Neural Networks explainability: report

Nursulu Sagimbayeva, Dinesh Adhithya

July 2024

1 CLIP-dissect

For Network dissection, we compared 3 models: ResNet18 trained on Places dataset, and ResNet18 and Resnet50 trained on ImageNet. For probing, we used a subsample of ImageNet consisting of 1000 images and inspected layer 3, layer 4, and fully-connected (FC) layers of all the models above. As for concepts, we used a list of 20000 words provided in the repository of CLIP-dissect. We used a soft-wmpi similarity metric, since it was observed to produce the best estimations by the authors of CLIP-dissect paper.

As expected, the highest similarity scores with concepts were found in FC for all models, since this layer is responsible for the final class prediction and therefore should learn more specific concepts (see Fig 1). Therefore, in our further analysis, we will refer to the units from the Fully-Connected layer. Another observation is that ResNet18 trained on the Places dataset has notably lower similarity with concepts from the Imagenet probing dataset than ResNet18 trained on Imagenet.

Highest similarity score by layer	ResNet18-Places	ResNet18-Imagenet	ResNet50-Imagenet
FC	0.149	**0.223**	0.201
Layer 4	0.156	0.144	0.183
Layer 3	0.132	0.167	0.148

On the concept analysis, for each model we took top-50 neurons from FC layer with the highest similarity scores, and analyzed the distribution of concepts learned (see Appendix A). Interestingly, in both ResNets trained on ImageNet, the concepts learned by most neurons were ”Insect” and ”Terrier”, as well as other objects (”Spider”, ”Pelican”, etc.). However, for ResNet18 trained on the Places dataset, the most frequent concepts were ”Wildlife” and ”Bedroom”, probably because of the specifics of the Places dataset - it captures more scenes than objects. Interestingly, ”Insects” and ”Spider”, as well as ”Dog” and ”Terrier” were learned as distinct concepts. The concept of ”Dog” was learned by especially many neurons in both ResNets trained on ImageNet. Wildlife was also a popular concept, and interestingly enough, it also had the highest similarity for ResNet18-Places (Fig A.

There was another concept that caught our attention, namely, ”Deg”, which we couldn’t interpret. Possibly, it is a variation of ”Dog”, but we don’t dive deeper into the analysis.

Regarding the number of unique concepts, as expected, ResNet50 captured the most concepts (more than 2 times as much as ResNet18 Places). ResNet18 trained on ImageNet also captured notably more concepts than the one trained in Places, which is explained by the difference in training and probing dataset distributions.

Additionally, we observed that similarity values in neurons increased as we increased the size of the probing dataset from 20 to 1000 samples. This is probably explained by the fact that the more images the model is exposed to, the more pronounced and differentiated neuron similarities to concepts are.

2 LIME

We ran LIME analysis on 10 images from ImageNet. Mostly, explanations seemed interpretable to the human eye. However, we observed that for several images, the model’s explanations seemed to focus on spurious features, such as background or objects often seen together with the target.

For example, predictions for Goldfish, Tigershark, and American coot seem to rely on background, such as water and seaweed. In predicting a terrier’s explanations, the model relied on the background too, possibly expecting that terriers are usually captured in a home environment. 2

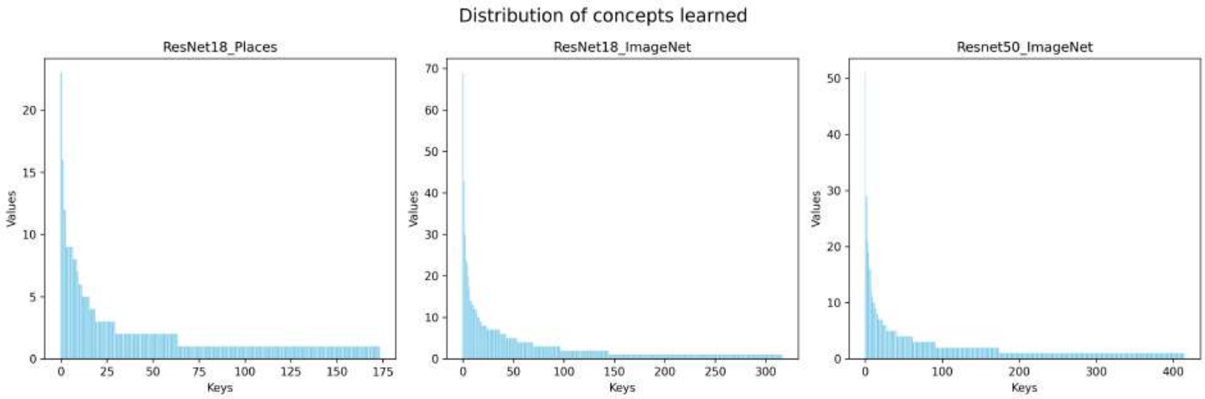

Figure 1: Distribution of unique concepts learned

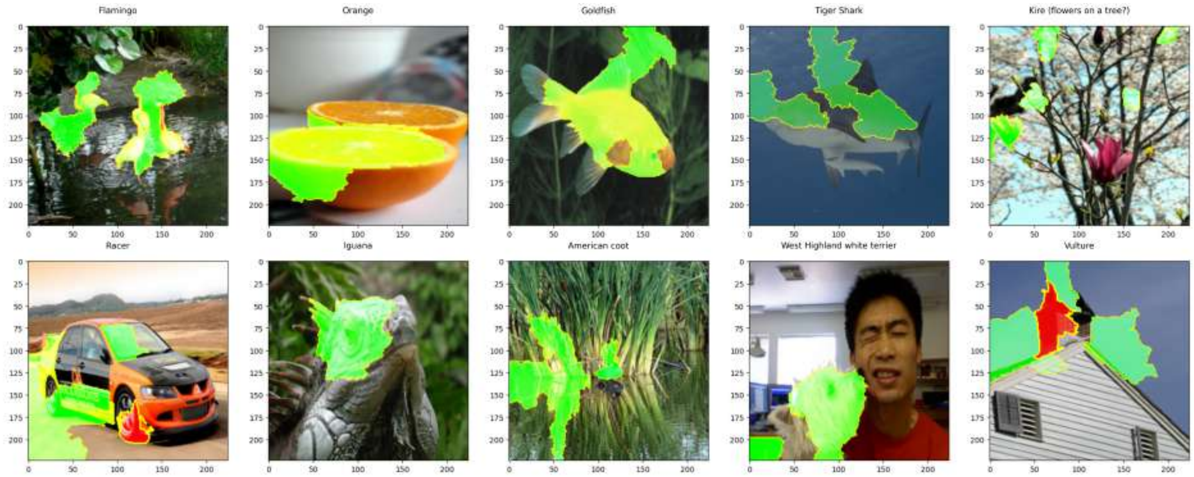

Figure 2: Explanations produced by LIME

On the other hand, the ”Vulture” (right lower corner in Fig2) in the image was captured on a house, which is arguably rather unusual or maybe rarely met in the dataset. Therefore, the red spot on the image covering the chimney of the house signifies an area that was most adversarial for the prediction. Nonetheless, the model predicted the ”Vulture” class correctly.

We also observed a funny error. While all class predictions were generally correct, there was one image of a tree with flowers (for some reason named ”Kite” in the ImageNet, upper right corner in Fig 2. On that image, the top-5 predictions were:

1. ”Proboscis monkey” (confidence: 0.1497)
2. ”Yellow lady’s slipper” (0.1259)
3. ”Pomegranate” (0.0979)
4. ”Macaw” (0.0666)
5. ”Orangutan” (0.0655)

We find images corresponding to the model’s prediction on the internet to understand why such unusual predictions were made and plot it on Fig 2. Indeed, there is something similar about the proboscis monkey’s nose and the flowers (if we disregard colors and context). However, this is not convincing if we want to claim the explainability of the model.

3 GradCAM

We run analysis on three different variations of class activation map methods. ScoreCAM has been said to produce more precise and less noisy explanations than GradCAM, and it doesn’t rely on gradients.

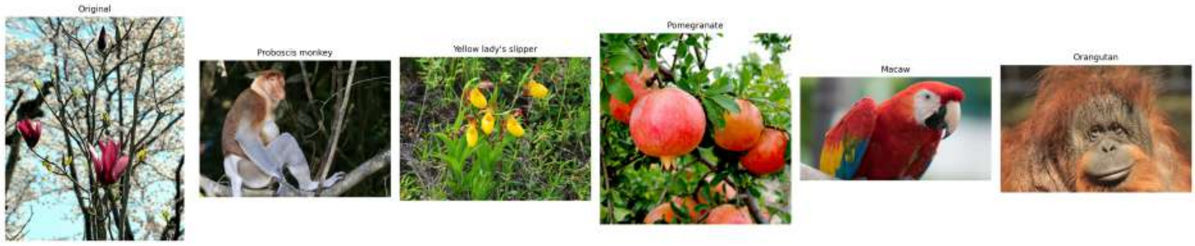

Figure 3: Original image Vs. Predicted classes

Similar benefits are promised by AblationCAM. However, the latter two took us much longer to compute, while producing no significant improvement over GradCAM results at least on the 10 probing images.

One exclusion is an image of an orange (Fig 5, last row), where ScoreCAM is able to locate the object more precisely on a saliency map. However, in the majority of other examples, the difference is marginal. Interestingly enough, all three methods fail to detect vultures and show the roof of the house as a salient feature (Fig 4, row 4).

4 Comparison: LIME Vs. GradCAM

While LIME focuses on superpixels and randomly samples them to obtain explanations, this might lead to predictions being suboptimal and changing from run to run. What we observe is that often regions selected as significant for model's prediction look somewhat abrupt, not coherent, and even spurious. In this regard, all three CAM-related methods seem to outperform LIME by producing more precise and smooth explanations.

For example, in predicting goldfish or tiger shark, CAM-methods focused on the fish itself, while LIME assigned a high weight to the background water (Fig 4). The same background inclusion problem that was present in LIME's explanations was less notable in CAM methods predictions for other objects as well, including orange, racer, and terrier (Fig 5). In Fig 5, however, we observed that a prediction on American coot (row 2) was reliant on background grass and water for all methods.

Overall, CAM-methods seemed to be more convincing in their explanations, but the difference between them wasn't immediately clear based on a limited sample we tested the methods on.

A CLIP-Dissect: Count of concepts in top-50 neurons with the highest similarity in FC layer

ResNet18-Places	ResNet18-Imagenet	ResNet50-Imagenet
wildlife: 13	**insect: 17**	terrier: 6
bedroom: 7	terrier: 9	**insect: 6**
food: 6	dog: 6	snorkeling: 5
car: 5	spider: 5	spider: 4
ferries: 3	pelican: 3	cobra: 3
dog: 2	ibis: 2	lizard: 3
nature: 2	basename: 2	dog: 3
insect: 1	bird: 1	monkeys: 3
insects: 1	birding: 1	basename: 2
vehicle: 1	lizard: 1	dresser: 2
juvenile: 1	insects: 1	pitcher: 2
foods: 1	wildlife: 1	bedroom: 2
deg: 1	mating: 1	bookshop: 1
bookshelf: 1		pelican: 1
vehicles: 1		snake: 1
desert: 1		cup: 1
sailing: 1		bookshelf: 1
vendors: 1		desk: 1
beverage: 1		tractors: 1
		birding: 1
		crab: 1

B Comparison of LIME and CAM variations

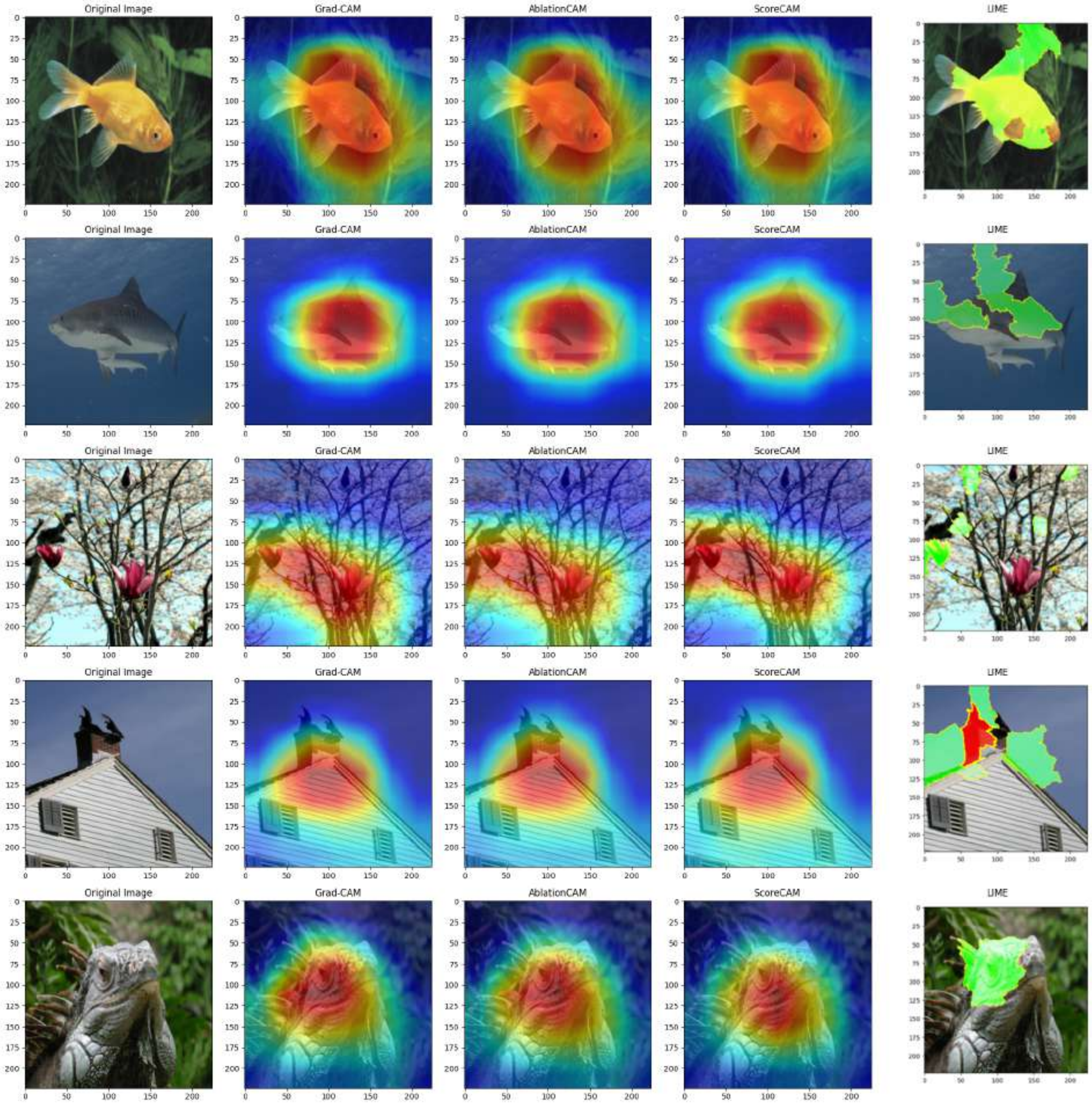

Figure 4: Comparison of explanations produced by GradCAM, AblationCAM, ScoreCAM, and LIME. Part 1

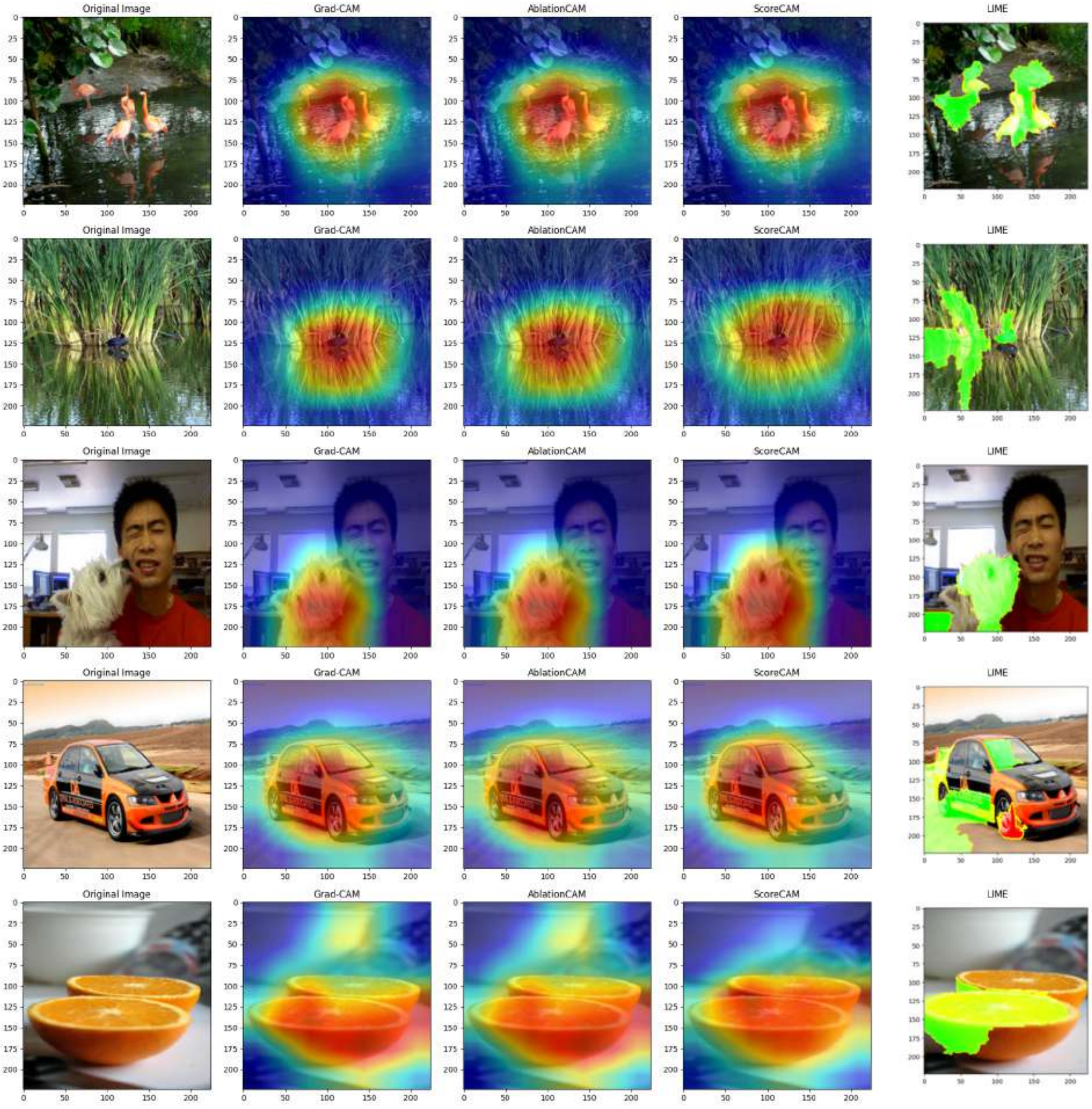

Figure 5: Comparison of explanations produced by GradCAM, AblationCAM, ScoreCAM, and LIME. Part 2