Deep Learning Lab-5

immediate

October 22, 2024

Abstract

This Lab focuses to implement two Whitebox explainer models GRADCAM and FEM. Whitebox explainer models explore the internal mechanisms of neural networks, leveraging the available knowledge of the model's architecture to provide a clearer understanding of its predictions. By analyzing how the network processes information, these models offer valuable insights into the underlying decision-making logic. This is particularly important in applications where interpretability is critical, as it allows for more transparent explanations of how and why specific predictions are made.

1 Explanation Models

Explanation models clarify the decision-making process of complex models, making their predictions more understandable. They provide insights in a way humans can easily interpret, helping to make blackbox models more transparent.

1.1 Whitebox models

They assist users in understanding how and why a model makes specific predictions, which is especially important in applications where interpretability is critical

1.1.1 GRADCAM (Gradient Class Activation Mapping)

Grad-CAM is a post-hoc visualization technique used to explain the class-discriminative regions of an image that contribute to a modelâs prediction. It leverages the structure of Convolutional Neural Networks (CNNs) to create a heat map highlighting the important pixels that influence the prediction for a particular class. Grad-CAM works by analyzing the feature maps from the last convolutional layer, as deeper layers in CNNs capture high-level features. It computes the gradient of the class score with respect to these feature maps, which helps identify the regions of the image most relevant to the prediction.

1.1.2 FEM (Feature-based Explanation Method)

FEM is similar to Grad-CAM in that it also utilizes the high-level features extracted by deeper convolutional layers in a CNN. However, FEM focuses on identifying strong and rare features that influence the final prediction. It assumes that these features follow a Gaussian distribution and employs K-sigma filtering to detect them from the last convolutional layer. By isolating these influential features, FEM provides a clearer understanding of how the model arrives at its decisions.

2 Dataset

The dataset consists of 102 images, equally split between two classes: African Elephants and Black Bears, captured in various environments.

3 Project Structure

The project is divided into three main files: `main.py`, `utils.py`, and `representations.py, FEM.py and GRADCAM.py`.

The `main.py` file is the core script that runs the pipeline. It loads the image, processes it through either GRAD-CAM or FEM methods, and displays the resulting saliency maps. The code implements two key methods for generating saliency maps: **GRAD-CAM** and **FEM (Feature Map Explanation)**. These methods highlight the regions of an image that a deep learning model (such as ResNet or Xception) focuses on when making predictions.

`utils.py` contains helper functions like `normalize_matrix`, which normalizes data, and `load_model_and_last_layer`, which loads pre-trained models (ResNet or Xception) for analysis. It also includes `display_grid` to present the visual outputs in a grid layout.

`representations.py` focuses on visual representation, providing functions like `create_heatmap` to convert saliency maps into heatmaps, and `overlay_heatmap_on_image` to blend heatmaps with the original image for easy interpretation of model attention.

4 GRADCCAM

4.1 Comparison of GRADCAM with Xception and RESNET models on an African Elephant and Black Bear images.

The ResNet saliency map for the elephant shows a wider spread around the body, focusing on the central parts and extending slightly to the head and tusks, indicating ResNet captures a broader range of features. In contrast, Xception has a more concentrated focus around the head and tusks, with deeper saliency, emphasizing specific features for classification.

For the black bear, ResNet highlights the central region, especially the head and shoulders, while Xception focuses more tightly on the head, indicating it prioritizes specific areas with higher impact for classification.

give me

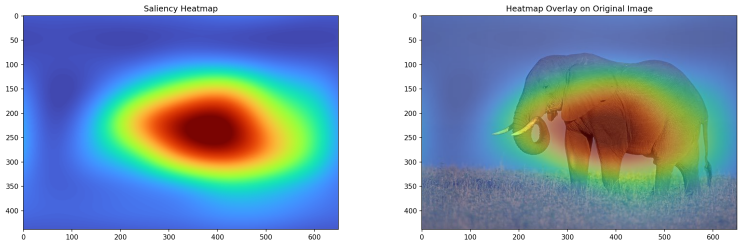

Fig. 1: GradCam with RESNET.

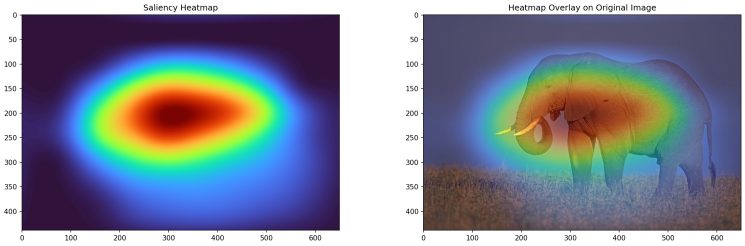

Fig. 2: GradCam with Xception.

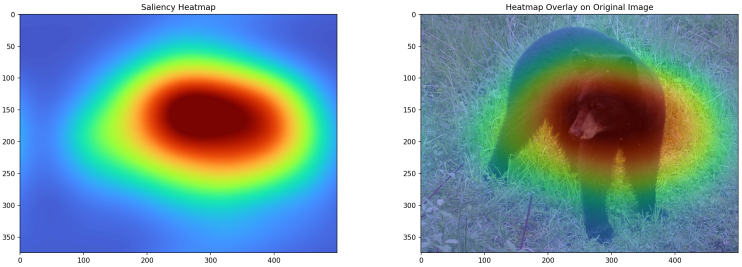

Fig. 3: GradCam with RESNET.

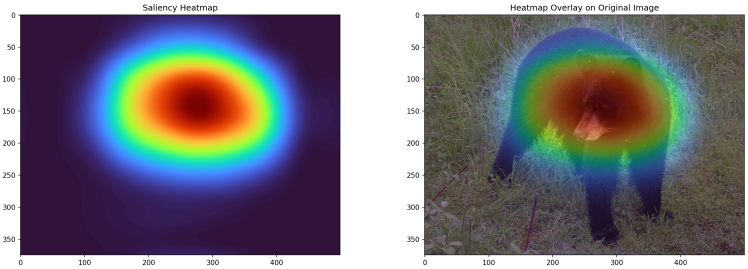

Fig. 4: GradCam with Xception.

5 FEM

For the elephant images, the **FEM with ResNet** saliency map shows a broader focus, highlighting a larger region around the body, including the head and trunk. **FEM with Xception**, however, concentrates more around the head and tusks, focusing on specific key features.

In the black bear images, **FEM with ResNet** highlights a more central region, primarily focusing on the bearâs head. In contrast, **FEM with Xception** has a more concentrated saliency map, tightly focusing on the bear's head and upper body, indicating it captures finer details for classification.

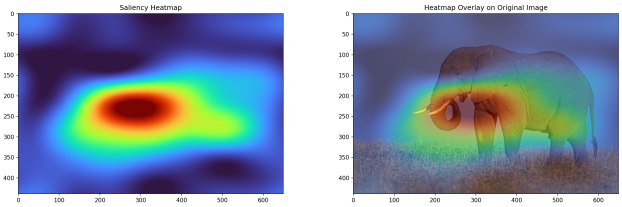

Fig. 5: FEM with RESNET.

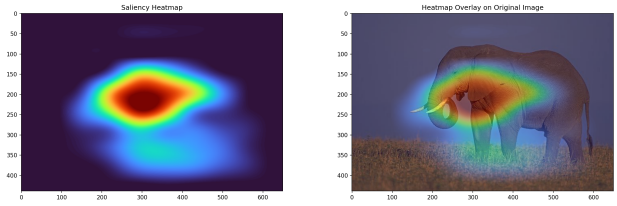

Fig. 6: FEM with Xception.

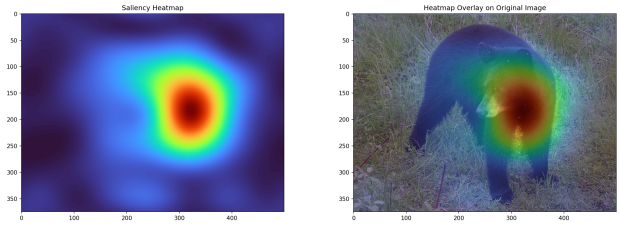

Fig. 7: FEM with RESNET.

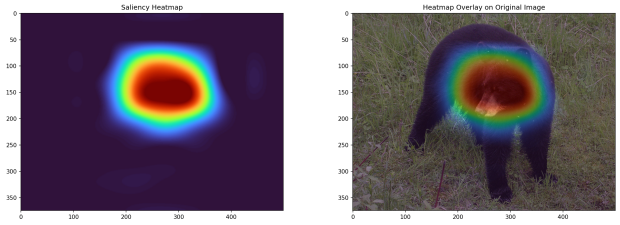

Fig. 8: FEM with Xception.

6 Conclusion

From the two explainer methods implemented (GRADCAM and FEM), the explanation results appear to depend more on the explainer method than on the classifier model (ResNet or Xception). Both methods highlight different regions of the object, with GRADCAM typically focusing on broader areas and FEM targeting more specific regions. Despite this, the classification outcomes are similar across both models, suggesting that the choice of explainer plays a greater role in determining the focus of the explanation, rather than the architecture of the classifier. This is in contrast to previous labs, where factors such as model generalization, the number of samples, and explainer-specific parameters had a stronger influence on explanation quality.

7 References

- Ayyar, M. P., Benois-Pineau, J., & Zemmari, A. (2021). White Box Methods for Explanations of Convolutional Neural Networks in Image Classification Tasks. *arXiv preprint arXiv:2104.02548*.
- Fuad, K. A. A., Martin, P. E., Giot, R., Bourqui, R., Benois-Pineau, J., & Zemmari, A. (2020, November). Features understanding in 3D CNNs for actions recognition in video. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1-6). IEEE.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618-626).