Learning to Understand: Identifying Interactions via the Mobius Transform

Justin S. Kang[1], Yigit E. Erginbas[1], Landon Butler[1], Prof. Ramtin Pedarsani[2], Prof. Kannan Ramchandran[1]





[1]UC Berkeley

[2]UC Santa Barbara

NeurIPS 2024

Motivation: Sentiment Analysis

Glowing Paper Review

The concept of information entropy was introduced by Claude Shannon in his 1948 paper "A Mathematical Theory of Communication", and is also referred to as Shannon entropy. Shannon's theory defines a data communication system composed of three elements: a source of data, a communication channel, and a receiver. The "fundamental problem of communication" – as expressed by Shannon – is for the receiver to be able to identify what data was generated by the source, based on the signal it receives through the channel …

Sentiment score

3/10

Motivation: Sentiment Analysis

Glowing Paper Review

The concept of information entropy was introduced by Claude Shannon in his 1948 paper "A Mathematical Theory of Communication", and is also referred to as Shannon entropy. Shannon's theory defines a data communication system composed of three elements: a source of data, a communication channel, and a receiver. The "fundamental problem of communication" – as expressed by Shannon – is for the receiver to be able to identify what data was generated by the source, based on the signal it receives through the channel …

→ →

Sentiment score

3/10

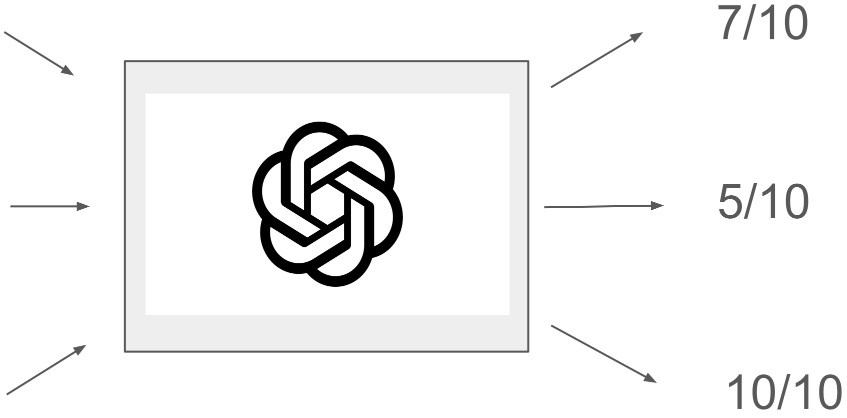
Can we understand what part of the text triggers the model to produce the (erroneous) output?

Typical Solution: Mask and Try Again

The concept of information [redacted] by Claude Shannon in his 1948 paper "A Mathematical Theory of Communication" and is also referred to as Shannon entropy. [redacted] communication [redacted] source of data, a communication channel, and a receiver. The "fundamental problem of communication" – as expressed by Shannon – is for the receiver to be able to identify what data was [redacted] …

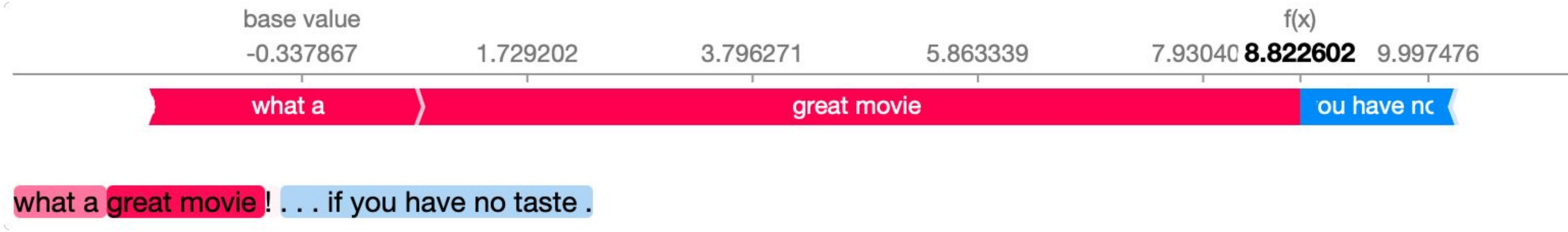
[redacted] was introduced by Claude Shannon in his 1948 paper "A Mathematical Theory of Communication"[redacted]of three elements: a source of data, a communication channel, and a receiver. The [redacted] of communication" – as expressed by Shannon – is for the receiver to be able to identify what data was generated by the source …

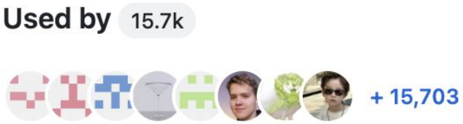
The concept [redacted]"A Mathematical Theory of Communication" and is also referred to as Shannon entropy. S[redacted] data communication system composed of three elements: a source of data, a communication channel, and a receiver. The "fundamental problem of [redacted]hannon – is for the receiver to be able to identify what data was generated by the source …



Shapley Value and SHAP

- SHAP software package: game-theoretic ***Shapley Value***
- Assigns a score to each (group of) word related to its average marginal contribution to the overall score

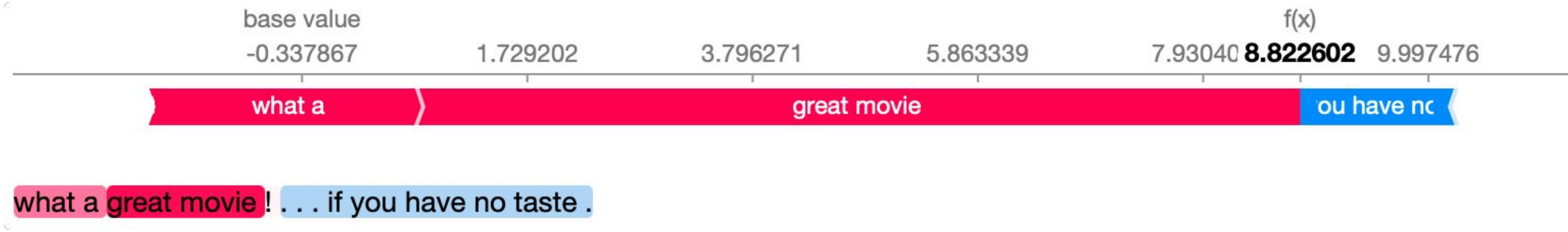




Shapley Value and SHAP

- SHAP software package: game-theoretic ***Shapley Value***
- Assigns a score to each (group of) word related to its average marginal contribution to the overall score

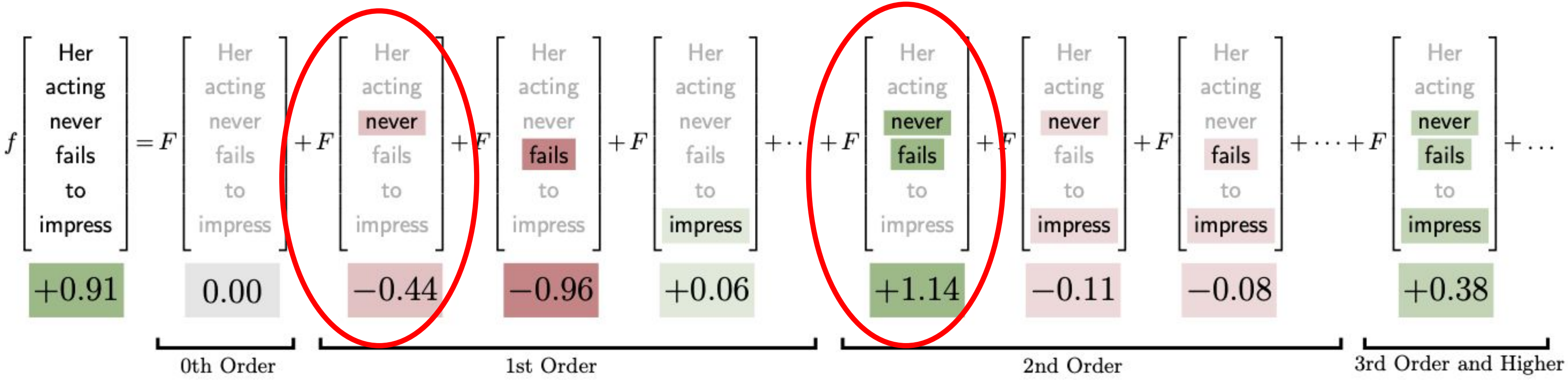
SHAP



Improving on SHAP:

1. Faster - Fewer masking patterns
2. Higher order information

Higher order information is useful

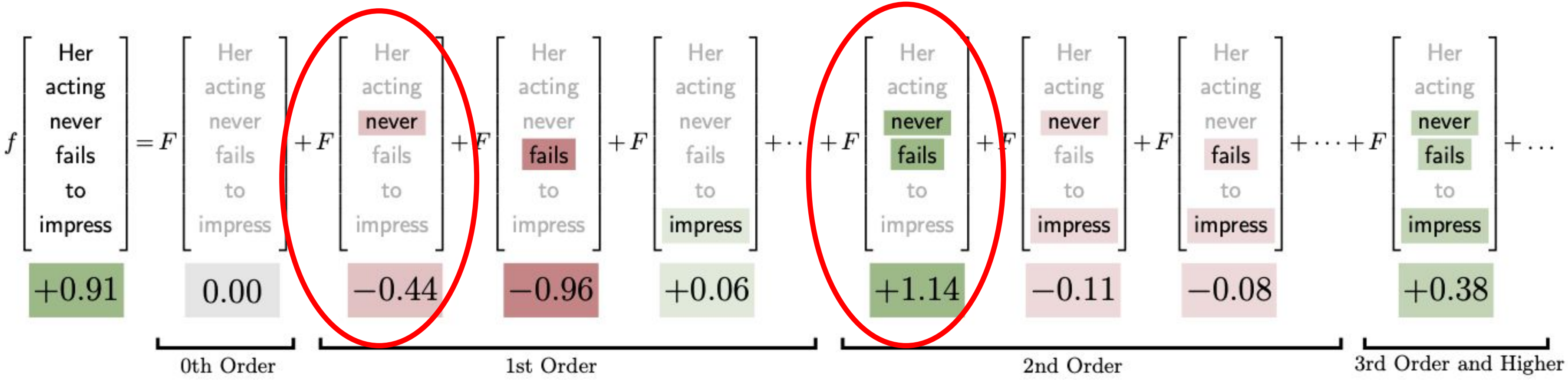


- First order information is deceptive: “**never**” is negative on its own.
- If “**never**” appears before “**fails**” connotation is positive.

Sentiment analyzer uses pretrained BERT fine-tuned on IMDB review dataset

Note: The version of the function used in this presentation differs slightly from the final version in the paper

Higher order information is useful



- First order information is deceptive: “**never**” is negative on its own.
- If “**never**” appears before “**fails**” connotation is positive.

Decomposing a function into constituent parts - **summing over all interactions**

The Mobius Transform (AND basis)

- The Mobius transform is defined as:

$$F(\mathbf{k}) = \sum_{\mathbf{m} \leq \mathbf{k}} (-1)^{\mathbf{1}^{\mathrm{T}}(\mathbf{k}-\mathbf{m})} f(\mathbf{m})$$

- The “Backwards” transform is:

$$f(\mathbf{m}) = \sum_{\mathbf{k} \leq \mathbf{m}} F(\mathbf{k})$$

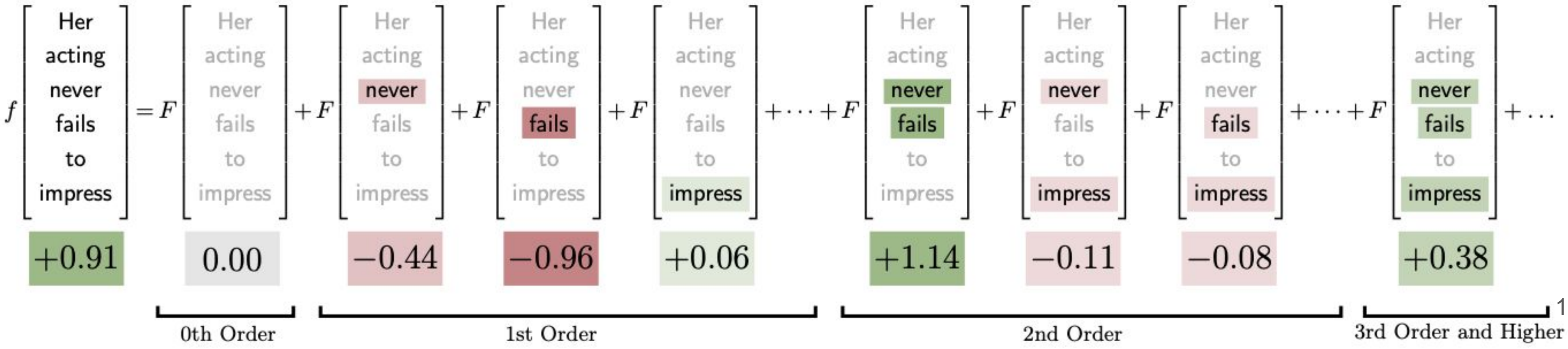
August Möbius

Gian-Carlo Rota

Signal Processing Approach for Explanations!

- Decompose the function in terms of effects of sets of inputs: polynomial

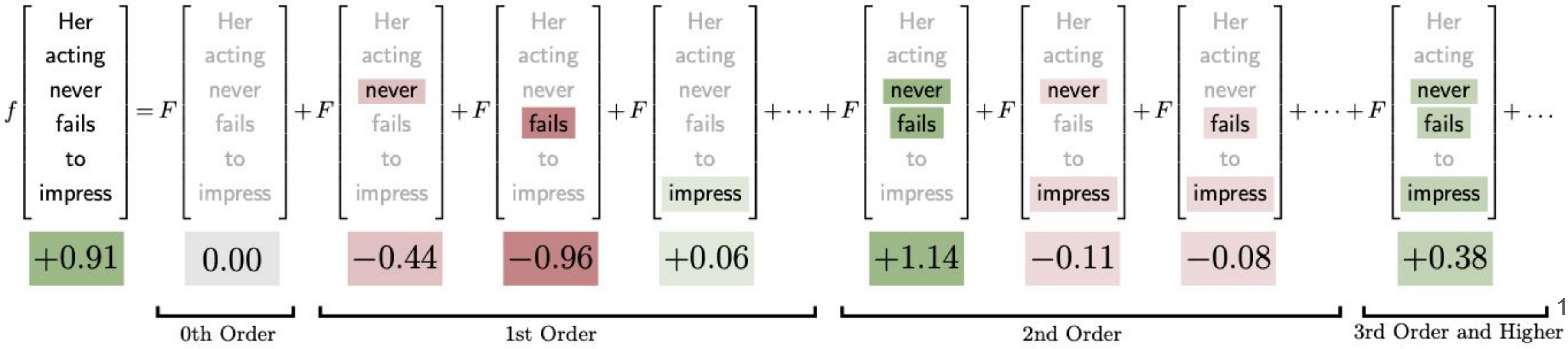
$$f(\mathbf{m}) = -0.44m_3 + -0.96m_4 + 0.06m_6 + \cdots + 1.14m_3m_4 + -0.11m_3m_6 + -0.18m_4m_6 + 0.38m_3m_4m_6 + \ldots$$



Signal Processing Approach for Explanations!

- Decompose the function in terms of effects of sets of inputs: polynomial

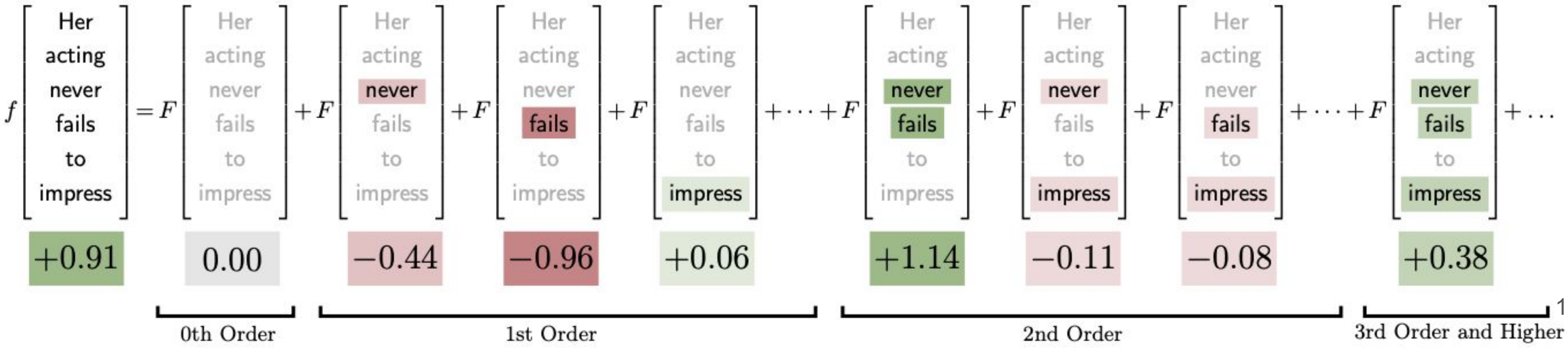
$$f(\mathbf{m}) = -0.44(1) + -0.96(1) + 0.06(1) + \cdots + 1.14(1)(1) + \\ -0.11(1)(1) + -0.18(1)(1) + 0.38(1)(1)(1) + \ldots$$



Signal Processing Approach for Explanations!

- Decompose the function in terms of effects of sets of inputs: polynomial

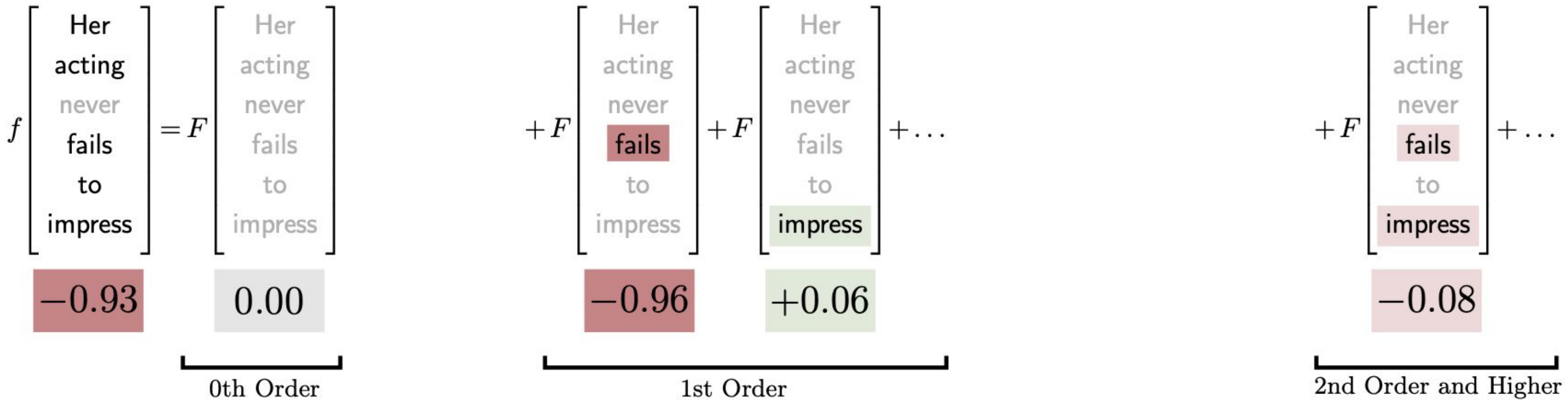
$$f(\mathbf{m}) = 0.91$$



Signal Processing Approach for Explanations!

- Decompose the function in terms of effects of sets of inputs: polynomial

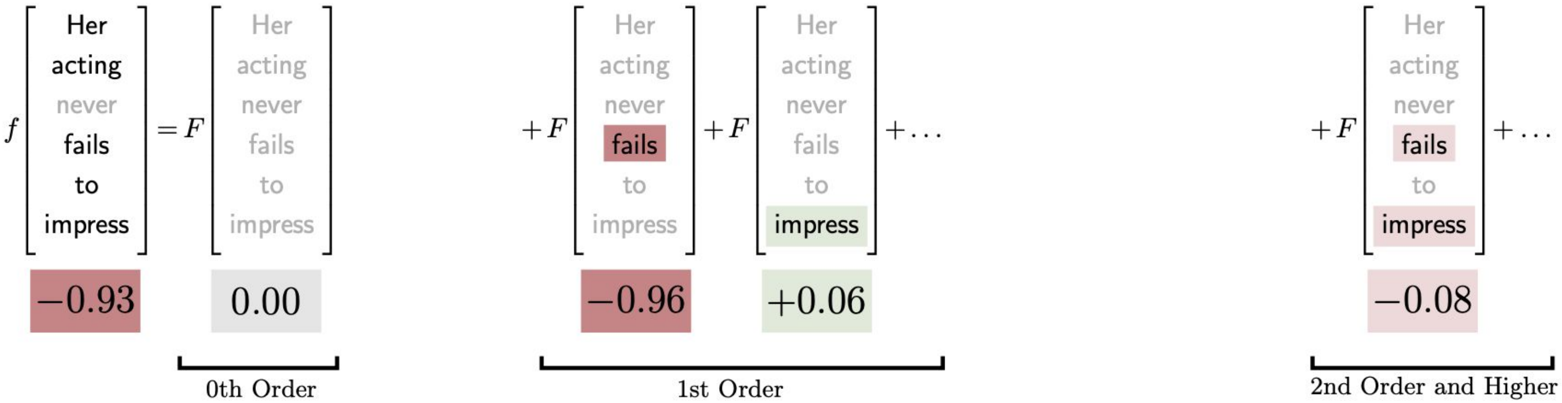
$$f(\mathbf{m}) = -0.44(0) + -0.96(1) + 0.06(1) + \cdots + 1.14(0)(1) + \\ -0.11(0)(1) + -0.18(1)(1) + 0.38(0)(1)(1) + \ldots$$



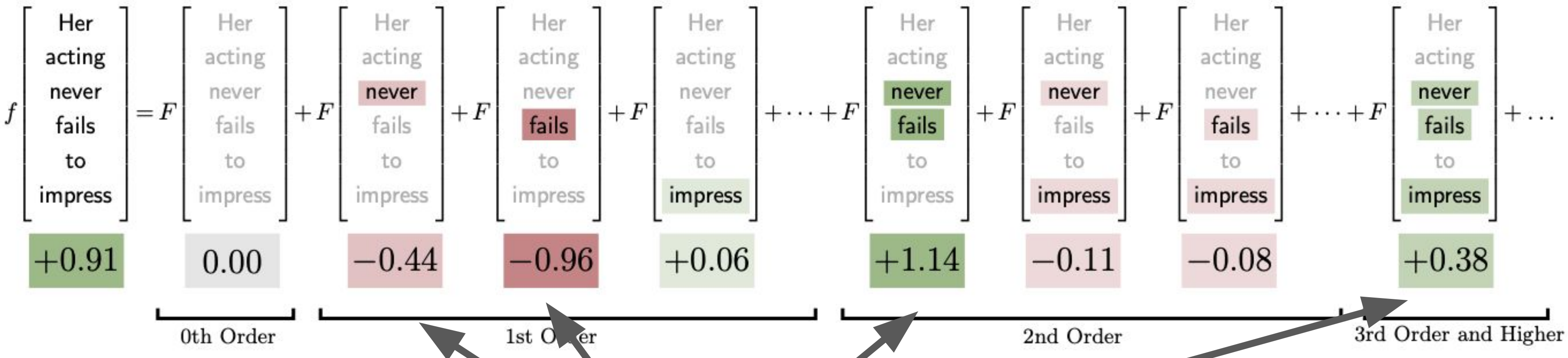
Signal Processing Approach for Explanations!

- Decompose the function in terms of effects of sets of inputs: polynomial

$$f(\mathbf{m}) = -0.93$$



“Signal” Model - Structure of Explainable Representations



1. Only a small number of coefficients are large

2. Big terms are generally low-order

Theorems

Theorem 1. (Noiseless Decoding) When there are K non-zero interactions chosen uniformly at random from all 2^n interaction, with $K = O(2^{n\delta})$ for $\delta < 1/3$ our algorithm exactly computes the Mobius transform:

- with sample complexity $O(Kn)$ and
- with time $O(Kn^2)$

with probability $1 - O(1/K)$.

Theorem 2. (Robust Low-Decoding, Informal) When there are K non-zero interactions chosen uniformly over all $|\mathbf{k}| \leq t$, with $t = \Theta(n^\alpha)$, $\alpha \leq 0.407$ our algorithm computes the Mobius transform:

- with sample complexity $O(Kt\log(n))$ and
- with time $O(K\text{poly}(n))$

with probability $1 - O(1/K)$ with any fixed SNR.

Step 1 - Subsampling for Optimal Aliasing/Hashing

- **Inescapable fact** of signal processing (Nyquist Sampling Theorem):

Harry Nyquist

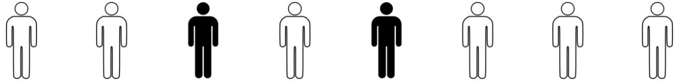
Subsampling causes aliasing

Embrace and understand the aliasing!

$$U(\mathbf{j}) = \sum_{\mathbf{Hk}=\mathbf{j}} F(\mathbf{k})$$

Step 2: Group Testing

- Originally proposed by Dorfman (1940s)
- Finds efficient ways to test soldiers for syphilis
- Pooling test allows you to identify infected individuals with fewer tests



								Outcome
1	1	1	1	0	0	0	0	Positive
0	0	0	0	1	1	1	1	Positive
1	1	0	0	0	0	0	0	Negative
0	0	1	0	0	0	0	0	Positive
0	0	1	0	1	1	0	0	Positive
0	0	0	0	1	0	0	0	Positive

t infected individuals

n total individuals

Step 2: Group Testing

- Originally proposed by Dorfman (1940s)
- Finds efficient ways to test soldiers for syphilis
- Pooling test allows you to identify infected individuals with fewer tests

$\mathbf{k}_1 =$	Her	acting	never	fails	to	impress	$\mathbf{j}$
	0	0	0	1	1	1	0
$\mathbf{H} =$	0	1	1	0	0	1	1
	1	0	1	0	1	0	1

t important words

n total words

Step 3: Message Passing - (Peeling Decoder)

Non-zero Interactions

$$\begin{bmatrix} \text{Her} \\ \text{acting} \\ \text{never} \\ \text{fails} \\ \text{to} \\ \text{impress} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \mathbf{k}_1 \qquad \begin{bmatrix} \text{Her} \\ \text{acting} \\ \text{never} \\ \text{fails} \\ \text{to} \\ \text{impress} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \mathbf{k}_2$$

$$\begin{bmatrix} \text{Her} \\ \text{acting} \\ \text{never} \\ \text{fails} \\ \text{to} \\ \text{impress} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \mathbf{k}_3 \qquad \begin{bmatrix} \text{Her} \\ \text{acting} \\ \text{never} \\ \text{fails} \\ \text{to} \\ \text{impress} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \mathbf{k}_4$$

Transform

Aliasing 1

$$\begin{aligned} u_1(00) &= f(110011) \\ u_1(01) &= f(110111) \\ u_1(10) &= f(111011) \\ u_1(11) &= f(111111) \end{aligned} \longrightarrow \begin{aligned} U_1(00) &= 0 \\ U_1(01) &= F(\mathbf{k}_3) \\ U_1(10) &= F(\mathbf{k}_1) \\ U_1(11) &= F(\mathbf{k}_2) + F(\mathbf{k}_4) \end{aligned}$$

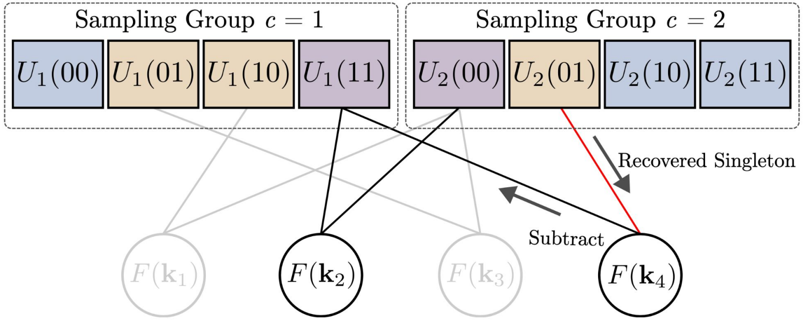
Zeroton

Singleton

Multiton

Aliasing 2

$$\begin{aligned} u_2(00) &= f(111100) \\ u_2(01) &= f(111101) \\ u_2(10) &= f(111110) \\ u_2(11) &= f(111111) \end{aligned} \longrightarrow \begin{aligned} U_2(00) &= F(\mathbf{k}_1) + F(\mathbf{k}_2) + F(\mathbf{k}_3) \\ U_2(01) &= F(\mathbf{k}_4) \\ U_2(10) &= 0 \\ U_2(11) &= 0 \end{aligned}$$



Conclusion

- Explaining deep models can be cast as **functional decomposition**
- **Aliasing, Group Testing** and **Message Passing** play a central role.

- Lots of open problems:
  - How do we improve robustness in real-world models?
  - Can we leverage white-box access to the model?
  - Can we exploit the connection between attention and Mobius transform?