Towards Aggregating Weighted Feature Attributions

Umang Bhatt, Pradeep Ravikumar, and José Moura
Carnegie Mellon University

Carnegie Mellon University

Electrical & Computer ENGINEERING

Overview

- We propose a method to combine feature attributions via [1, 2] with a local neighborhood influence measure proposed in [3]. Specifically, we weight feature attributions of k training points by their importance to a test point and aggregate the k attributions into a consensus attribution.
- We also explore aggregating various feature attribution techniques in order to maximize a pre-selected evaluation criteria.

Weighting Explanations

We can explain a test point, x_{test}, by analyzing and aggregating attributions of training points near the test point. Using the approximation in [3], we define the influence weight, $\rho_j \in \mathbb{R}_{\geq 0}$, of training point, $x^{(j)}$, on test point, x_{test} as:

$$\rho_j = \frac{d}{d\epsilon}\mathcal{L}(f_{\epsilon,x^{(j)}}, x_{\text{test}})\big|_{\epsilon=0}$$

We then select the local neighborhood, $\mathcal{N}_k$, of the k most influential training points on x_{test}.

$$\mathcal{N}_k(x_{\text{test}}, \mathcal{D}) = \underset{\mathcal{N}\subset\mathcal{D}, |\mathcal{N}|=k}{\arg\max} \sum_{x^{(j)}\in\mathcal{N}} \rho_j$$

Suppose we get a Shapley value explanation, ϕ^j, for every point in $\mathcal{N}_k$. [4] proposed the weighted Shapley value which would weigh every contribution by a player's weight. In our case, we weigh each feature's contribution from every influential point ($x^{(j)}$) by its influence weight (ρ_j).

$$\phi_i(x^{(j)}) = \sum_{S\subseteq F\setminus\{i\}} \frac{\rho_j}{\rho} R\left(f_T(x_T) - f_S(x_S)\right)$$

Let $\rho = \sum_{i\in S}\rho_i$. Since Shapley values allow for scaling and additivity, we can sum attributions across all influential datapoints and simplify.

$$\mathcal{A}_{\text{SHAP}}(\phi, \mathcal{N}_k) = \sum_{x^{(j)}\in\mathcal{N}_k} \frac{\rho_j}{\rho}\phi^j$$

A similar derivation can be followed for Integrated Gradients. We could have also leveraged traditional rank aggregation techniques (i.e., Borda Count and Markov Chains) to combine the k attributions.

Experimentation

We run tabular experiments to show the utility of weighted explanations (particularly weighted Shapley values) and to show the intuitive results of aggregating various explanations with images.

MIMIC-III

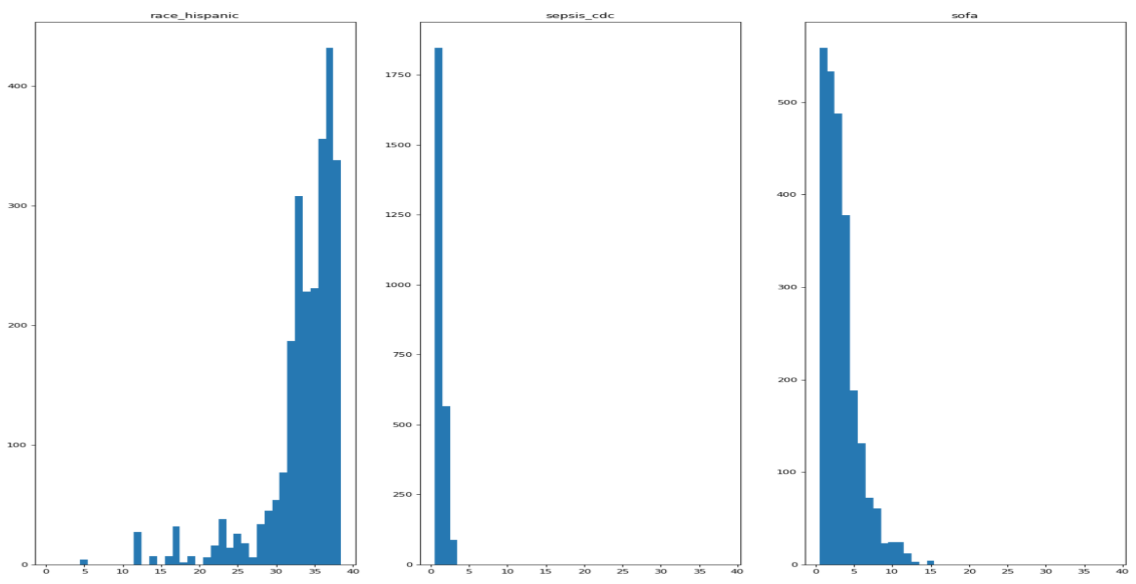
We explain a sepsis prediction model trained on a dataset [5] consisting of 11,791 hospital admissions with 38 semantically meaningful features (physical descriptors, lab results, indicators).

Faithfulness via recall: Let $F' \subset F$ be the top b features of an interpretable model h. Let S_i be the top b features from ε_A. We measure:

$$\mathsf{faithfulness} = \frac{1}{N}\sum_{i=1}^{N}\frac{|S_i \cap F'|}{|F'|}$$

Model	Acc.	SHAP	IG	$\mathcal{A}_{\text{SHAP}}$	$\mathcal{A}_{\text{IG}}$
1 HL-S	85.3	60	29	**68**	37
1 HL-R	82.8	62	33	**69**	47
2 HL-S	86.7	61	34	**75**	41
2 HL-R	87.2	55	35	**64**	35
3 HL-S	83	64	31	**68**	41
3 HL-R	87	55	38	**65**	48

Histogram of accumulated rankings for representative MIMIC features:



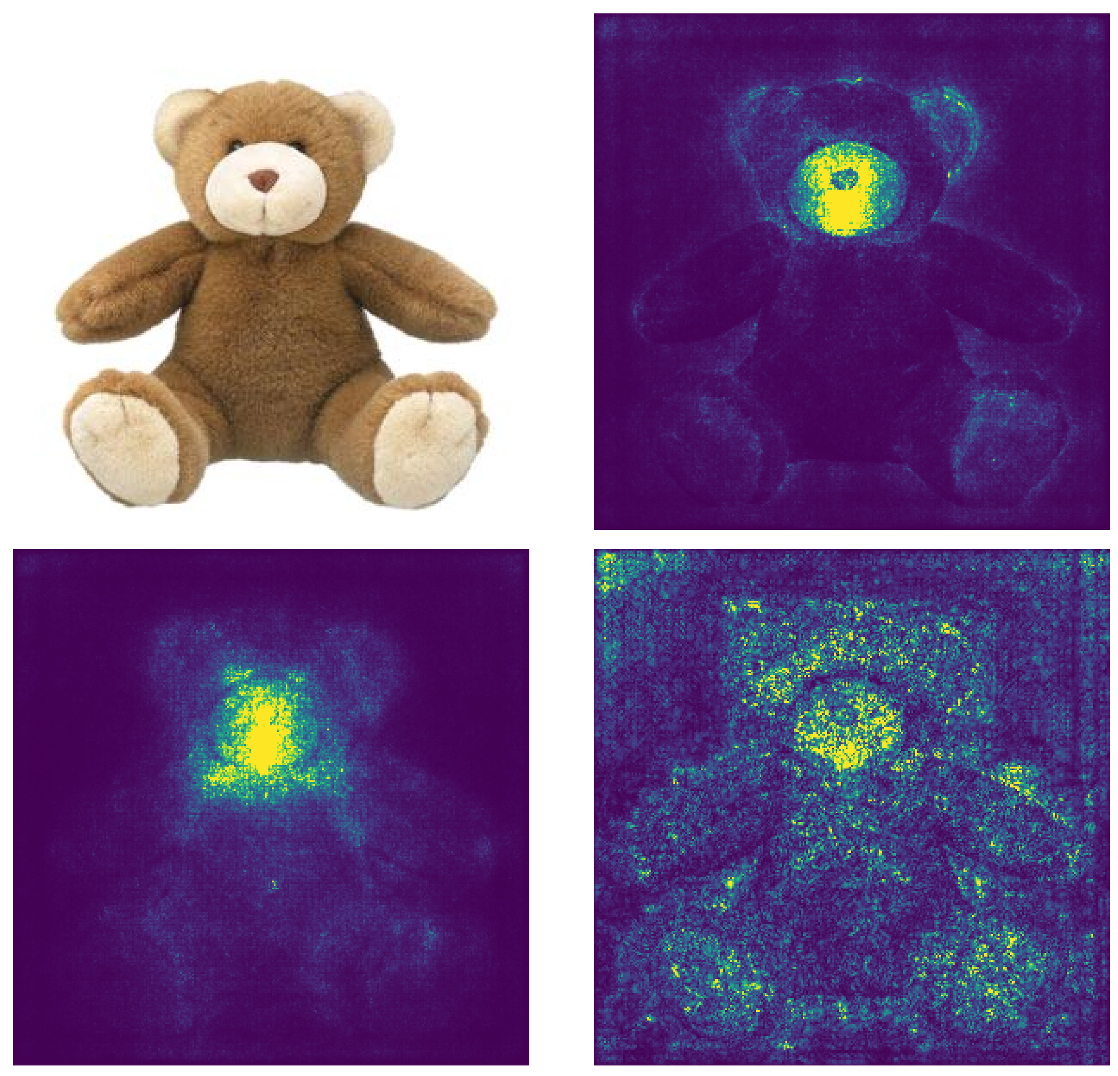
Cumulative count vs Ranking of Features

ImageNet

We attempted to learn an aggregate explanation that maximized *sensitivity* [6].

We define sensitivity as the Pearson correlation coefficient between the sum of the attributions ($\sum_{i=1}^{d}\varepsilon_i$) and the residual effect on the model output of randomly zeroing out pixels in the original image $f(x) - f(x_{[S=0]})$.

Below is the result of aggregating saliency maps subject to maximizing sensitivity.



Clockwise from Top Left: Original, Aggregate, Integrated Gradients, SmoothGrad

Aggregating Across Explanations

We also explore aggregating different explanation techniques to maximize user-defined criteria. Suppose a user wants to find an aggregate explanation, ε_{agg}, that maximizes both faithfulness and sensitivity equally. Alternatively, users can add weights on individual criteria. The simplest form of ε_{agg} would be a convex combination of the different explanation techniques.

$$\varepsilon_{agg} = w^T\Phi$$

$$\Phi^T = \begin{pmatrix} | & | & & | \\ \text{LIME} & \text{IG} & \cdots & \text{SHAP} \\ | & | & & | \end{pmatrix}$$

To learn ε_{agg}, we can maximize the two criteria as follows.

$$\underset{w}{\arg\max}\sum_{i=1}^{N} \text{faithfulness}(w^T\Phi_i) + \text{sensitivity}(w^T\Phi_i)$$

Alternatively, we can use traditional rank aggregation to aggregate Φ into a singular explanation ε_{agg}. We use the following formulation based on centroids [7, 8] with respect to some distance $d : \mathcal{E}\times\mathcal{E} \mapsto \mathbb{R}$ and then change the criteria maximization accordingly for any arbitrary metric.

$$\varepsilon_{agg} = \mathcal{A}(g, \mathcal{N}_k) \in \underset{\varepsilon\in\mathcal{E}}{\arg\min}\sum_{x\in N_k} d(\varepsilon, g(x))$$

$$\max\sum_{i=1}^{N}\text{metric}(\varepsilon_{agg})$$

References

[1] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

[2] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*. 2017.

[3] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*, 2017.

[4] E. Kalai and D. Samet. On weighted shapley values. *Int. J. Game Theory*, 16(3):205–222, September 1987.

[5] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmark of deep learning models on large healthcare mimic datasets, 2017.

[6] Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018.

[7] Nina Narodytska and Toby Walsh. The computational impact of partial votes on strategic voting. In *ECAI*, 2014.

[8] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2474–2482. 2012.