Explaining Explanations: : Axiomatic Feature Interactions for Deep Networks

Joseph D. Janizek, Pascal Sturmfels, Su-In Lee

Presenter: Arshdeep Sekhon

https://qdata.github.io/deep2Read

Introduction

- for many tasks, simply knowing which features were important to a model's prediction may not provide enough insight to understand model behavior.
- The interactions between features within the model may better help us understand the model, and why certain features are more important than others.
- This paper: Extension of Integrated Gradients: Integrated Hessians

Method: Integrated Gradients Review

- Model is a function $f(x) : \mathbb{R}^d \rightarrow \mathbb{R}$
- Say x' is some baseline value
- We want to explain feature i for sample x
- Feature attribution of feature i using Integrated Gradients:

$$\phi_i(x) = (x_i - x_i') \int_0^1 \partial \frac{f(x' + (\alpha(x - x')))}{\partial x_i} d\alpha \qquad (1)$$

- To use Integrated Gradients method, the only requrement is the function be differentiable from x' to x.

Method: Integrated Hessians

- Integrated Gradients

$$\phi_i(x) = (x_i - x_i') \int_0^1 \partial \frac{f(x' + (\alpha(x - x'))}{\partial x_i} d\alpha \tag{2}$$

- The Integrated Gradients for a differentiable model $f(x) : \mathbb{R}^d \to \mathbb{R}$ is itself a differentiable function $\phi_i(x) : \mathbb{R}^d \to \mathbb{R}$.
- Now change problem to feature attribution of j to function $\phi_i(x)$.

$$\Gamma_{i,j}(x) = \phi_j(\phi_i(x)) \tag{3}$$

Method: Integrated Hessians

explanation of the importance of feature i in terms of the input value of feature j. If $i \neq j$:

$$\Gamma_{i,j}(x) = (x_i - x_i')(x_j - x_j') \times$$
$$\int_{\beta=0}^{1} \int_{\alpha=0}^{1} \alpha\beta \frac{\partial^2 f(x' + \alpha\beta(x - x'))}{\partial x_i \partial x_j} d\alpha d\beta \tag{4}$$

Some caveats:

- proof requires f must satisfy Leibniz Integral Rule: so that integration and differentiation are interchangeable
- requires function and derivative are continuous over x in the integration region: ReLU can't be explained

Integrated Hessians Derivation

$$\Gamma_{i,j}(x) := (x_j - x'_j) \times \int_{\beta=0}^{1} \frac{\partial \phi_i(x' + \beta(x - x'))}{\partial x_j} d\beta \tag{5}$$

Consider the function $\frac{\partial \phi_i}{\partial x_j}(x)$, and we first assume that $i \neq j$

$$\frac{\partial \phi_i}{\partial x_j}(x) = \tag{6}$$

$$(x_i - x'_i) \times \frac{\partial}{\partial x_j} \left(\int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha \right) = \tag{7}$$

$$(x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial}{\partial x_j} \left(\frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} \right) d\alpha = \tag{8}$$

$$(x_i - x'_i) \times \int_{\alpha=0}^{1} \alpha \frac{\partial^2 f(x' + \alpha(x - x'))}{\partial x_i \partial x_j} d\alpha \tag{9}$$

Integrated Hessians Derivation

We can proceed by plugging equation 9 into the original definition of $\Gamma_{i,j}(x)$:

$$\Gamma_{i,j}(x) := (x_j - x'_j) \times \int_{\beta=0}^{1} \frac{\partial \phi_i(x' + \beta(x - x'))}{\partial x_j} d\beta \tag{10}$$

$$= (x_j - x'_j) \times \int_{\beta=0}^{1} (x'_i - \beta(x_i - x'_i) - x'_i) \tag{11}$$

$$\int_{\alpha=0}^{1} \alpha \frac{\partial^2 f(x' + \alpha(x' - \beta(x - x') - x'))}{\partial x_i \partial x_j} d\alpha d\beta \tag{12}$$

$$= (x_j - x'_j)(x_i - x'_i) \int_{\beta=0}^{1} \int_{\alpha=0}^{1} \alpha\beta \frac{\partial^2 f(x' + \alpha\beta(x - x'))}{\partial x_i \partial x_j} d\alpha d\beta \tag{13}$$

Fundamental Axioms for Interaction Values: Interaction Completeness

$$\sum_i \sum_j \Gamma_{i,j}(x) = f(x) - f(x') \tag{14}$$

Satisfying interaction completeness is important because it demonstrates a relationship between model output and interaction values. Without this axiom, it is unclear how to interpret the scale of interactions.

Fundamental Axioms for Interaction Values: Self Completeness

$$\Gamma_{i,i}(x) = \phi_i(x) - \sum_{j \neq i} \Gamma_{i,j}(x) \tag{15}$$

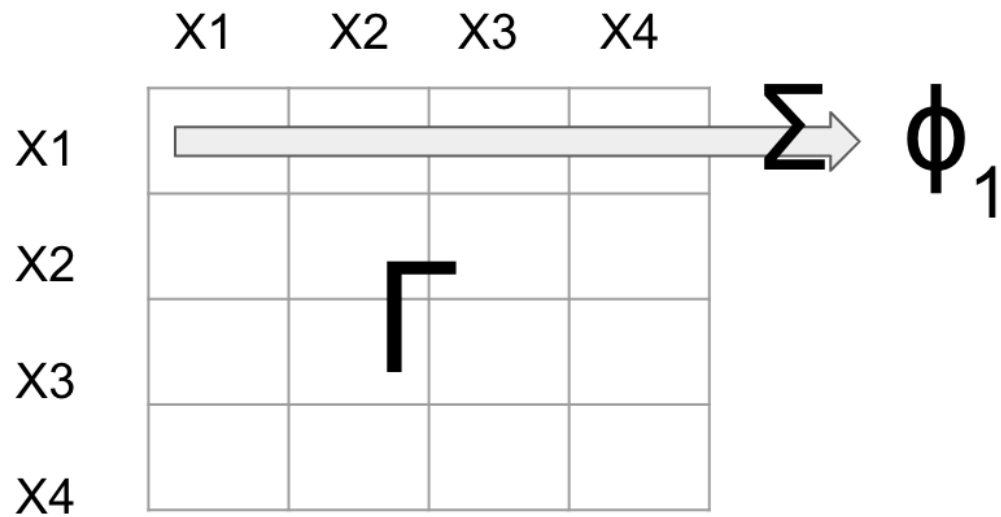


Figure: Self Completeness

main effect of feature i after interactions with all other features have been subtracted away. So if only one feature, or $\Gamma_{i,j} = 0$: $\Gamma_{i,i}(x) = \phi_i(x)$

Other Axioms

- interaction symmetry $\Gamma_{i,j} = \Gamma_{j,i}$
- interaction sensitivity and
- interaction linearity

Computing Integrated Hessians

- discrete sum approximation of the integral, similar to how Integrated Gradients
- 50 to 300 discrete steps suffice to approximate the double integral in most cases.

Smoothing ReLU Networks

$$\mathrm{SoftPlus}(x) = \log(1 + e^x). \tag{16}$$

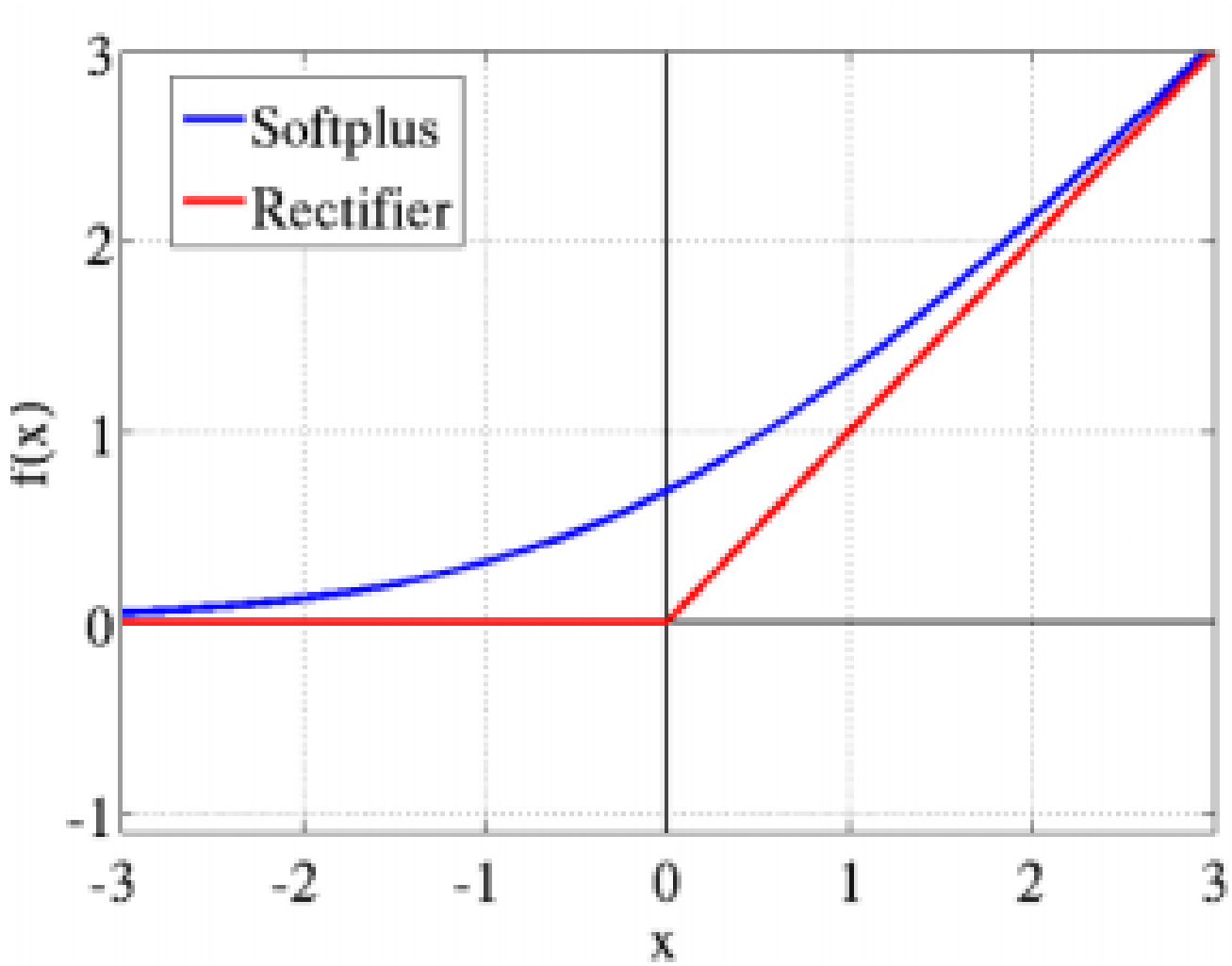


Figure: smooth approximation of ReLU

Smoothing ReLU Networks

- ReLU: second partial derivatives equal to zero in all places
- the ReLU activation function has a smooth approximation – the SoftPlus function:

$$\mathrm{SoftPlus}_\beta(x) = \frac{1}{\beta}\log(1 + e^{\beta x}). \tag{17}$$

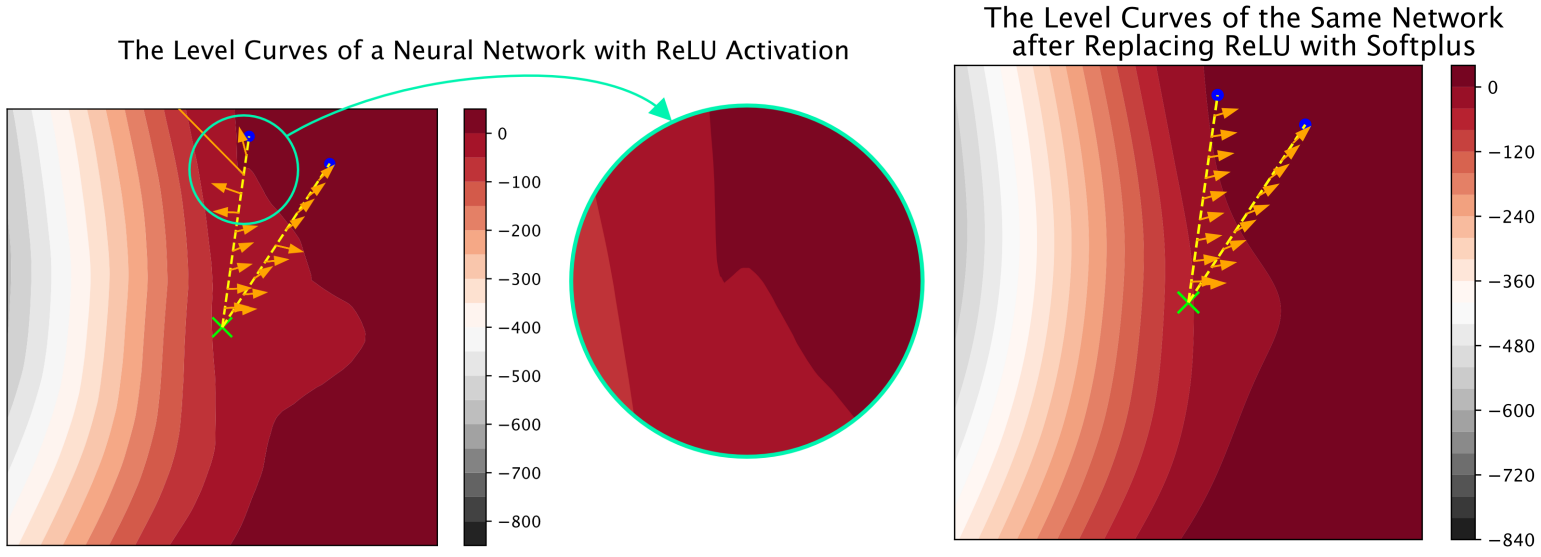
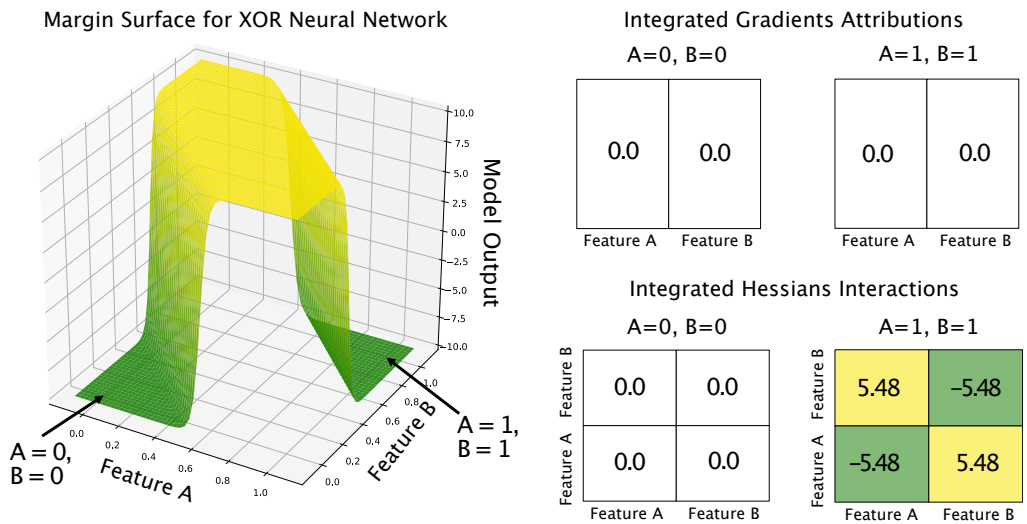


Figure: Replacing ReLU activations with $\mathrm{SoftPlus}_\beta$ activations with $\beta = 10$ smooths the decision surface of a neural network: gradients tend to be more homogeneous along the integration path. Orange arrows show the gradient vectors at each point along the path from the reference (green x) to the input (blue dots).

Experiments: Explaining XOR

- Consider an XOR neural network
- because both features are on, which on their own should increase the model output, but in interaction with each other cancel out the positive effects and drive the model's output back to the baseline.



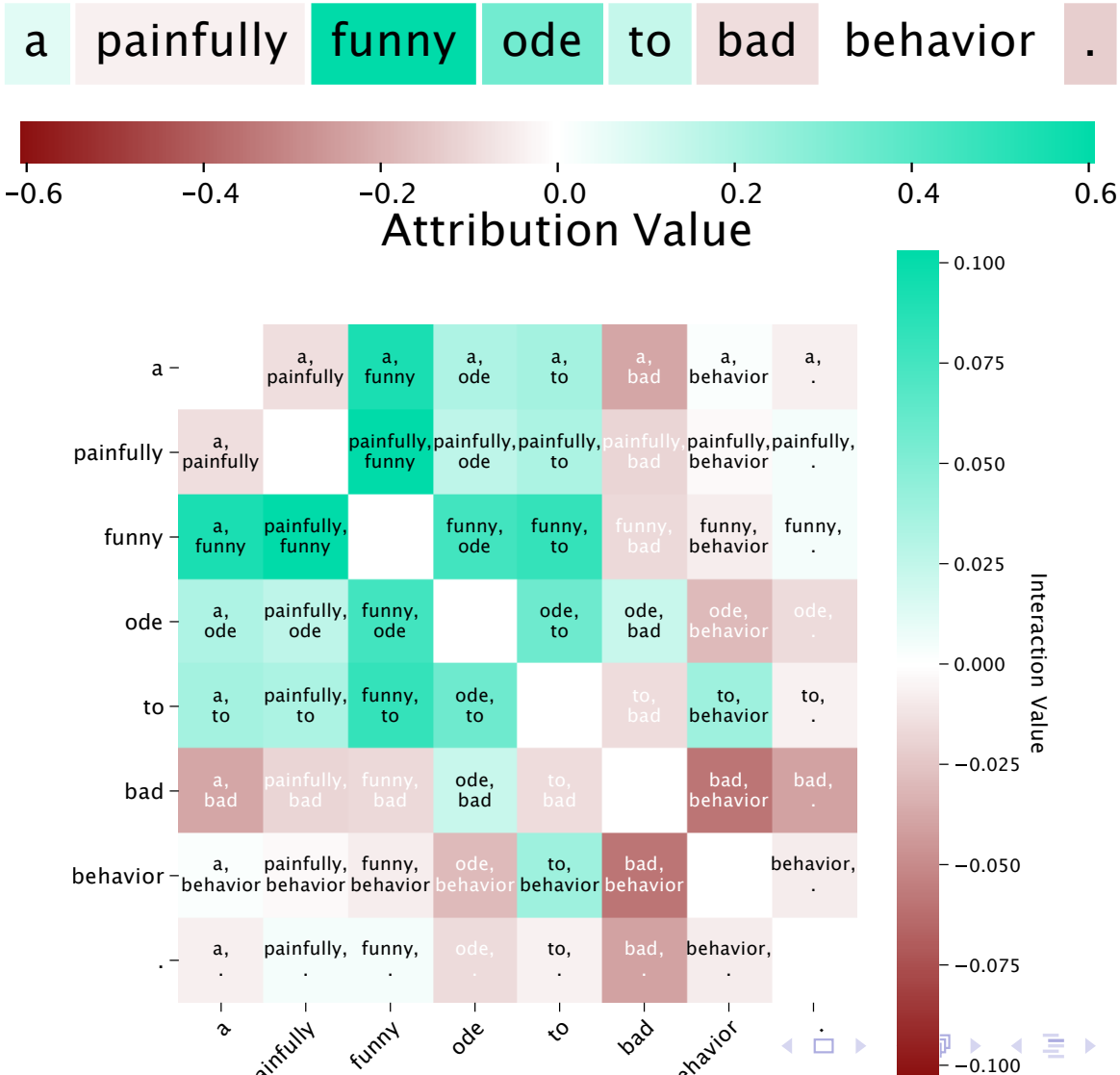
Figure

Experiments: Explaining XOR

- input gradients and input hessians have completely flattened (saturated) at all points
- By integrating between the baseline and the samples, Integrated Hessians is capable of correctly detecting the negative interaction between the two features.

NLP

- pre-trained weights for DistilBERT
- fine-tune the model on the Stanford Sentiment Treebank dataset: predict movie review as positive or negative sentiment



NLP: Saturation Effects

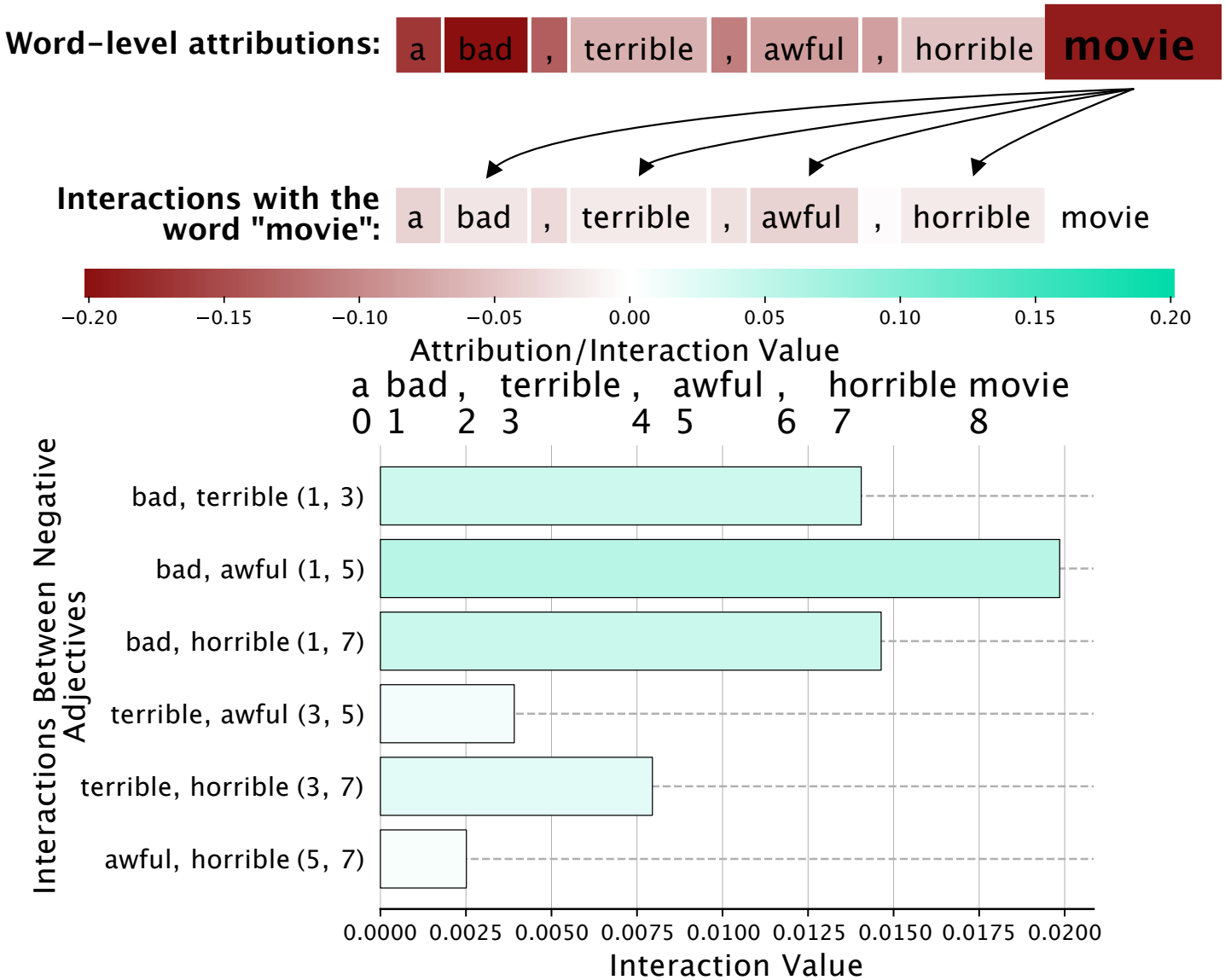


Figure: Although the word "movie" interacts negatively with all negative modifying adjectives, those negative adjectives themselves all interact positively. The more negative adjectives are in the sentence, the less each individual negative adjective matters towards the overall classification of the sentence.

NLP: Trying to understand why some models perform better than others

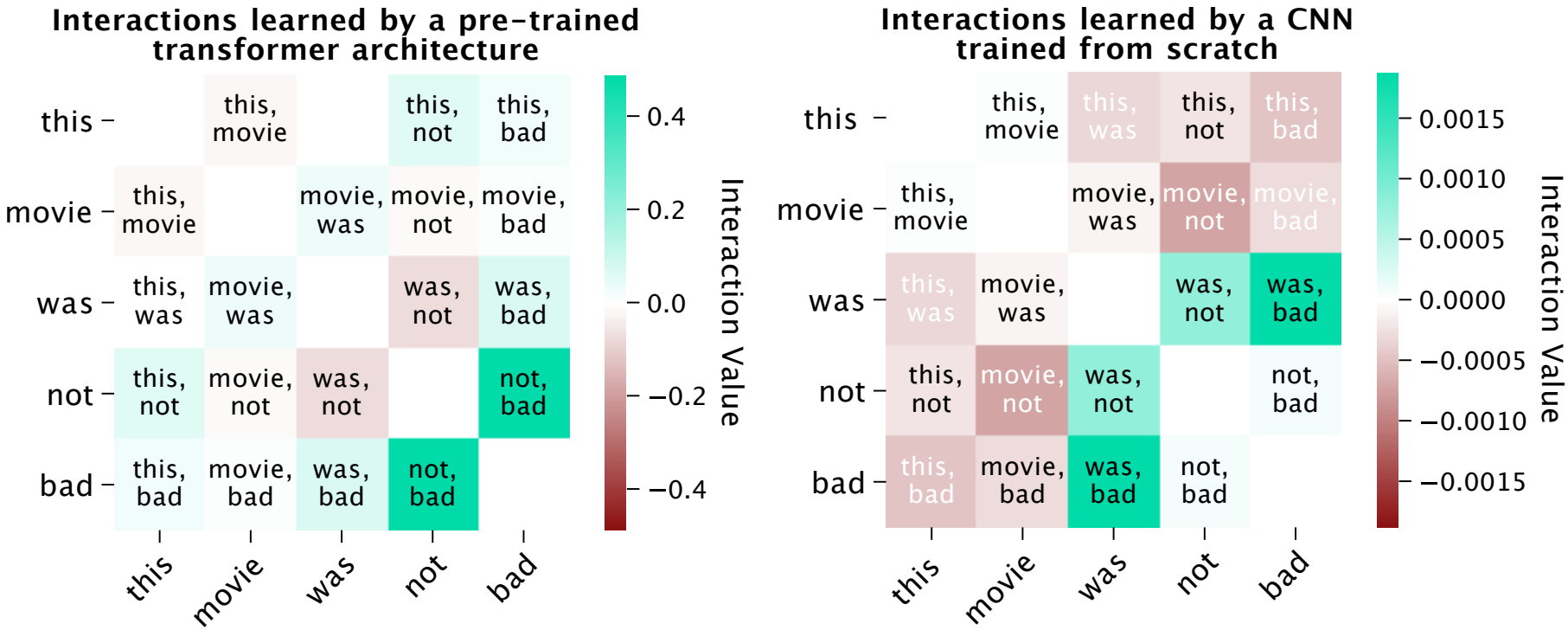
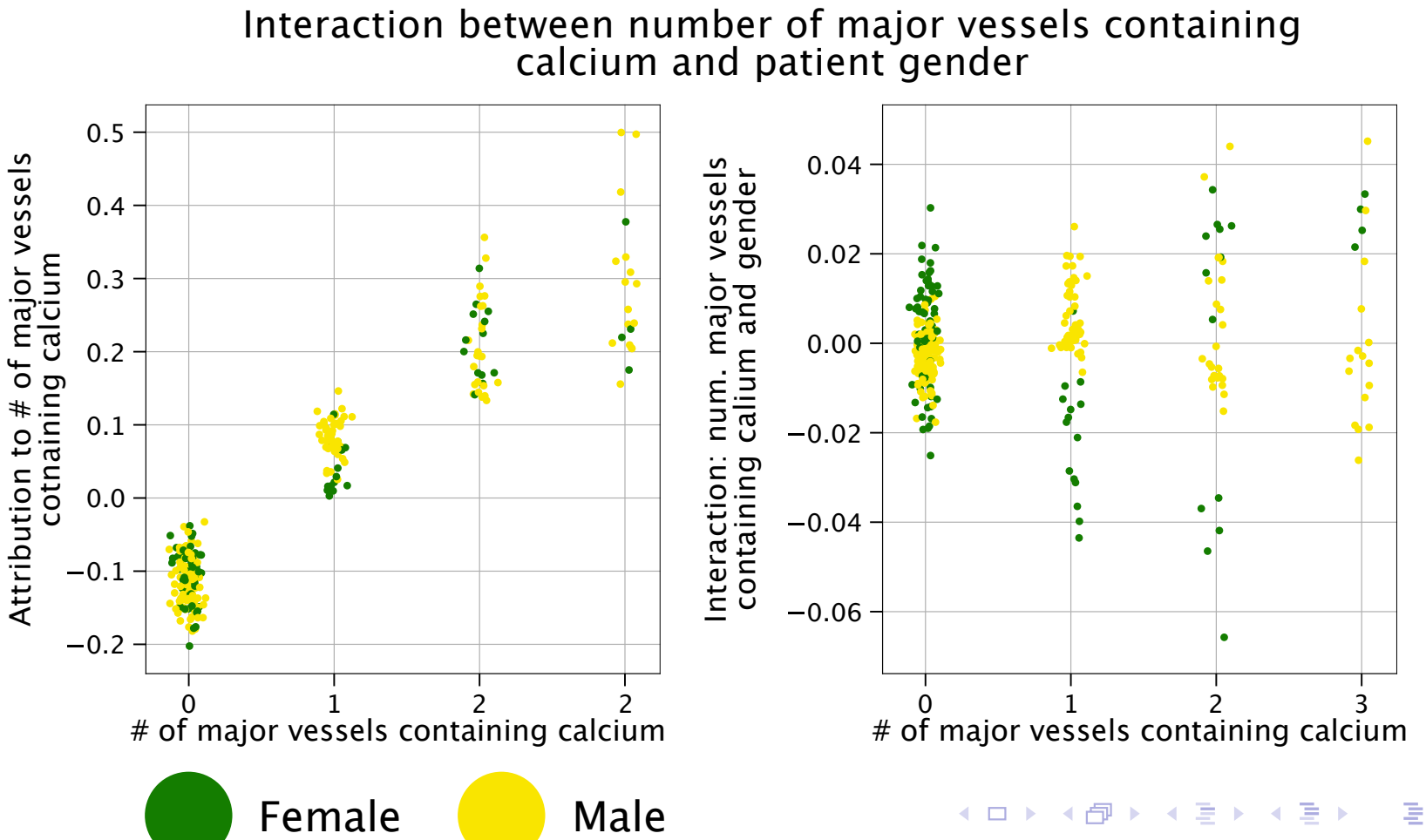


Figure: Here, we examine interactions on the sentence "this movie was not bad." We compare two models trained to do sentiment analysis on the Stanford Sentiment dataset: a pre-trained transformer, DistilBERT (left), (98.2% confidence), and a convolutional neural network trained from scratch (97.6% confidence). The transformer picks up on negation patterns: "not bad" has a positive interaction, despite the word "bad" being negative. The CNN mostly picks up on negative interactions like "movie not" and "movie bad".

Heart Disease Prediction

- 298 patients with 13 associated features
- When the Expected Hessians interactions are aggregated across the dataset, they reveal that our model has learned that women with calcium deposition in one coronary artery are less likely than men to be diagnosed with coronary artery disease



Drug combination response prediction

- drug combination response in acute myeloid leukemia
- Features: drugs and targets and gene expression for cancer cells
- presence or absence of the drug Venetoclax in the drug combination is the most important feature
- not enough: while the presence of Venetoclax is generally predictive of a more responsive drug combination, the amount of positive response to Venetoclax is predicted to vary across samples.
- variablity in drug respose: which drug it is combined with

Drug combination response prediction

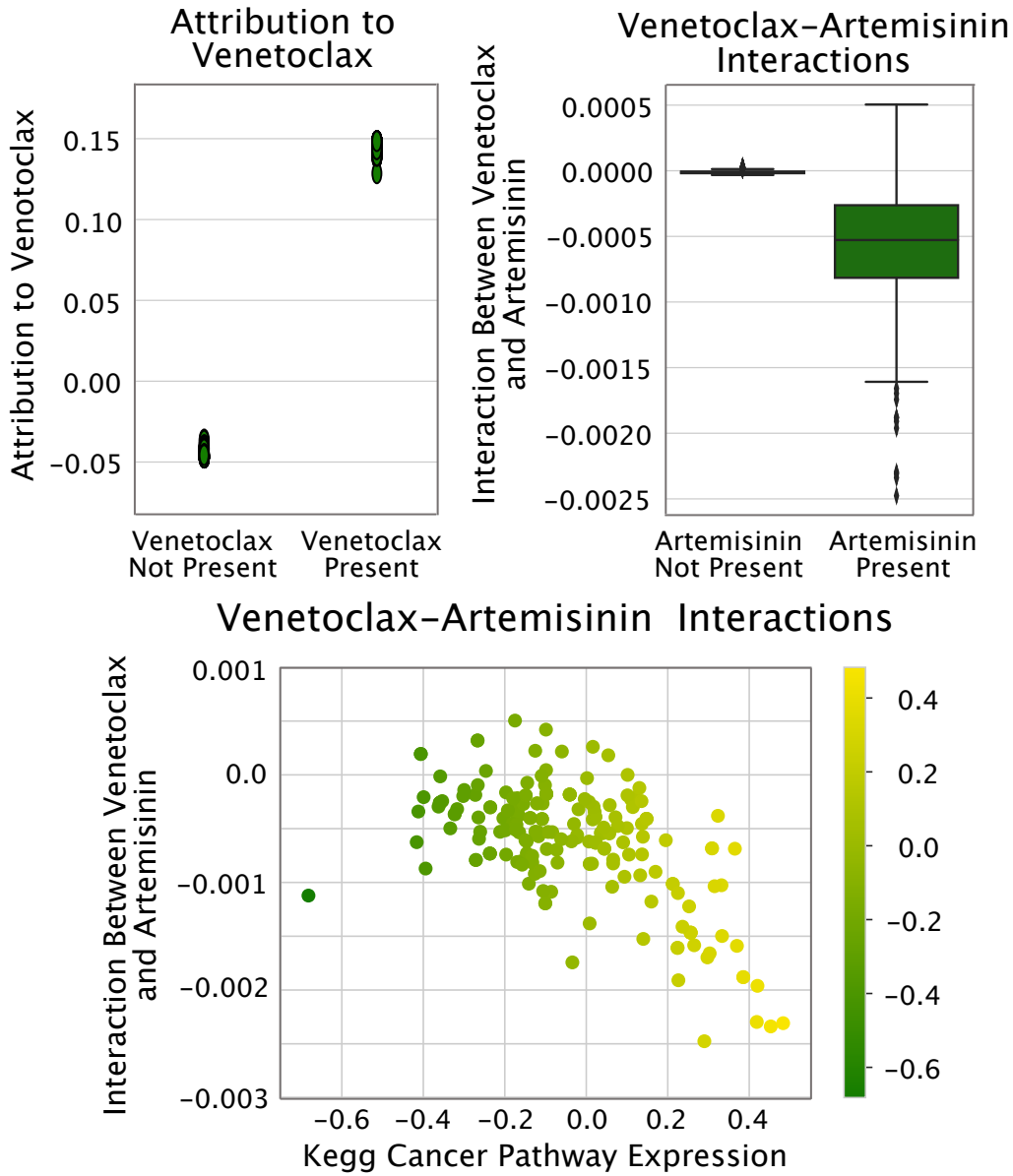


Figure: Top Left: Integrated Gradients values for Venetoclax. Top Right: Venetoclax interactions with Artemisinin across all samples. Bottom: Venetoclax and Artemisinin interaction is driven by expression of genes in cancer samples.

Conclusions

- Interaction as the combined effect of two features to the output of a model,
- the explanation of one feature's importance in terms of another.

Baselines and Expected Hessians

- Baseline: some data that is uninformative
- Single Baseline is challenging in some datasets:
- Use Expected Gradients Instead:

$$\phi_i^{EG}(x) = \mathbb{E}_{x',\alpha}\left[(x_i - x_i')\frac{\partial f(x' + \alpha(x - x')}{dx_i}\right] \tag{18}$$

where the expectation is over both $x' \sim D$ for the training distribution D and $\alpha \sim U(0,1)$. We can apply Expected Gradients to itself to get Expected Hessians:

$$\Gamma_{i,j}^{EG}(x) =$$
$$\mathbb{E}\left[(x_i - x_i')(x_j - x_j')\alpha\beta\frac{\partial^2 f(x' + \alpha\beta(x - x')}{\partial dx_i \partial dx_j}\right] \tag{19}$$

where the expectation is over $x' \sim D$, $\alpha \sim U(0,1)$ and $\beta \sim U(0,1)$.