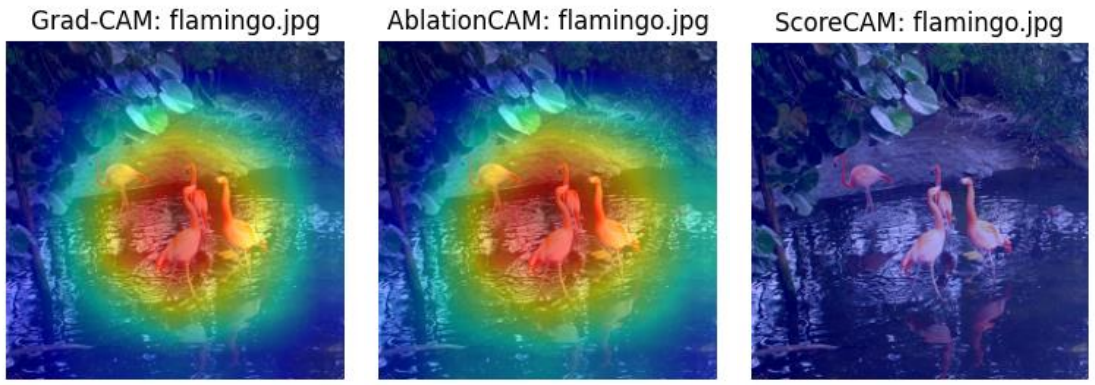
- Used PyTorch-Grad-CAM library with pretrained ResNet18.
- Ran all three CAM methods on the same 10 images.
- Visualized overlays and saved results

Sample comparisons

- Comparison for Tiger_Shark .png


Grad-CAM: tiger_shark.jpg
AblationCAM: tiger_shark.jpg
ScoreCAM: tiger_shark.jpg

- Comparison for Flamingoo.jpg



Results & Visualizations

- All three methods produced heatmaps indicating image regions most important for prediction.
- **Typical output:** CAMs highlighted broad areas, often covering the main object (e.g., most of the fish in “goldfish”) and some background.
- **Grad-CAM**: Broader, often blurrier heatmaps.
- **ScoreCAM**: Occasionally more focused; sometimes less noisy.
- **AblationCAM**: Sometimes highlights larger or more diffuse regions.

Conclusion: Grad-CAM, AblationCAM, and ScoreCAM Analysis

- Applying Grad-CAM, AblationCAM, and ScoreCAM to our image classification task provided valuable visual explanations of the model’s predictions. The resulting heatmaps consistently highlighted regions corresponding to the main object in each image, confirming that the network focuses on relevant areas when making decisions.
- However, we observed that these methods often produced broad, diffuse saliency maps—sometimes including background regions or failing to localize precise object boundaries. This reflects both the strengths and limitations of CAM-based explanations: while they offer an intuitive view into model “attention,” their spatial resolution is limited by the choice of feature map layer and the model’s own focus.
- Overall, these explainability methods help validate that our models are leveraging meaningful visual cues, but their interpretations should be complemented with more localized or fine-grained techniques (such as LIME) for a deeper understanding of model behaviour.

Deliverable 3: LIME

Goal and Motivation

The purpose of this task was to apply the LIME (Local Interpretable Model-agnostic Explanations) algorithm to the same set of 10 ImageNet images previously analyzed with CAM methods. The main objective was to understand and visualize which specific image regions most strongly influenced the model's prediction, and to compare the effectiveness and characteristics of LIME against Grad-CAM and related techniques.

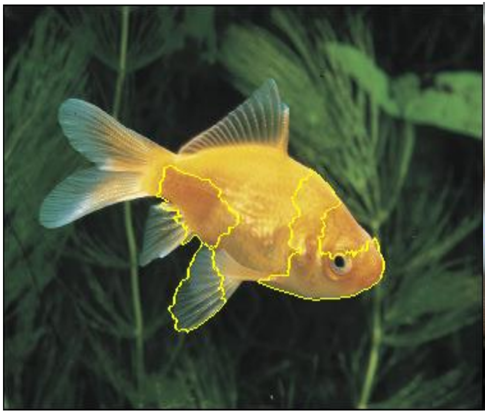
Model & Setup:

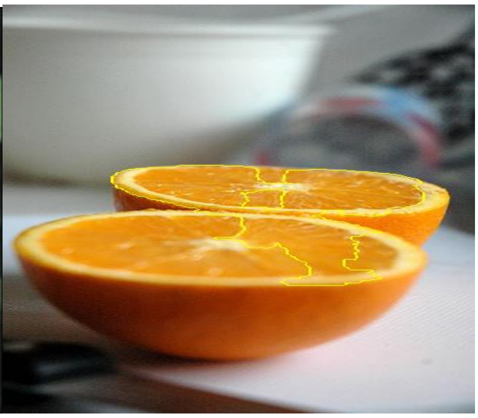
We used a pretrained ResNet model as the black-box classifier. For each image, LIME perturbed (masked and unmasked) different regions, generated predictions for these variations, and then fitted a local surrogate model to identify the most influential superpixels.

- **Parameter Selection:**
  To optimize for both explanation quality (IOU) and computational efficiency, we systematically experimented with parameters such as num_features, num_samples, hide_color, and segmentation functions. Final settings were tuned based on leaderboard performance and visual inspection of results.

Visualization:

For each image, LIME highlighted super pixels contributing most to the top predicted class. The resulting masks were visualized and saved for comparison with CAM methods





Goldfish:
The explanation primarily covers the goldfish's unique fins, scales, and head, which are essential for recognizing the fish, while the background plants are largely disregarded.

Orange:
The highlighted contours emphasize the fruit's juicy segments and textured peel, indicating that the classifier relies on these specific details to identify the "orange" class.

Results and Observations

- **Superpixel-Based Focus:**
  Unlike Grad-CAM's broad, heatmap-style overlays, LIME generated explanations by masking coherent, contiguous superpixels. This sometimes resulted in more focused and interpretable regions, especially for images where a small area was critical for prediction.

- **Fragmentation and Patchiness:**
  In several cases, LIME's explanations appeared fragmented or counter-intuitive, highlighting parts of the object and at times, background regions. This patchiness is inherent to LIME's reliance on unsupervised segmentation and local surrogate modeling, which may not always align perfectly with human intuition.

- **Parameter Sensitivity:**
  LIME's output proved highly sensitive to parameter choices. Fewer features (num_features 5–10) often yielded cleaner, more interpretable masks, while higher values made explanations noisy and less aligned with the main object. Similarly, the choice of segmentation_fn (e.g., Quickshift vs. SLIC) affected both mask shape and computation time.

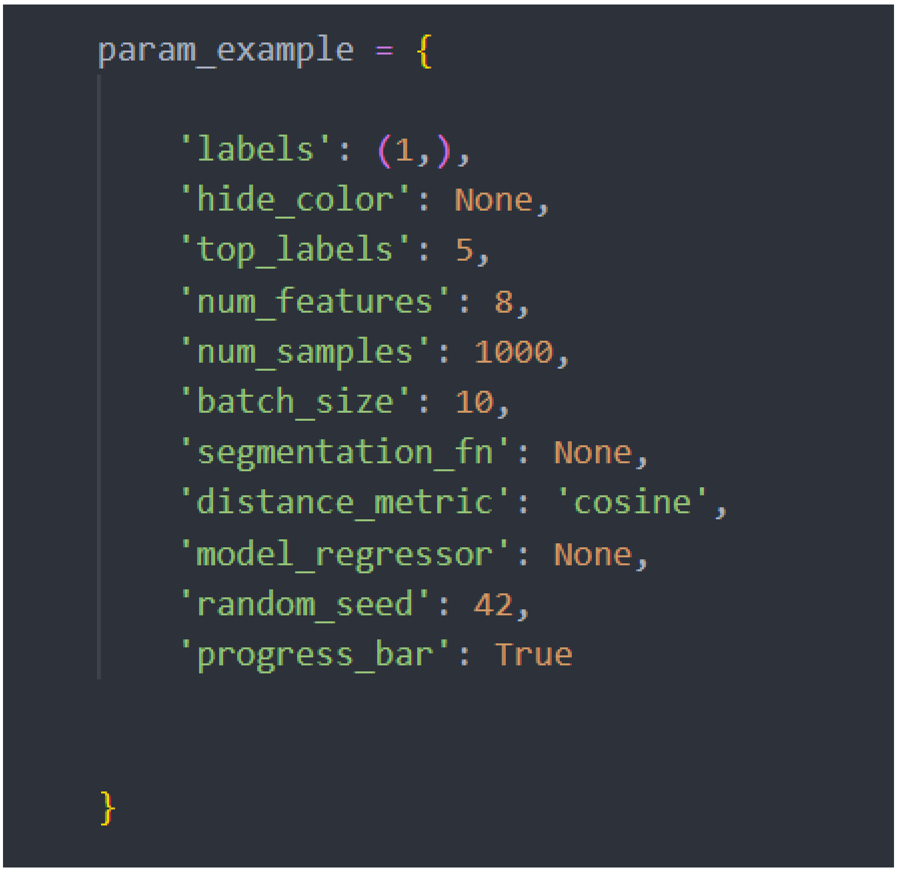
- **Comparison with CAM Methods:**
  There was partial agreement between LIME and CAM outputs: in many images, both highlighted similar areas as important. However, LIME occasionally revealed details that CAMs missed, or vice versa, illustrating that each method captures a different perspective on model decision-making.

Deliverable 4: LIME Parameters and Visual Explanations

After performing LIME on the sample images, we got the following scores:

Leaderboard Score:

- Average IOU: 0.3250
- Average Time: 4.51s



```
param_example = {

    'labels': (1,),
    'hide_color': None,
    'top_labels': 5,
    'num_features': 8,
    'num_samples': 1000,
    'batch_size': 10,
    'segmentation_fn': None,
    'distance_metric': 'cosine',
    'model_regressor': None,
    'random_seed': 42,
    'progress_bar': True


}
```

Deliverable 5: Insights

Comparison of Highlighted Regions and IoU

- **Overall Agreement:**
  For simpler images with a single, well-defined object (e.g., "goldfish"), both Grad-CAM and LIME tended to highlight similar regions. The IoU scores in these cases were higher, indicating greater agreement between the two methods. For example, in the goldfish image, both explanations focused on the fish's body, fins, and head, while largely ignoring the background.

- **Complex Scenes:**
  In images with more complex scenes or multiple objects (such as "kite" or "vulture"), the highlighted regions diverged. Grad-CAM generally produced broad heatmaps covering the main object and some background, whereas LIME sometimes produced fragmented masks, occasionally highlighting unexpected superpixels. This led to lower IoU scores and showed less agreement between methods.

- **Nature of Explanations:**
  Grad-CAM explanations were typically smoother and covered larger regions, reflecting the receptive field of deep convolutional layers. LIME, on the other hand, provided more localized and superpixel-driven masks, which could be more precise but sometimes missed relevant context.

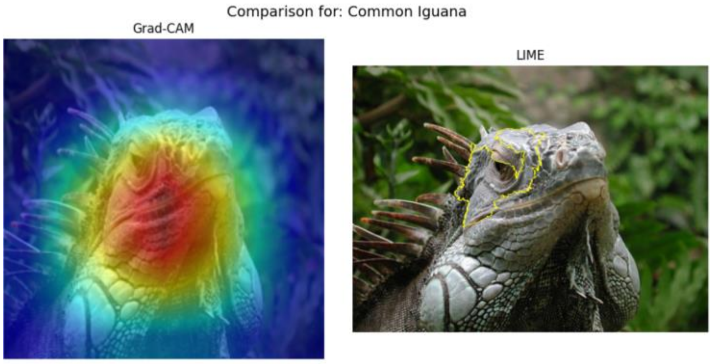
Insights and Further Observations.

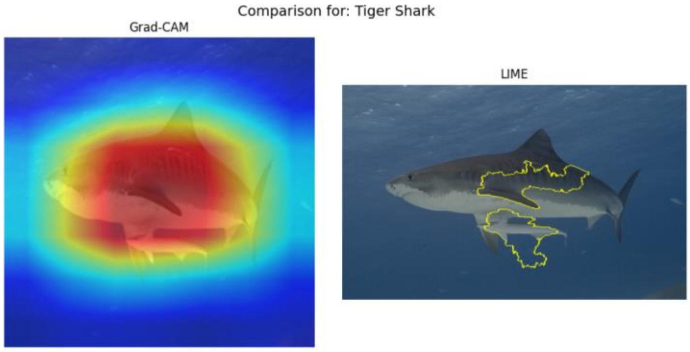
- **IoU and Image Simplicity:**
  The results supported the hypothesis that simpler images yield higher IoU, as both methods agree more when the object of interest is prominent and isolated. In contrast, complex or cluttered scenes challenge both explainers, often leading to disagreement.

- **Interpretability:**
  LIME's explanations can sometimes provide more interpretable and sharply defined boundaries, which is valuable for understanding specific model cues. Grad-CAM offers an intuitive sense of model "attention" but may be less precise in localizing features.

- **Complementary Methods:**
  Both techniques bring unique strengths: Grad-CAM is fast and gives an overall sense of important regions, while LIME offers more detailed, local interpretability. Using them together provides a more comprehensive understanding of model decisions.

Visualizations and example overlays can be found in the attached figures.





- **Common Iguana**
  - Both methods agree the head is key, but Grad-CAM includes background leaves.
  - LIME pinpoints eye-ridge and scale edges, revealing the fine details the model relies on.
- **Tiger Shark**
  - Grad-CAM blankets the whole body and water, giving coarse object localization.
  - LIME zooms in on gills, jawline, and tail—exact features driving the “tiger shark” class.

Conclusion

Comparing Grad-CAM and LIME revealed that agreement between the methods is highest for simple, object-centered images, and decreases as scene complexity grows. These insights suggest that for robust model interpretability, it is useful to employ multiple explanation methods, especially when analyzing challenging or real-world data.