Explaining Bird Species Prediction Using Saliency Maps and Concept Bottleneck Model

Albert Kjøller Jacobsen, Carolina Lopez Olmos, Rasmus J. Pedersen, Naja Jean Larsen, Nina Weng

DTU Compute, Technical University of Denmark

Introduction

Explaining the predictions of AI systems is relevant and an important field of interest. In this project, we implemented **two explainability methods**, namely **Saliency Map** and **Concept Bottleneck Models (CBM)**. These are implemented using two modeling frameworks: a transfer learning **ResNet50** architecture and a **CBM,** respectively. The focus is on bird species prediction using the **CUB-200-2011 dataset** which contains 11,788 images of 200 subcategories belonging to birds. To conclude, a quantitative approach has been designed to validate the XAI methods' robustness and explainability capabilities. The code of this project can be found here:
https://github.com/albertkjoller/XAI-ResponsibleAI

Explainability Methods

Choice of Saliency Map

The post-hoc method **GradCAM** [1] [3] was selected for obtaining saliency maps from the ResNet50-based model. GradCAM was selected for emphasizing salient regions while being computationally efficient. Compared to e.g. the CAM-approach, GradCAM does this by weighting the CAM with the gradients of the selected class wrt. the final convolutional layer. We emphasize that gradient-based methods might not always locate the import salient regions of the image, e.g. in case of noise or adversarial attacks.

Concept Bottleneck Model (CBM)

Concept Bottleneck Models [2] obtains explanations by using an architecture with a bottleneck layer designed to learn concepts occurring in the input data. As such, CBMs provide *intrinsic explanations.* In this project, the *independent* version of the CBM was trained in order to avoid data leakage.

Model performance

Figure 1) presents the accuracy grouped by bird species for the two trained architectures. In terms of accuracy, the model is sensitive to the class in its' prediction.

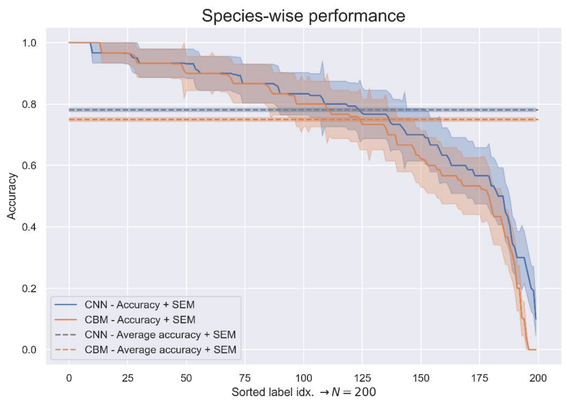

Figure 1. Species-wise performance.

Visualizations of Explanations

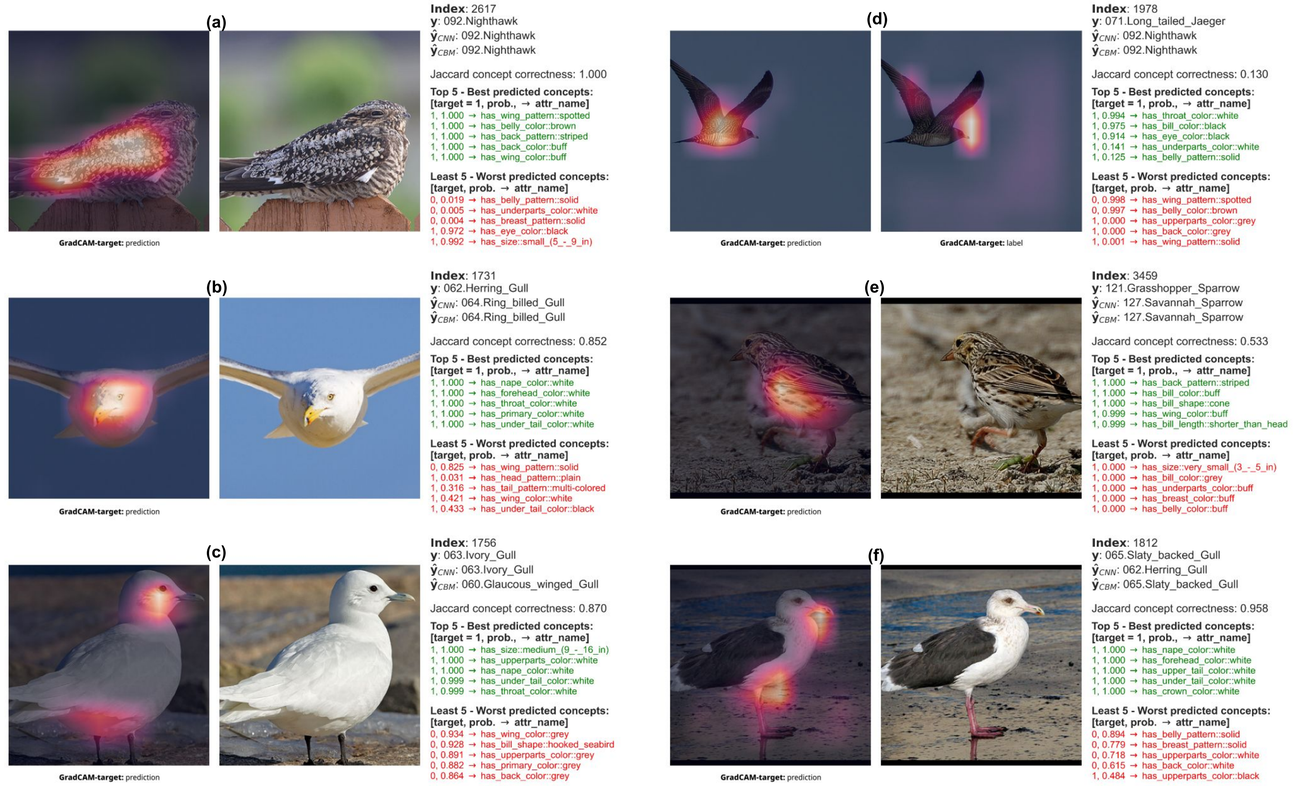

Figure 2. Examples of explanations from saliency map and concept bottleneck models

Quantitative Validation Strategy of Explanations - Robustness

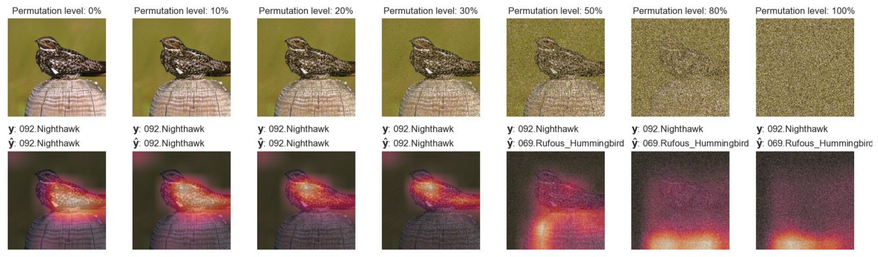

Figure 3. The illustration of pixel flipping robustness with different perturbation level.

Figure 3) presents a suggested strategy for quantitatively assessing robustness of explanations. This project deals with robustness of explanations as *invariance* to *slight* perturbations of the input data.

With this definition, **pixel flipping** was applied for permuting a selected subset of pixels in the input image. As seen in the bottom row, this affects the saliency map for large proportions of pixel flips.

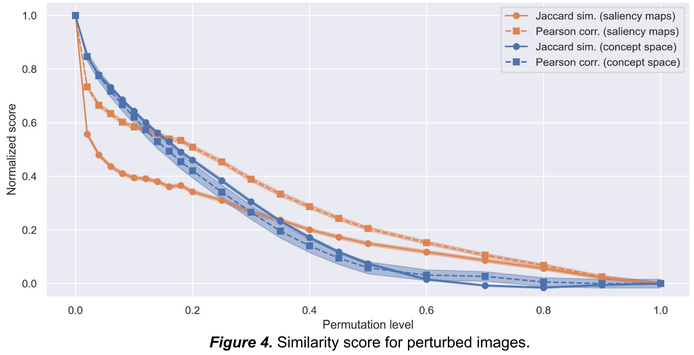

Figure 4. Similarity score for perturbed images.

For determining robustness of explanations, a distance metric between the original explanation and explanations obtained for modified input images is required. In Figure 4) multiple robustness scores are reported for a variety of pixel flipping ratios.

It could be seen from Figure 4 that:

- **Saliency maps are less robust** than concept-based explanations considering all metrics for slight perturbation. E.g. **2%** pixel flipping → Normalized SSIM score drops to **0.55**
- **Robustness scores** depend on the metric and dimensionality of the inputs.
  E.g. saliency maps $\in [0, 1]^{224 \times 224}$ and concepts $\in [0, 1]^{112}$

Discussion

Saliency maps can be very intuitive, e.g. in Figure 2) for the Nighthawk the wing pattern is highlighted **(a)**, similarly when the Long Tailed Jaeger is misclassified as a Nighthawk **(d)**, it also highlights the pattern on the bird making it very intuitive why this misclassification occurred. However, it is less intuitive what has to change for it to predict correctly, when looking at the saliency map for the true label. **Concepts** on the other hand are more informative allowing a more detailed understanding of the prediction decision. Through them the problems in reasoning can be explained, e.g. it can be seen CBM suffers from the influence of lighting conditions since in **(b)** the head pattern is explained as *non-plain* or in **(c)** the wing is said to be *grey*. Similarly, in **(e)** the image focus on the bird influences the model thinking the bird has a *large size.* The possibility of concept intervention, while not studied in this work, can be of great usefulness to handle these misperceptions.

Limitations

Limitation of saliency maps - as GradCAM uses a target class for computing the gradient, the saliency map is dependent on the confidence level of the predictions. Saliency maps *might* highlight salient regions but domain knowledge is often needed for understanding explanation.
Limitation of CBM - CBM learns which concepts typically connect to which species and might suffer from correlated concepts as well as misalignment between predicted and true concepts (due to the independent architecture).
Limitation of the quantitative strategy - comparison of similarity scores across modalities is difficult, because of differences in metrics and dimensions. As such, the conclusion is highly dependent on the choice of score. One important aspect of the strategy is to consider wrt. *what* the model is robust. Perturbations of the input data can take many forms - e.g. using rotations might reveal different findings while being more suitable for real-life applications.

An example is provided in Figure 5) showing robust predictions, however without maintaining a robust saliency map after the different rotations. This shows how robustness in predictive performance is not equivalent to robustness in explainability.

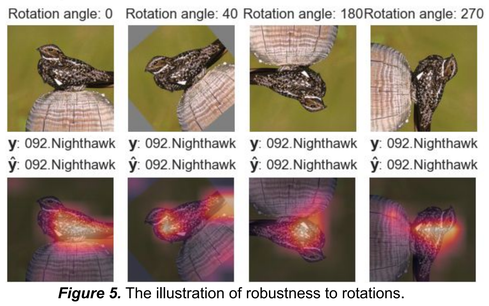

Figure 5. The illustration of robustness to rotations.

References

[1] Ramprasaath R. Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. 2017
[2] Pang Wei Koh et al. Concept Bottleneck Models. 2020
[3] https://github.com/jacobgil/pytorch-grad-cam