Background and Problem Definition

Using various proposed methods of post-hoc interpretability we can test the relevance of one or more concepts in a classification task. These methods work by testing intermediate layers for their ability to separate classes based on some concept. These methods include Concept Activation Vectors (CAV) [1], Concept Gradient (CG) [2] which is a generalization of CAVs and Hierarchical Concept based explanations [3]. These methods allow for a human-friendly interpretation of the internal state of a deep learning model, so that questions about model decisions may be answered in terms of natural high-level concepts [1].

However, the main problem we propose to work on in this project, is in using these high-level concepts in order to train a model more efficiently. This problem has been explored using both loss based [5] and feature mapping approaches [4]. Where they use a learned feature map in order to more closely align the raw input with a conceptual input that highlights a specific concept. However, as this is a pre-print and we do not have access to their datasets, nor their code we are unable to implement a similar solution to that which they have implemented.

Instead we propose a similar solution to those which have been attempted by Ross et al. [5]. This centers on the concept of using a joint loss in order to train the model on a specific concept (or explanation). To do this we will modify the methods that interpret the relevance of a concept (CAV and CG if time permits) in a particular layer towards a classification task and use this to calculate a joint loss.

$$L_{joint} = L_{label} + \lambda L_{expl}$$

Where the additional loss is computed by taking the difference between the conceptual sensitivity $S_{C,k,l}(x)$ [1] and the target concept for the given class $c(x)$.

$$L_{expl} = \left| S_{C,k,l}(x) - c(x) \right|$$

If time permits, we will also look at other ways to learn from concepts as explained in [4].

Datasets

We propose to use the same datasets that were used in the Kim et al [1]. We will follow the steps outlined here for downloading these datasets. These datasets include the Broden dataset for image concepts: striped, zigzagged, and dotted. The original Broden dataset has 63,305 images with 1197 visual concepts. There are 6 categories of concepts: textures, colors, materials, parts, objects, and scenes. We will sample 120 images from each concept class.

We will also sample 4 different sets of 120 images from the Imagenet dataset. These images will be distinct from the concept and zebra images. The testing will then be performed on zebra images. This dataset contains over 14 million images, which have been hand annotated. These are both public datasets.

References

[1] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2673–2682. ICML, 2018.

[2] Andrew Bai, Chih-Kuan Yeh, Pradeep Ravikumaret, Neil Y. C. Lin, Cho-Jui Hsieh. Concept Gradient: Concept-based interpretation without linear assumption. https://doi.org/10.48550/arXiv.2208.14966

[3] Mohammad Nokhbeh Zaeem, Majid Komeili. Cause and Effect: Hierarchical Concept-based Explanation of Neural Networks. https://doi.org/10.48550/arXiv.2105.07033

[4] Unknown. Machine Learning from explanations. Submitted to ICLR 2023

[5] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: training differentiable models by constraining their explanations. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 2662–2670, 2017.