Crossing Cross-Entropy:

The Power of Provably Faithful Interpretability

J. Setpal

October 10, 2024

MACHINE LEARNING @ PURDUE

Outline

1. The Approach

2. Results & What's Next

What is Cross Entropy?

Cross-Entropy is the *premier* cost function to quantify classification error:

$$H(y, \hat{y}) = -\sum_{c \in C} y_c \log(\hat{y}_c) \tag{1}$$

where y is a one-hot-encoded vector of the target class, & $\hat{y}$ is the model prediction.

What is Cross Entropy?

Cross-Entropy is the *premier* cost function to quantify classification error:

$$H(y, \hat{y}) = -\sum_{c \in C} y_c \log(\hat{y}_c) \tag{1}$$

where y is a one-hot-encoded vector of the target class, & $\hat{y}$ is the model prediction.

Despite being performant, optimization is still *very* non-convex.

What is Cross Entropy?

Cross-Entropy is the *premier* cost function to quantify classification error:

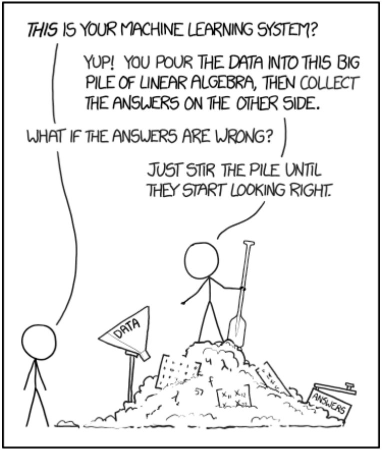
$$H(y, \hat{y}) = -\sum_{c \in C} y_c \log(\hat{y}_c) \tag{1}$$

where y is a one-hot-encoded vector of the target class, & $\hat{y}$ is the model prediction.

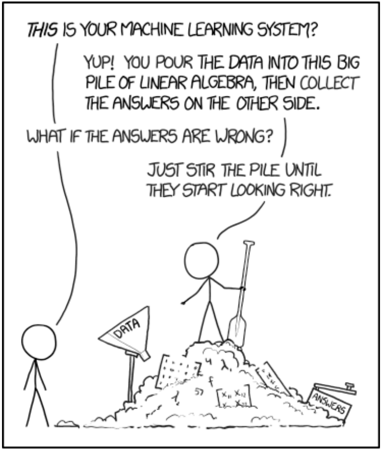
Despite being performant, optimization is still *very* non-convex.

Can we do better?

What is Interpretability?

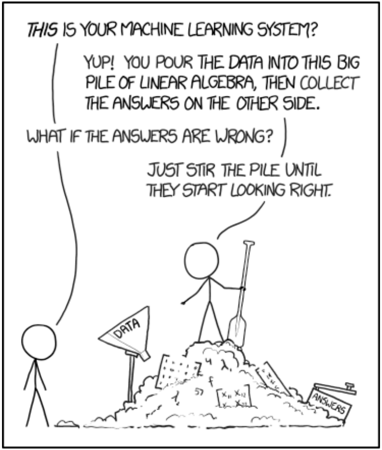


What is Interpretability?



Interpretability within Machine Learning is the **degree** to which we can understand the **cause** of a decision, and use it to consistently predict the model's prediction.

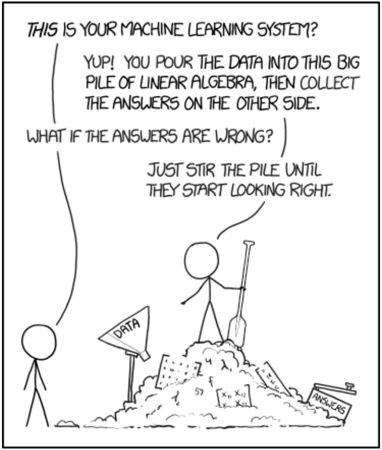
What is Interpretability?



Interpretability within Machine Learning is the **degree** to which we can understand the **cause** of a decision, and use it to consistently <u>predict the model's prediction</u>.

Traditionally, interpretability & performance is seen as a trade-off.[a]

What is Interpretability?



Interpretability within Machine Learning is the **degree** to which we can understand the **cause** of a decision, and use it to consistently predict the model's prediction.

Traditionally, interpretability & performance is seen as a trade-off.[a]

Our work demonstrates a deep intersect between these two *seemingly* orthogonal research foci.

[a] sd

Outline

Contrastive Activation Maps

HiResCAMs are a provably faithful interpretability technique:

$$\mathcal{A}_c^{\text{HiResCAMs}} = \sum_{f=1}^{F} \frac{\partial \hat{y}_c}{\partial A_f} \odot A_f \tag{2}$$

Contrastive Activation Maps

HiResCAMs are a provably faithful interpretability technique:

$$\mathcal{A}_c^{\text{HiResCAMs}} = \sum_{f=1}^{F} \frac{\partial \hat{y}_c}{\partial A_f} \odot A_f \tag{2}$$

Provably faithful because:

$$\hat{y}_c = \sum_{d_1=1, d_2=1}^{D_1, D_2} \mathcal{A}_{c,d_1,d_2}^{\text{HiResCAMs}} + b_c \tag{3}$$

Contrastive Activation Maps

HiResCAMs are a provably faithful interpretability technique:

$$\mathcal{A}_c^{\text{HiResCAMs}} = \sum_{f=1}^{F} \frac{\partial \hat{y}_c}{\partial A_f} \odot A_f \tag{2}$$

Provably faithful because:

$$\hat{y}_c = \sum_{d_1=1, d_2=1}^{D_1, D_2} \mathcal{A}_{c,d_1,d_2}^{\text{HiResCAMs}} + b_c \tag{3}$$

However, softmax-activated multi-class classification relies on **inter-class logit differences**!!!, while HiResCAMs re-construct *absolute values*.

Contrastive Activation Maps

Therefore, we define **ContrastiveCAMs**:

$$\tilde{\mathcal{A}}^{\text{contrastive}}_{(c_t, c_{t'})} := \left\{ \tilde{\mathcal{A}}^{\text{HiResCAM}}_{c_t} - \tilde{\mathcal{A}}^{\text{HiResCAM}}_{c_{t'}} \right\}^{|c|-1}_{c_{t'} \in c \setminus c_t} \tag{4}$$

This creates a new objective function, equivalent to cross-entropy:[1]

$$\max_{\theta} \sum_{d_1, d_2}^{D_1, D_2} \tilde{\mathcal{A}}^{\text{contrastive}}_{(c, c'), d_1, d_2} \ \forall c' \in \mathbb{Z}_+(|c| - 1) \tag{5}$$

With one key difference: we've preserved spatial information.

[1] with subtle changes to the architecture

The Fault in Our Cross-Entropy

Outline

1. The Approach

2. Results & What's Next

Results (so far)

What's to Come

Next, we are going

Thank you!

Have an awesome rest of your day!

Slides: `https://cs.purdue.edu/homes/jsetpal/slides/cont-opt.pdf`
Code: `https://dagshub.com/jinensetpal/contrastive-optimization`