

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

James Callanan¹, Carles Garcia-Cabrera², Niamh Belton¹, Gennady Roshchupkin³,
Kathleen M Curran¹

¹ *University College Dublin*, ² *Dublin City University*, ³ *Erasmus University Rotterdam*

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 classifications. 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 MRI slices. Moreover, HiResCAM heatmaps suggest that these models relied to a greater extent on regions of the MRI slices 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 images of a certain class when making classifications. This suggests that heatmap loss functions could be used to prevent models from learning dataset biases by directing where the 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 it possible to integrate them into a model's loss function. Models were successfully trained with both Grad-CAM (Selvaraju et al., 2017) and HiResCAM (Draelos and Carin, 2020) integrated into their loss functions. However, 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 served to disincentivize the classifier from relying on irrelevant portions of images when making classifications and the MSE component acted to incentivize the model to make correct class classifications. A training, validation and testing dataset of synthetic cardiac MRI slices were generated along with their corresponding segmentation masks. The areas of the MRI slices outside of the heart were deemed irrelevant for cardiac disease 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.


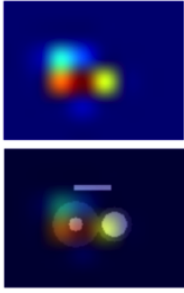
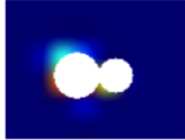


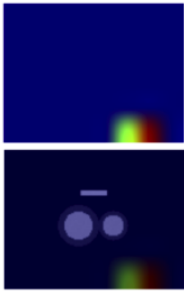


Cardiac MRI cross sections	HiResCAM heatmaps	Heatmap regions outside of heart	Heatmap regions inside of heart
			
			

Figure 1: **Visual explanation of the heatmap loss component:** The heatmap loss component is calculated by summing the portion of the HiResCAM heatmap that lies outside of the heart (shown in column 3 above). This definition is imperfect as both MRI slices above incur equivalent heatmap losses despite the classifier relying on more information within the heart in the top MRI slice. Thus, there is likely scope to define a better heatmap loss function.

2 Methods and Results

Balanced datasets of MRI slices through the center of the heart were generated. These slices were meant to mimic short axis cardiac MRI cross sections through the center of the heart. Four classes of cardiac MRIs were generated, these included; 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. Below are sixteen sample MRI slices taken from the training dataset used in the first experiment, these correspond to sixteen different ‘patients’.

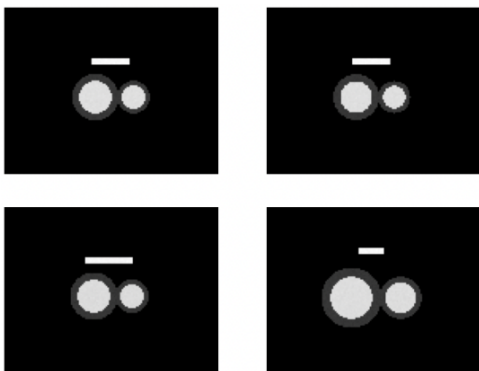


Figure 2: Exp 1: Normal MRI slices

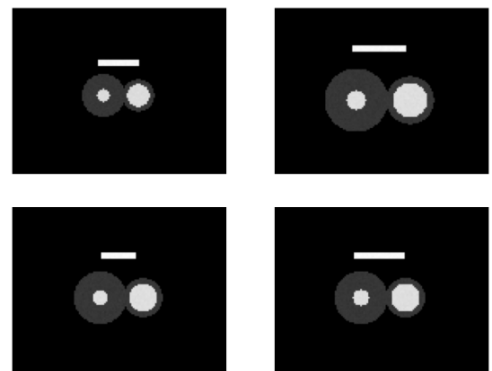


Figure 3: Exp 1: HCM MRI slices

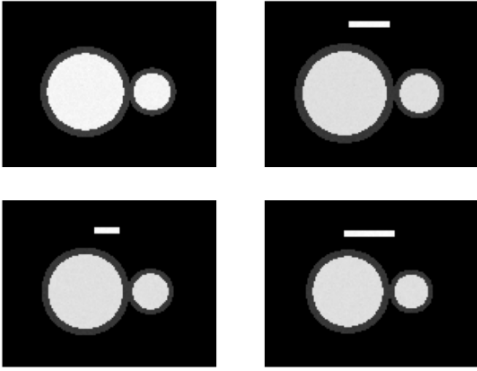


Figure 4: Exp 1: DCM MRI slices

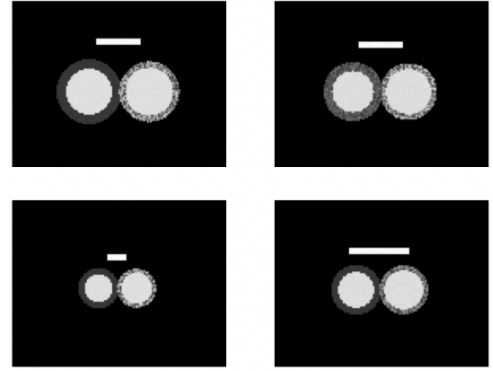


Figure 5: Exp 1: ARV MRI slices

In the first experiment, ~ 30 models were trained with a heatmap loss function and ~ 30 models with CCE. All models had identical VGG16 inspired architectures. Learning rates were varied using Keras-Tuner within a range that exhibited good convergence of training and validation loss and accuracies. Of the ~ 60 models trained, models with a classification accuracy of less than 95% on a validation dataset were discarded. This meant 23 models trained with a heatmap loss function and 25 models trained with CCE remained. HiResCAM-heart overlap metrics were computed on a test dataset comprising of 300 MRIs for models in both groups. The results of which can be seen in Figure 6. A Shapiro-Wilkes test confirmed the distributions of the overlap 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 would rely on knowledge of the patient's sex when making classifications. It was hypothesized that a model may base classifications off knowledge of a patient's sex because many cardiac diseases occur disproportionately among the sexes. For example, for every female case of ARV there are ~ 2.7 male cases. The datasets used in these experiments were designed to mirror real world differences in disease prevalence among the sexes. In 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 to distinguish them. This enabled us to separate the concept of sex from the heart. Thus, we could test for a model's reliance on sex when making classifications by calculating the degree of overlap between the HiResCAM heatmap and sex label. Approximately 50 models were trained with both loss functions, after discarding those with a validation accuracy $< 95\%$, 22 models trained with the heatmap loss function and 23 models trained with the CCE loss function remained. HiResCAM-heart overlaps as well as HiResCAM sex label overlaps 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 7. These models 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 on an unseen testing dataset.

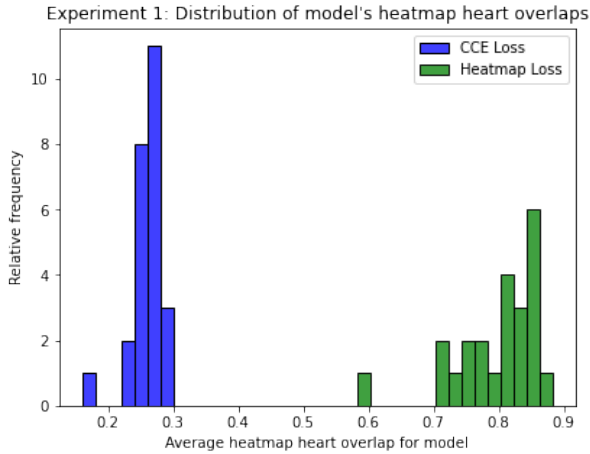


Figure 6: Exp 1: Distribution of overlaps

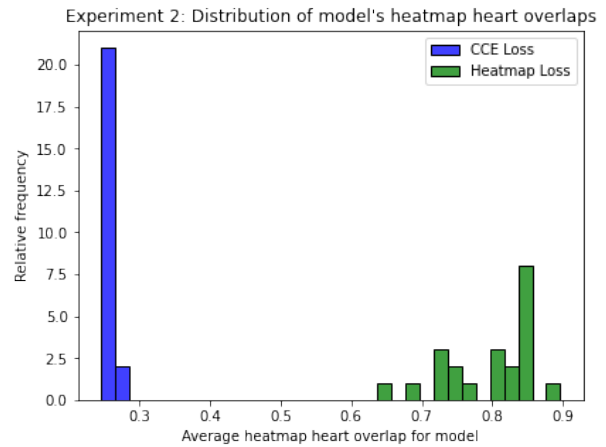


Figure 7: Exp 2 (sex label): Distribution of overlaps

3 Discussion and Conclusion

We have demonstrated that a model trained with a heatmap loss function can achieve high classification accuracies on a synthetic dataset. However, we envision this loss function being used to overcome an issue that affects existing high performing classifiers, the issue of learned bias (as outlined by Kim et al. (2018), Selvaraju et al. (2017) and Ghorbani et al. (2019)). Eliminating learned bias using techniques such as data augmentation, oversampling and undersampling is likely infeasible, if not impossible. The disproportional co-occurrence of non-causal concepts within images of a given class seems inevitable, especially when the set of possible concepts that a classifier can detect is extremely large. Thus, we believe heatmap loss functions warrant further investigation. Future research should test the feasibility of heatmap loss functions on a real world dataset. Ideally the chosen dataset would be large enough to train deep learning classifiers using conventional loss functions and would contain both classification labels and segmentation masks. Moreover, several obstacles need to be investigated further such as; the intrinsic limitations of the feature attribution methods used, the requirements of regions to be separable from the object being classified and the increased training times required when using a heatmap loss function. The code associated with this project along with a more in depth discussion can be found at <https://jamescallanan.github.io/HeatmapLossFunction>.

Acknowledgements

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6183.

References

- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.
- Draeos, R. L. and Carin, L. (2020). Hirescam: Faithful location representation in visual attention for explainable 3d medical image classification. *arXiv preprint arXiv:2011.08891*.
- Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. (2019). Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.
- 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.