Neural Network Attributions: A Causal Perspective

Aditya Chattopadhyay, Vineeth N Balasubramanian

Visual Learning and Intelligence Lab, Indian Institute of Technology Hyderabad, Near NH-65, Sangareddy, Kandi, Telangana 502285

IIT Hyderabad
Indian Institute of Technology Hyderabad

Introduction

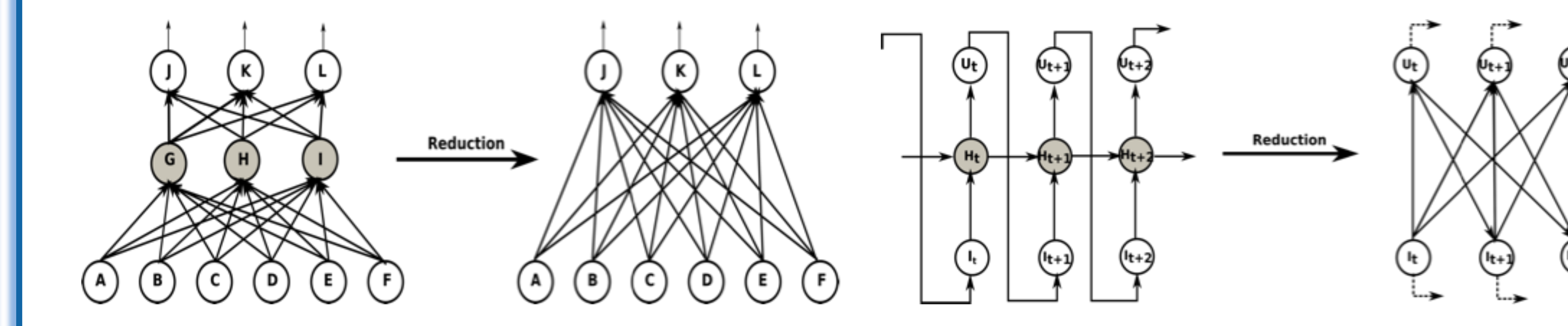
- Over the last decade, deep learning models have been highly successful in solving complex problems. However, the real bottleneck in accepting most of these techniques for real-life applications is the "interpretability problem".
- Over the years, three broad approaches towards "Explainable AI" have started to emerge: (i) optimization-based methods [Yosinski *et al.* 2015]; (ii) attribution-based methods [Sundararajan *et al.*, ICML 2017]; and (iii) supplanting black-box models with more interpretable learning machines [frost *et al.* 2017].
- In this work, we focus on "attribution-based methods". Harnessing theories from causal inference, we show that it is possible to obtain a global picture of a neural network's decision-making process along with local justifications. Our main contributions include:
  i. An interpretation of neural network architectures in terms of Structural Causal Models (SCMs).
  ii. Proposing a method to efficiently calculate interventional expectations, causal attributions and subsequently the causal effect of input neurons on the output.
  iii. Learning causal regressors to explain neural networks globally.
  iv. A discussion about the inherent biases prevalent in all current attribution-based methods.
  v. Experimental results exhibiting the efficacy of our proposed method.

Preliminaries

- **Definition 2.1 (Structural Causal Models)**. A Structural Causal Model (SCM) is a 4-tuple X, U, f, P_u where,
  i. X is a finite set of endogenous variables, usually the observable random variables in the system;
  ii. U is a finite set of exogenous variables, usually treated as unobserved or noise variables;
  iii. f is a set of functions $[f_1, f_2, \dots . f_n]$, where n refers to the cardinality of the set X. These functions define causal mechanisms, such that $\forall x_i \in X, x_i = f_i(Par, u\_i)$. The set Par is a subset of $X - \{x_i\}$ and $\forall u_i \in U$. We do not consider feedback causal models here;
  iv. P_u defines a probability distribution over U.
- An SCM $M(X, U, f, P_u)$ can be trivially represented by a directed graphical model $G = (V, U)$ where the vertices V represent the endogenous variables X (each vertex v_i corresponds to an observable x_i). The edges E denote the causal mechanisms f. Such a graph is called a **causal Bayesian network**. The distribution of every vertex in a causal Bayesian network depends only upon its parent vertices (local Markov property).
- **Proposition 1.** Two random variables a and b are said to be conditionally independent given a set of random variables Z if they are *d-separated* in the corresponding graphical model G.
- **Definition 2.2 (d-separation)**. Two vertices v_a and v_b are said to be *d-separated* if all paths connecting the two vertices are "blocked" by a set of random variables Z.
- A path is said to be "blocked" if either (i) there exists a *collider* that is not in $Anc(Z)$, or, (ii) there exists a *non-collider* $v \in Z$ along the path. $Anc(Z)$ is the set of all vertices which exhibit a *directed path* to any vertex $v \in Z$. A *directed path* from vertex v_i to v_j is a path such that there is no incoming edge to v_i and no outgoing edge from v_j.

Neural Networks as SCMs

- **Proposition 2**. An $l - layer$ feedforward neural network $N(l_1, l_2, \dots l_n)$ with l_i denoting the set of neurons in layer i has a corresponding SCM $M(l_1 + l_2 + \dots . + l_n, U, f_1 + f_2 + \cdots f_n, P_u)$, where l_1 refers to the input layer and l_n refers to the output layer. Corresponding to every l_i, f_i refers to the set of causal functions for neurons in layer i.
- **Corollary 2.1**. Every $l - layer$ feedforward neural network $N(l_1, l_2, \dots l_n)$, with l_i denoting the set of neurons in layer i, with a corresponding SCM $M(l_1 + l_2 + \dots . + l_n, U, f_1 + f_2 + \cdots f_n, P_u)$, can be reduced to an SCM $M'(l_1 + l_n, U, f', P_u)$ by marginalizing out the hidden neurons.



Neural Interpretability via Causal Effects

- This work tries to address the question: "What happens to an output value when one of the input features is changed by an external agent (the user)?" or more generally "What is the causal effect of a particular input neuron on a particular output neuron of the network?".
- Given a neural network with l_1 being the set of input features and l_n being the set of output features, we measure the Average Causal Effect (ACE) of an input feature $x_i \in l_1$ with value α on an output feature $y \in l_n$ as:

$$ACE^{y}_{do(x_i)=\alpha} = E(y|do(x_i) = \alpha) - baseline_{x_i}$$

- In this work, we propose the average ACE of x_i on y as the baseline value for x_i, i.e. $baseline_{x_i} = E_{x_i}(E_y(y|do(x_i) = \alpha)$. In absence of any prior information, we can assume that the "doer" is equally likely to perturb x_i uniformly in its range.

Calculating Interventional Expectations

- Given a neural network, the output neuron y can be expressed as the causal mechanism $f'_y(x_1, x_2, \dots, x_k)$, where x_i refers to neuron i in the input layer. Considering a quadratic approximation around the interventional means,

$$E(f'_{y|do(x_i)=\alpha}(l_1 - \mu + \mu)|x_i = \alpha) = f'_y(\mu_1, \mu_2, \dots, \mu_k) + Tr(\nabla^2 f'_y(\mu_1, \mu_2, \dots, \mu_k). E((l_1 - \mu)(l_1 - \mu)^T | do(x_i) = \alpha)).$$

- **Proposition 3.** Given an *l*-layer feed forward neural network $N(l_1, l_2, \dots l_n)$ with) with l_i denoting the set of neurons in layer i and its corresponding reduced SCM) with l_i denoting the set of neurons in layer i, the intervened input neuron is d-separated from all other input neurons.
- **Corollary 3.1.** Given an *l*-layer feedforward neural network $N(l_1, l_2, \dots l_n)$ with l_i denoting the set of neurons in layer i and an intervention on neuron x_i, the probability distribution of all other input neurons does not change, i.e. $\forall v_j \in V \text{ and } v_i \neq v_j \; P(v_j|do(x_i) = \alpha) = P(v_j)$.
- Note that here we have assumed causal independency between different input neurons of a feed forward network. This is violated in time-series models or sequence prediction tasks, in that case we have to iterate over the entire training data for every intervention.
- **Proposition 4.** Given a recurrent neural function, unfolded in the temporal dimension, the output at time t will only be dependent on inputs from timesteps t to $t - \tau$, where τ is given as $E_x(argmax_k(|\det(\nabla_{x^{t-k}} y^t)| > 0))$.

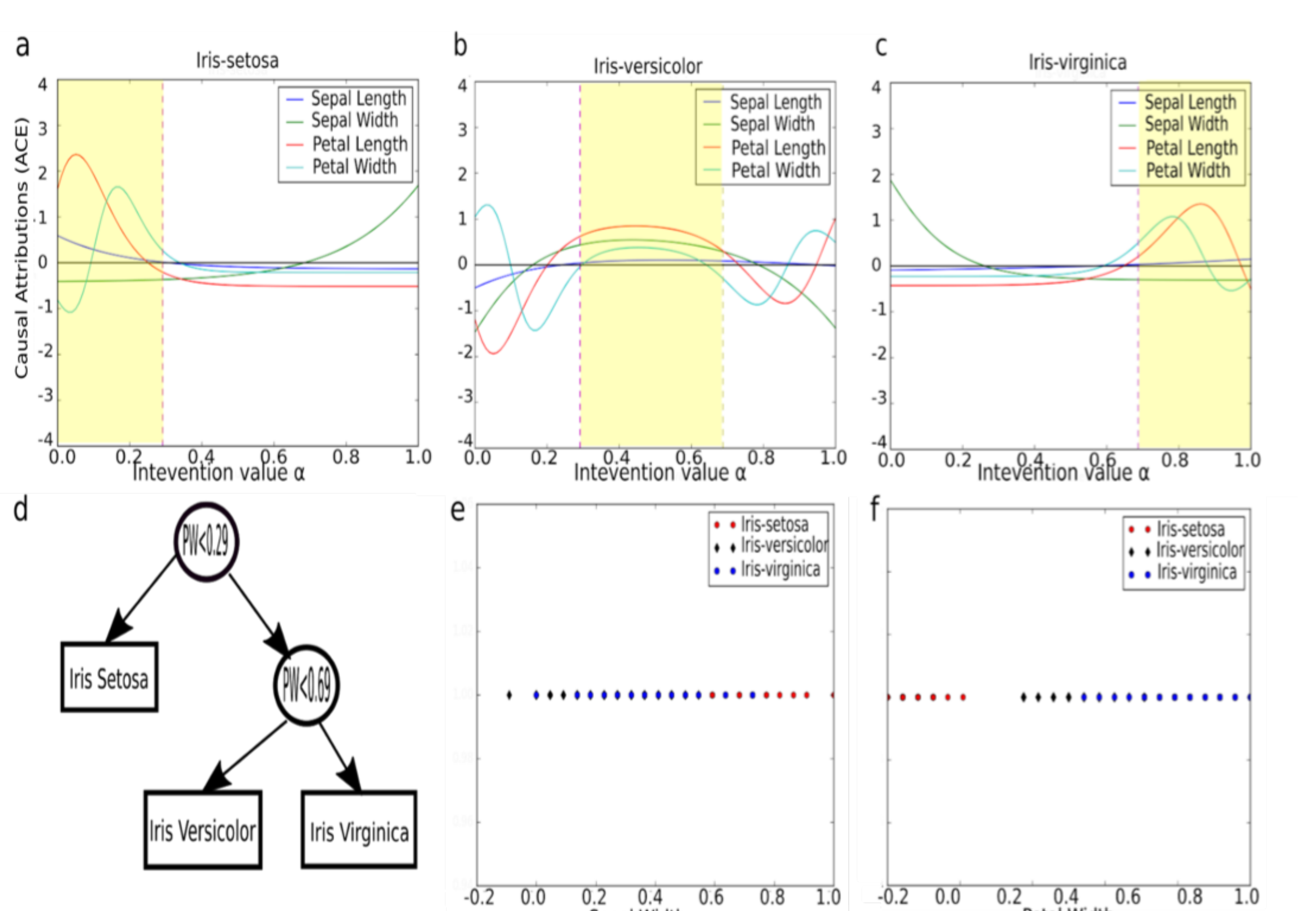
Causal Regressors

- The interventional expectation $E(y|do(x_i) = \alpha$ will only be a function of x_i as all the other variables have been marginalized out.
- We assume this function to be a member of the polynomial class of functions $\{f \mid f(x_i) = \sum_j^{order} w_j x_i^j\}$. Bayesian model selection was employed to determine the optimal order of the polynomial that best fits the given data by maximizing the marginal likelihood.
- Calculating interventional expectations for multiple input values is a costly operation. Learning the function, termed as **causal regressors**, allows one to estimate these values on-the-fly for subsequent attribution analysis.
- Furthermore, inspecting the nature of these causal regressors can give valuable insights into the global workings of the neural network.

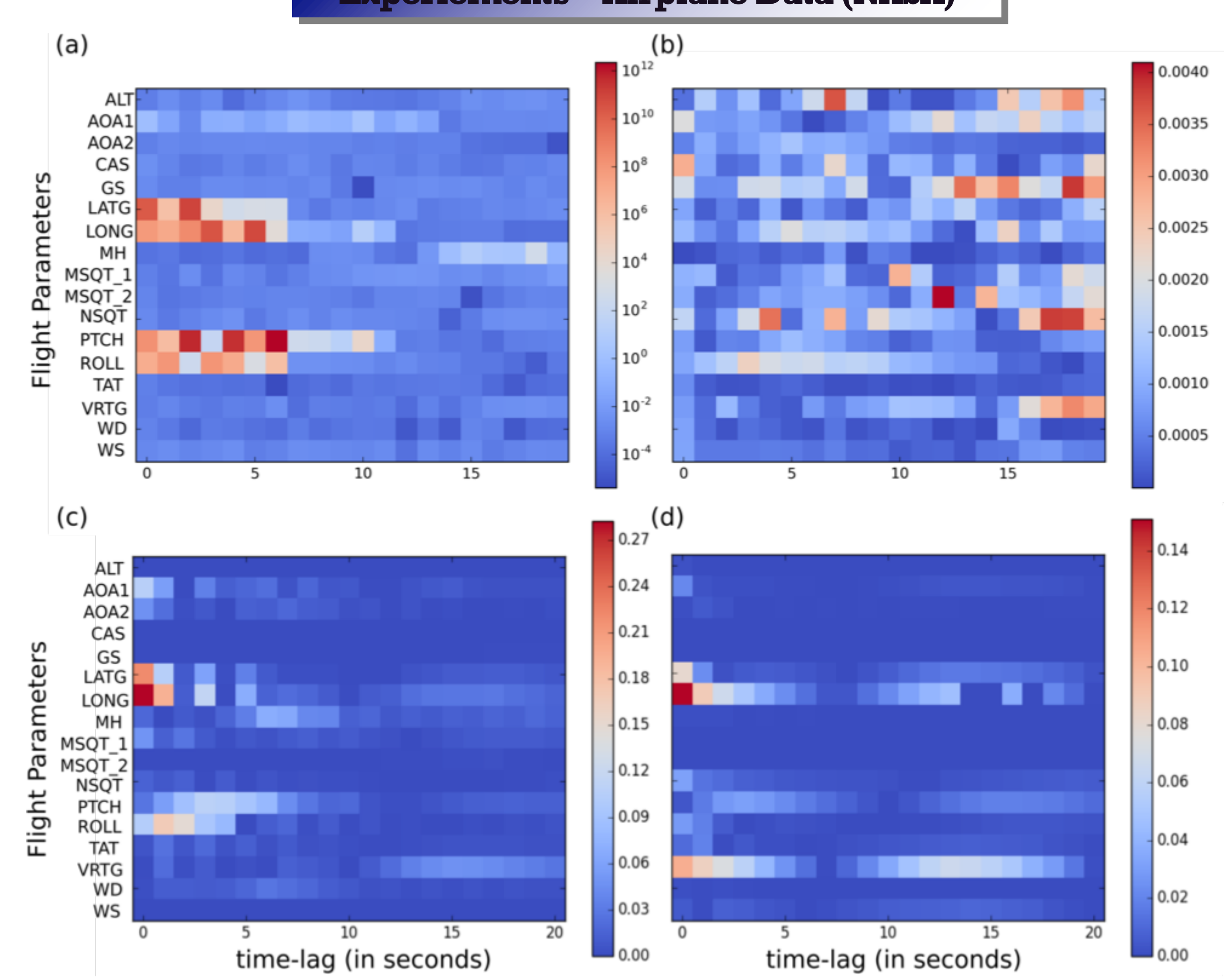
Previous Work

- Attribution methods are concerned with unravelling the importance of a particular input feature on the output of a network. Initial attempts involved perturbing regions of the input via occlusion maps or inspecting the gradients of an output neuron with respect to an input neuron.
- However, the unidentifiability of "source of error" is a central impediment to designing attribution algorithms for black-box deep models. It is impossible to distinguish whether an erroneous heatmap (given our domain knowledge) is an artifact of the attribution method or a consequence of poor representations learnt by the network. This resulted in development of newer methods guided by certain axioms: (i) Conservative (ii) Sensitivity, (iii) Implementation Invariance, (iv) Symmetry preserving, and (v) Input Invariance. Despite these axioms, the proposed methods are not really causal in nature.
- For example, consider the integrated gradients method. While this method satisfies the axioms, there exists an implicit bias in the attribution values (variable importance) obtained. Consider the function $f(a, b) = a.b$, and two input vectors $i_1 = [3,5]$ and $i_2 = [3,100]$.Integrated gradients assign attributions to $[a, b]$ as $[3.4985, 7.4985]$ for input i_1 and $[50.951, 244.951]$ for input i_2.
- This is an implicit bias which occurs because of not marginalizing other input variables while computing the attribution of a. Most current attribution methods are based on the gradient and suffer from this bias.

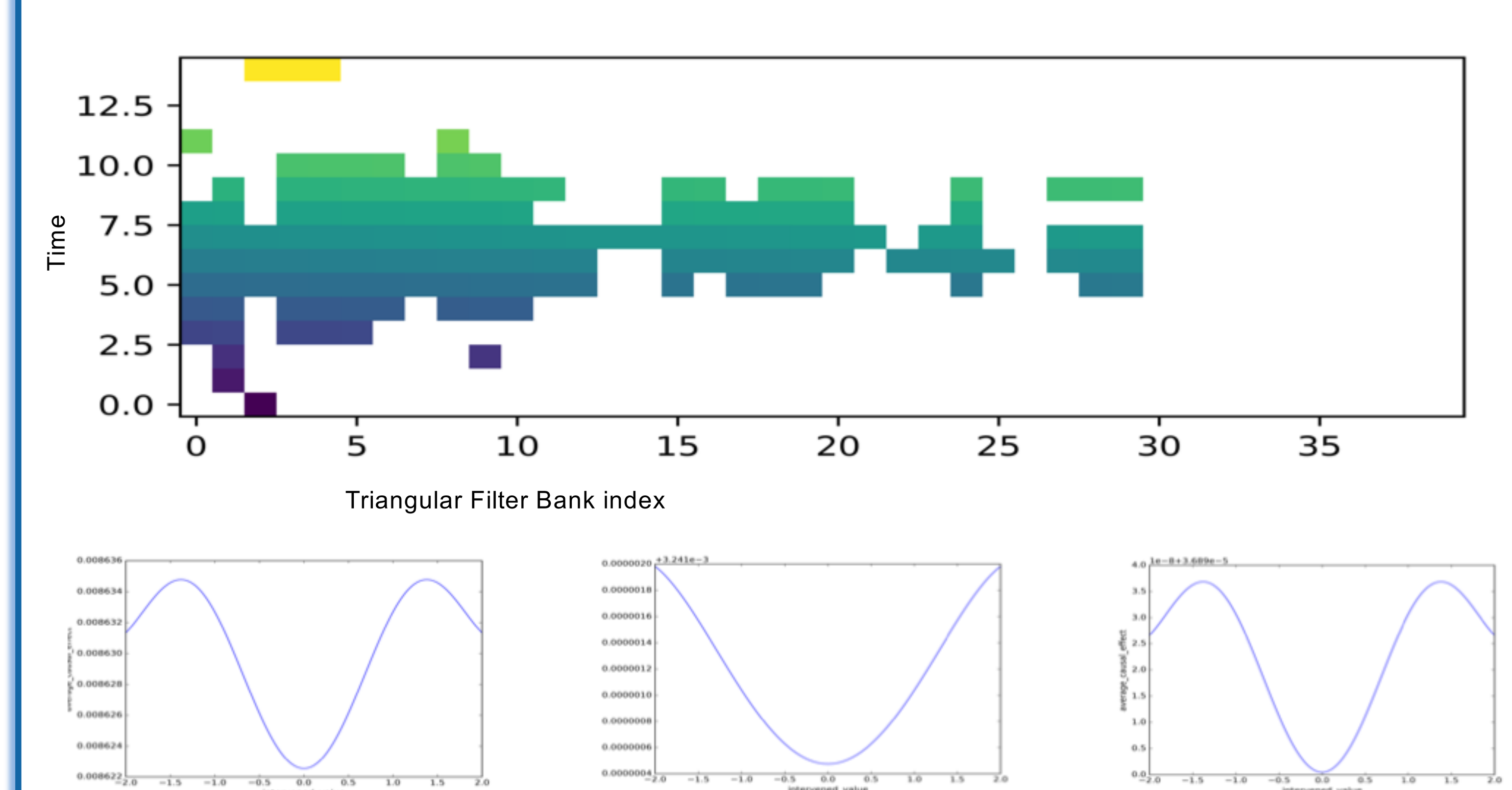
Experiment – Toy dataset



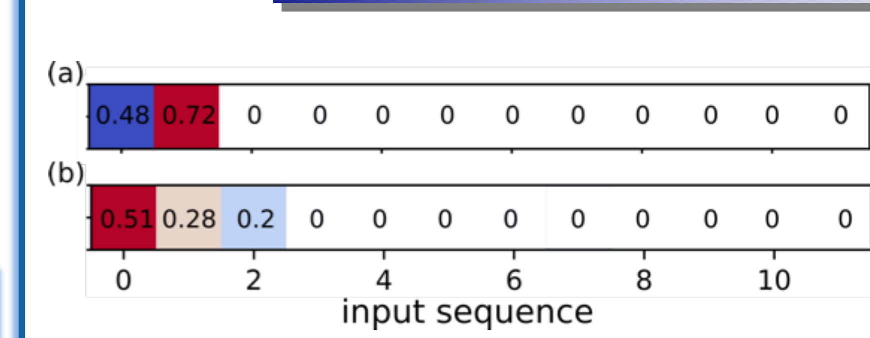
Experiements – Airplane Data (NASA)



Experiements – Speech Data (TIMIT)



Experiement – Simulated dataset



Occluded Feature	Test Error
2 ($x^2 \sim N(0,0.2)$)	0.01059
1 ($x^1 \sim N(0,0.2)$)	0.01072
0 ($x^0 \sim N(0,0.2)$)	0.01268
None *(Baseline)*	0.01059

References

1. Pearl, Judea. *Causality*. Cambridge university press, 2009.
2. Sundararajan, Mukund *et al.* "Axiomatic attribution for deep networks." *arXiv preprint arXiv:1703.01365*(2017).
3. Yosinski, Jason, et al. "Understanding neural networks through deep visualization." *arXiv preprint arXiv:1506.06579* (2015).

Acknowledgements

We thank Honeywell Aerospace for helping with the aircraft data and providing funding. We are also grateful to Anirban Sarkar and Kush Motwani for fruitful discussions.