Integrating feature attribution methods into the loss function of deep learning classifiers

James Callanan

Abstract

Feature attribution methods are typically used post-training to judge if a deep learning classifier is using meaningful concepts in an input image when making predictions. In this study, we propose using feature attribution methods to give a classifier automated feedback throughout the training process via a novel loss function. We call such a loss function, a heatmap loss function. Heatmap loss functions enable us to incentivize a model to rely on relevant sections of the input image when making classifications.

Two groups of models were trained, one group with a heatmap loss function and the other using categorical cross entropy (CCE). Models trained with the heatmap loss function were capable of achieving equivalent classification accuracies on a test dataset of synthesised cardiac MRIs. Moreover, HiResCAM heatmaps suggest that these models relied to a greater extent on regions of the input image within the heart.

A further experiment demonstrated how heatmap loss functions can be used to prevent deep learning classifiers from using non-causal concepts that disproportionately co-occur with certain classes when making classifications. We believe the heatmap loss function could be used to train more skillful classifiers and to prevent them from learning dataset biases by directing where a model should be looking when making classifications.

Keywords: Loss function, Dataset bias, Grad-CAM, HiResCAM, Deep learning

1 Introduction

Many feature attribution methods are differentiable with respect to the network's weights and biases. This makes them well suited to be integrated into a model's loss function. As part of this study, models were successfully trained by integrating Grad-CAM (Selvaraju et al., 2017) and HiResCAM (Draelos and Carin, 2020) into their loss functions.

The heatmap loss function used in the experiments below consisted of a weighted sum of a HiResCAM component and a mean squared error (MSE) component. The HiResCAM component serves to disincentivize the classifier from relying on irrelevant portions of images when making classifications and the MSE component acts to incentivize the model to make correct class classifications.

The models in this study were trained on a dataset of programmatically generated cardiac MRI cross sections to classify cardiac disease. Thus, the areas of the images outside of the heart were deemed irrelevant for making classifications. Consequently, the HiResCAM component of the loss function was set equal to the sum of the HiResCAM heatmap values that lay outside of the heart. Many other metrics have been proposed to evaluate the degree of overlap between feature attribution maps and segmentation masks in segmentation problems such as Dice (1945). There is potential for these to be adapted for use in a heatmap loss function.

2 Methods and Results

Balanced datasets containing 4 classes of cardiac MRIs were generated. The classes were; normal, hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM) and arrhythmogenic right ventricular cardiomyopathy

(ARV). Attempts were made to make the synthetic datasets representative of a real world cardiac dataset by injecting noise and taking into account disease biomarkers, aetiology and sex prevalence.

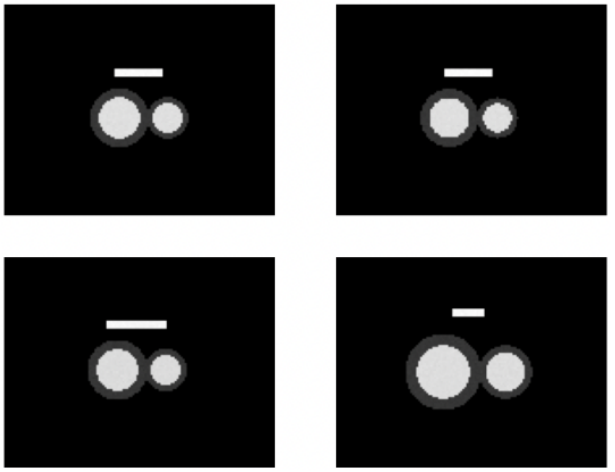

Figure 1: Exp 1: Normal MRIs

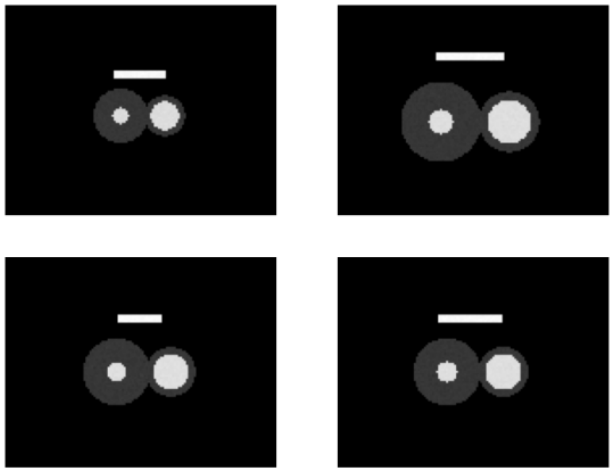

Figure 2: Exp 1: HCM MRIs

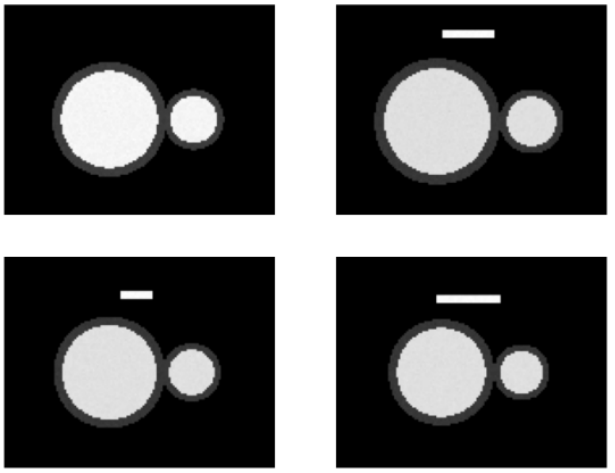

Figure 3: Exp 1: DCM MRIs

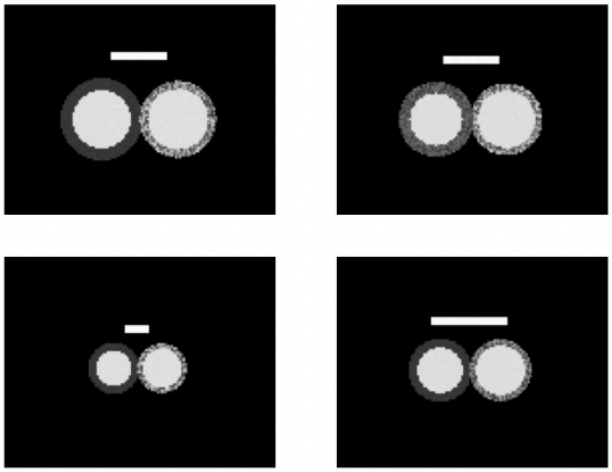

Figure 4: Exp 1: ARV MRIs

Two groups of models were trained, one with the heatmap loss function and one with CCE. Only models with a validation accuracy of greater than 95% were selected. All models had identical architectures. However, the learning rates varied from model to model. The model training process was automated using Keras-Tuner (O'Malley et al., 2019) and learning rates were varied within a range that exhibited good convergence of training and validation loss and accuracies. HiResCAM-heart overlap metrics were computed on a test dataset of 300 MRIs for the 25 selected models with CCE loss and 23 with heatmap loss as can be seen in Figure 5. A Shapiro-Wilkes test confirmed the distributions of these metrics were not normally distributed among groups. Consequently, a two-sided Mann-Whitney U-test ($\alpha = 0.05$) was performed to test for a statistically significant difference between the group's overlap metrics. The models trained with the heatmap loss function were found to have systematically higher degrees of heatmap-heart overlap, with a p-value $\approx 1 \times 10^{-9}$.

A second experiment was carried out to test whether models were relying on knowledge of the patient's sex when making classifications. It was theorised that a model may base predictions based off a patient's sex when classifying diseases such as arrhythmogenic right ventricular cardiomyopathy. This disease occurs a disproportionate amount in males in our training dataset and in real life. For this experiment, all systematic differences

between the MRIs of males and females were removed (i.e. size and body fat's sex dependence). However, a label was included in the bottom corner of male patient's MRIs. This enabled us to separate the concept of sex from the heart. Thus, we could test for a model's reliance on sex by calculating the degree of overlap between the HiResCAM heatmap and the sex label. Two groups of models were trained of which 23 models with CCE loss and 22 with heatmap loss were selected. HiResCAM-heart overlap as well as HiResCAM sex label overlap metrics were computed on a test dataset of 300 MRIs. Statistically significant differences were found in the distributions of the heatmap-heart and heatmap-sex label overlaps among both groups. The models trained with the heatmap loss function had higher degrees of heatmap-heart overlap (p-value $\approx 1 \times 10^{-8}$) as can be seen in in Figure 6, they also had lower degrees of heatmap-sex label overlap (p-value $\approx 1 \times 10^{-7}$).

Perfect classification accuracy was achieved by models in groups across all experiments. Figure 7 and Figure 8 show classifications with good and poor HiResCAM-heart overlaps.

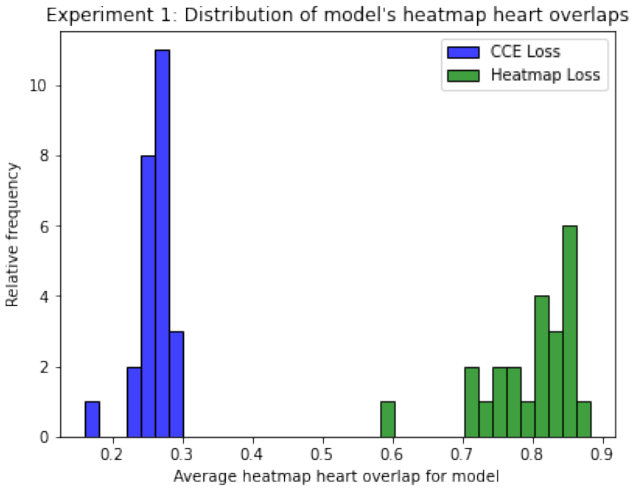

Figure 5: Exp 1: Distribution of overlaps

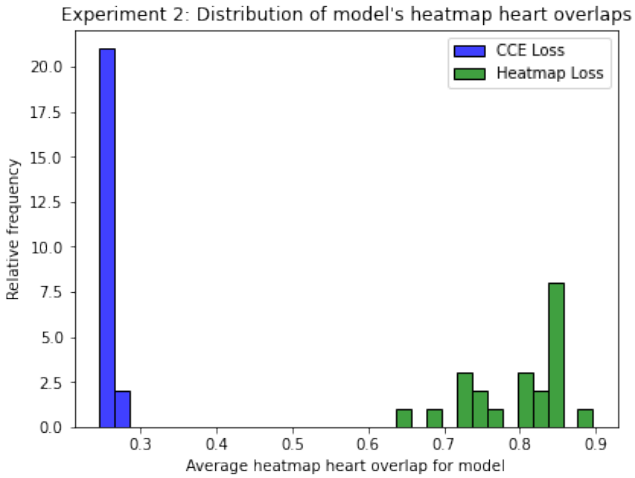

Figure 6: Exp 2: Distribution of overlaps

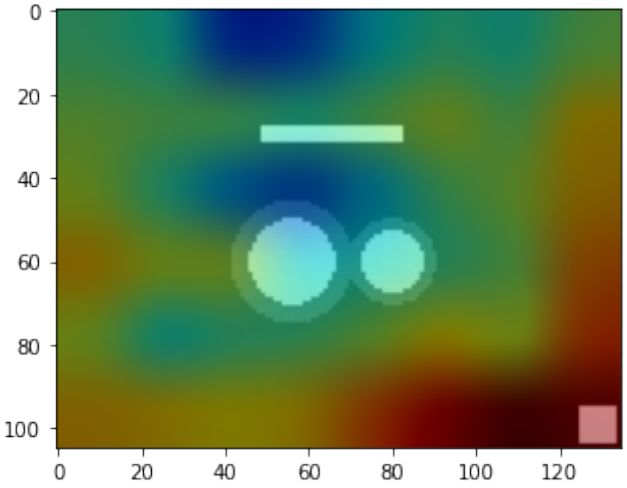

Figure 7: Low heatmap-heart overlap

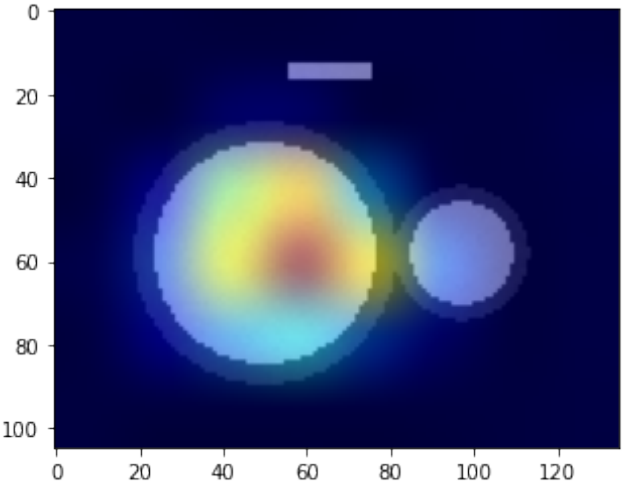

Figure 8: High heatmap-heart overlap

3 Discussion and Conclusion

We have demonstrated that a model trained with a heatmap loss function can achieve high classification accuracies. Further research needs to be carried out to test the feasibility of heatmap losses for harder classification

problems. Several obstacles need to be considered such as; the intrinsic limitations of the attribution methods used, the requirements of regions to be separable from the object being classified and the increased training times. The code associated with this project can be found here. A more in depth discussion of this project as well as the applications and future directions of heatmap loss functions can be found here.

References

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Draelos, R. L. and Carin, L. (2020). Hirescam: Faithful location representation in visual attention for explainable 3d medical image classification. *arXiv preprint arXiv:2011.08891*.

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al. (2019). Kerastuner. `https://github.com/keras-team/keras-tuner`.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.