Interpretable Machine Learning

Tyler Martin (tam66@cam.ac.uk)

Supervisor: Adrian Weller

Interpretability

Interpretability is defined as *presenting a rationale behind an algorithm's decision in terms understandable to humans*. For a machine learning algorithm b, both local and global and global interpretability exist for dataset D.

$$\mathcal{D} = (\mathcal{X}, \mathcal{Y})$$

Local

$$b(x) = \hat{y}$$

- Understanding why a *particular* decision or prediction was made

Global

$$b : \mathcal{X} \rightarrow \hat{\mathcal{Y}}$$

- General understanding of what predictions will be made for any input

Two approaches to interpretability are 1) constraining models to be inherently interpretable and 2) applying post-hoc explanation methods.

Inherently Interpretable

- Constrained model form (Bayes decision list)
- Accuracy tradeoff

Post-hoc

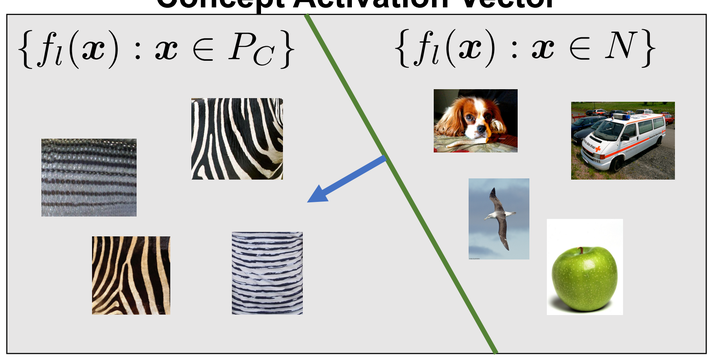
- Model agnostic
- Typically achieve local interpretability

TCAV

Testing with Concept Activation Vectors [1] is a post-hoc method with some degree of global interpretability. **Human-defined** concepts are used to create a set of images P_C that contain concept C. Another set of images N that do not contain concept C are generated randomly from ImageNet.

The activations $f_l(x)$ are found for layer l for each example image and a linear decision boundary is fit. The CAV is normal to the decision boundary hyperplane.

Concept Activation Vector



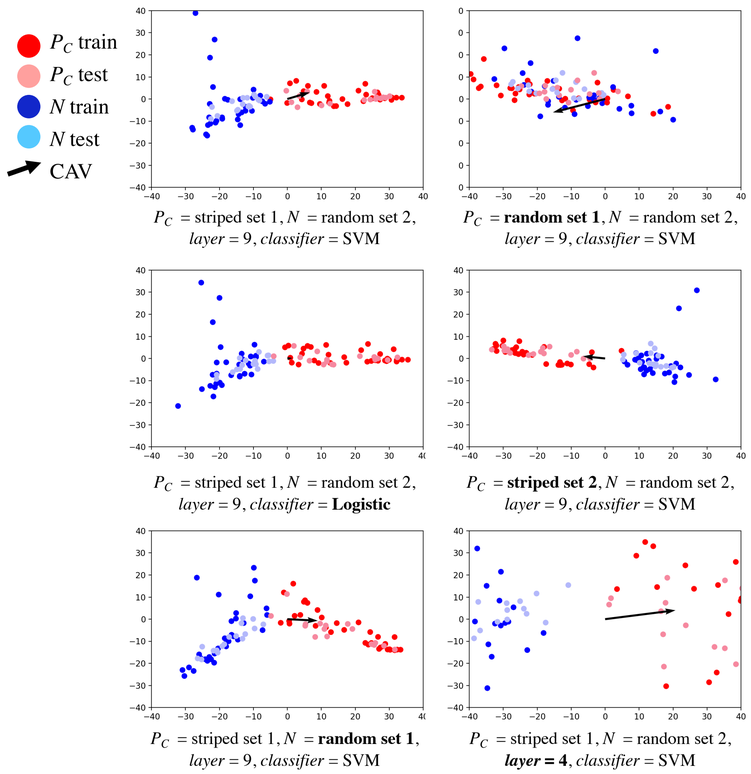
CAV Visualization

Questions about CAVs

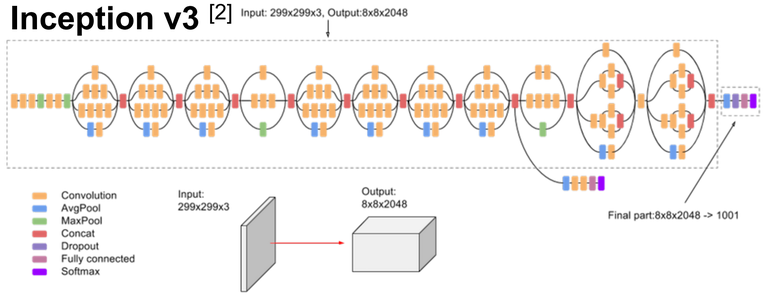
- Will two distinct sets of the same concept produce a similar CAV?
- How do CAVs from activations of different layers change?
- Does changing the negative class affect CAVs?
- How does the linear model selection affect CAVs?

PCA to visualize CAVs

- Principal component analysis ($n = 2$) was applied to the layer activations $f_l(x)$ and CAVs
- Both are high dimensional ($d > 10^6$) depending on chosen layer



Inception v3 [2]



Class Sensitivity

Using CAVs with image classes

- Which examples from a given class k are most/least similar to a concept C?

$$S_{C,k,l}(\boldsymbol{x}) = \nabla h_{l,k}\left(f_l(\boldsymbol{x})\right) \cdot \boldsymbol{v}_C^l$$

Most (left) and least (right) stripy zebra images based on human-defined striped class

Quantifying the sensitivity of a class to a concept

- The TCAV score computes the fraction of example images that have a positive directional derivative with respect to a concept

$$\text{TCAV}_{Q_{C,k,l}} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}$$

Future Work

TCAV

- Test preliminary results with more concepts and classes
- Show the change in TCAV score for how high/low level concepts throughout layers
- Use the deep dream method to visualized learned concepts
- Consider alternate computation of TCAV score

Generative Model

- Interpretable Lens Variable Model (ILVM) [3]
  - VAE with *side information* trained with a human in the loop process to maximize interpretability
  - Test the learned representations on downstream tasks

References

[1] Been Kim et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)". In: *arXiv preprint arXiv:1711.11279* (2017).

[2] Szegedy, Christian et al. "Rethinking the Inception Architecture for Computer Vision." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): n. pag. Crossref. Web.

[3] Tameem Adel, Zoubin Ghahramani, and Adrian Weller. "Discovering Interpretable Representations for Both Deep Generative and Discriminative Models". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmassan, Stockholm Sweden: PMLR, Oct. 2018, pp. 50-59. url: http://proceedings.mlr.press/v80/adel18a.html.