Towards a Deeper Understanding of Concept Bottleneck Models Through End-to-End Explanation

Jack Furby[1] **Daniel Cunnington**[2] **Dave Braines**[2] **Alun Preece**[1]

[1]Cardiff University, UK [2]IBM Research Europe

Introduction

In our paper, we evaluate **Concept Bottleneck Models (CBMs) [1]**, a type of Deep Neural Network that first maps raw input(s) to a vector of human-defined concepts, before using this vector to predict a final classification. We analyse what input features the model uses for **concept predictions** and which predicted **concepts contribute to the final classification**.

Contributions

1.Provide an understanding of the input features the models use from both the input image and predicted concepts.

2.Introduce a quantitative evaluation to measure the distance between the maximum input feature relevancy and the ground truth location and perform this with multiple relevancy techniques.

3.We propose using the proportion of relevance as a measurement for explaining concept importance.

Method

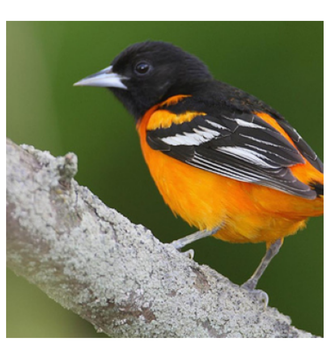
Our models were trained on a modified version of the CUB-200 2011 (CUB) dataset [1] which used **class-level concepts**; concepts are applied to classes.

We use **Layer-wise Relevance Propagation (LRP) [2]**, an explanation technique which highlights features of an input which were relevant to a prediction We generate **saliency maps** of both the relevance on the input image and concept prediction vector.

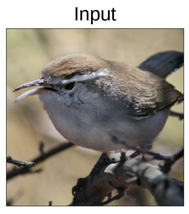
We compare concept saliency maps using by averaging the distance between the most salient points and the ground truth locations.

Concept contribution proportion calculations are made possible by LRP's requirement to conserve relevancy as it's propagated backwards through a model.

Code and additional examples available at: github.com/JackFurby/explainable-concept-bottleneck-models



Input



Input

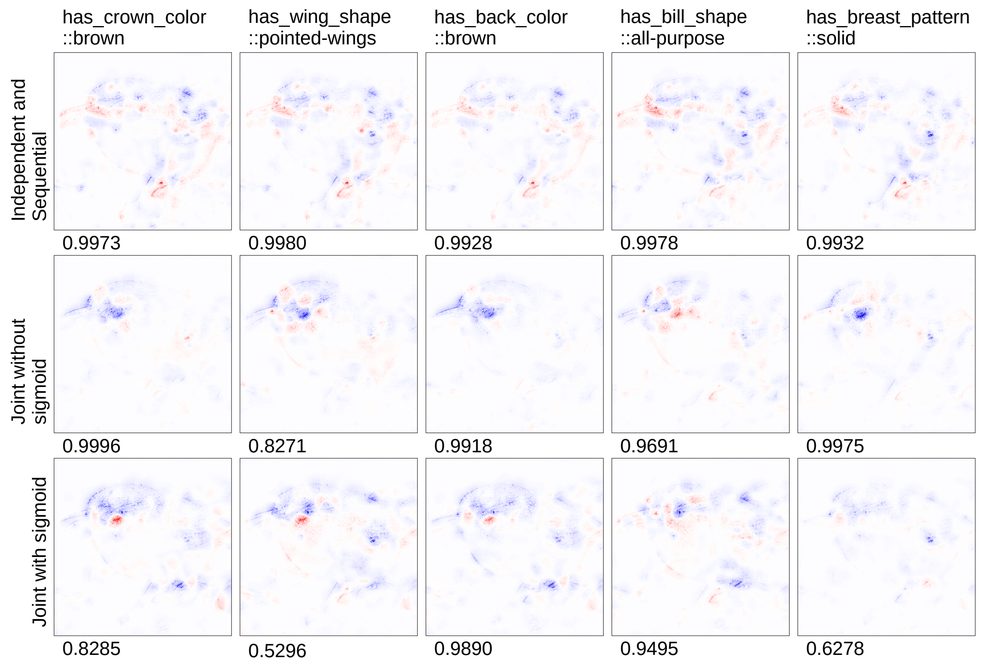


Figure 1: In general, relevance does not map to input features that a human would associate each concept with.

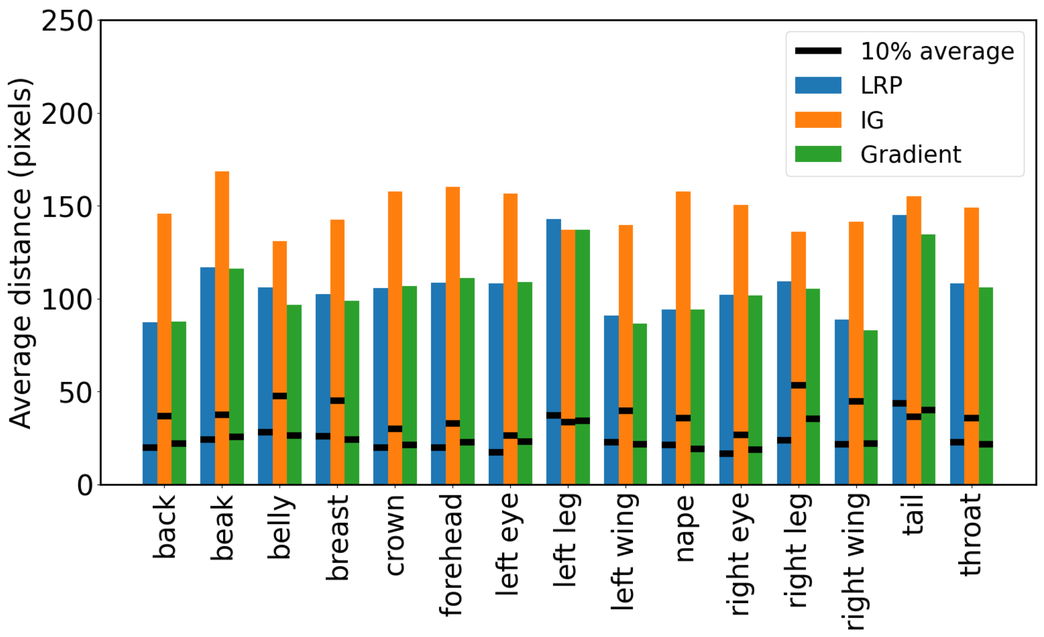


Figure 2: The average distance of all explanation techniques are too far from the ground truth locations for the model to be focusing on the correct input features.

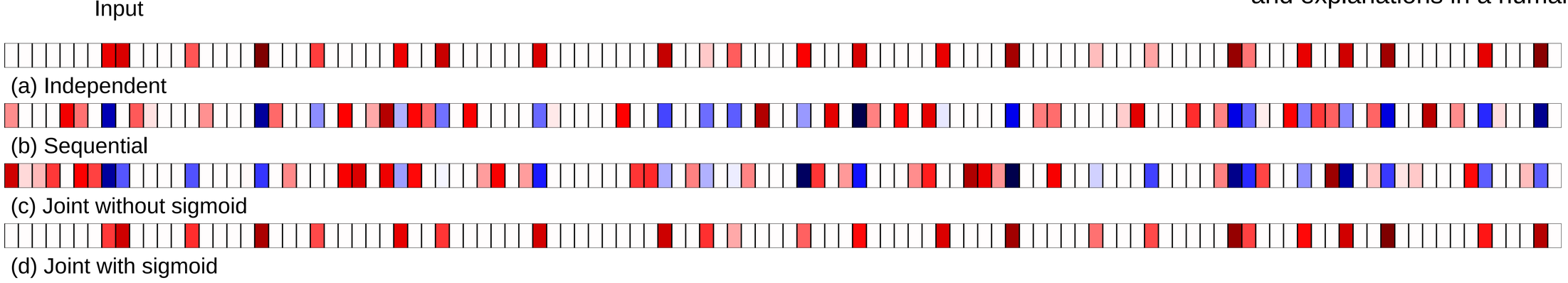


Figure 3: The concept to final classification model part primarily uses concept predicted as present although, for the sequential model and joint without sigmoid model, concepts predicted as present have negative relevance.

Results

Relevancy is generally distributed over the entire bird and the same input features can predict different concepts, as shown in figure 1.

We believe the dataset is not confining the model to learn concepts as intended.

Figure 2 shows the average distance of saliency to the ground truth location. The distance is around 100 pixels which reinforces **the model is not focusing on concepts** as 100 pixels could easily be outside of a concept. The input images are 299x299 pixels.

In figure 3 we produce saliency maps to show which concepts the model uses for the final classification. **Our CBMs mostly predict final classifications using concepts predicted as present.**

By calculating relevancy distribution we reveal the concepts used for final classification predictions.

Conclusion and Future Work

Our paper evaluates CBMs using the LRP explanation technique. These reveal that concepts do not map to distinct regions in the input space. However, relevance from the final classification back to the concept vector shows the model has mapped these as expected for some CBM training methods.

We demonstrate the ability to calculate proportional concept contribution to final classifications.

Future work

- CBMs trained on instance-level concepts and non 1 to 1 mapping of concepts to final classifications.
- Evaluating CBMs with a non-relevance-based method.
- Studying the effectiveness of CBMs and explanations in a human study.

References

[1] Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept Bottleneck Models. In III, H. D.; and Singh, A., eds., Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, 5338–5348. PMLR.

[2] Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE, 10(7): 1–46.

Acknowledgement

This research is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) and IBM UK via an Industrial CASE (ICASE) award.