CoFrNets: Interpretable Neural Architecture Inspired by Continued Fractions

Researchers
Isha Puri
Amit Dhurandhar
Tejaswini Pedapati
Karthikeyan Shanmugam
Dennis Wei
Kush Varshney

18 October 2021

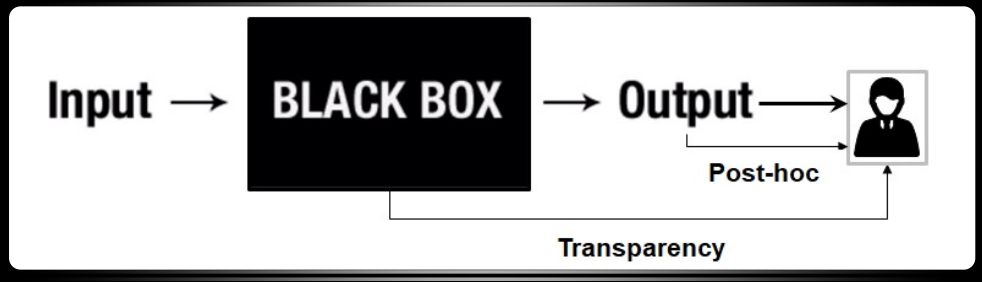
Research

IBM

Background: Problem Solved

Much research has been done on local post-hoc explanations for neural networks

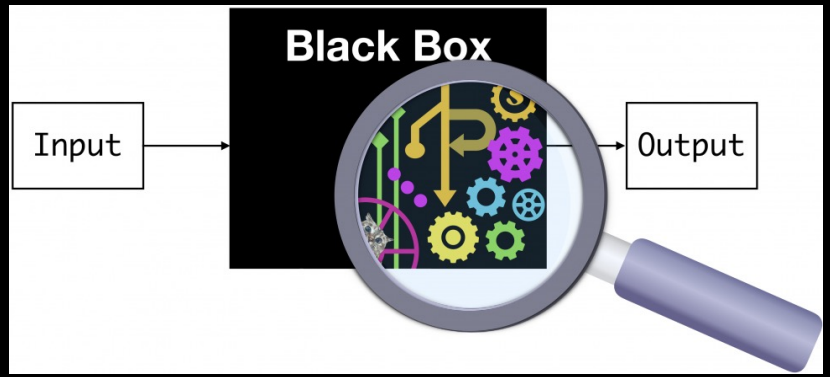
Despite this, architectures that are inherently interpretable are less common

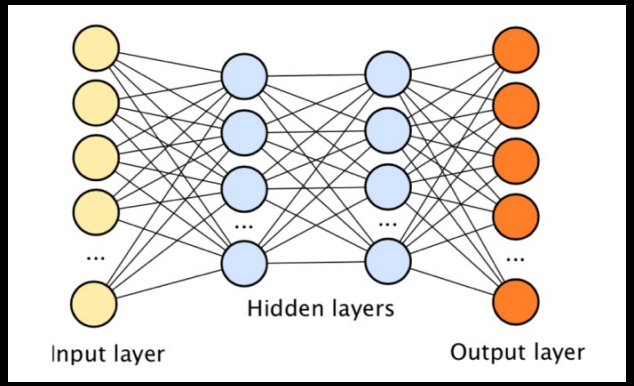


Background: Prior Approaches

1. Black Box Neural Architectures and Variations

- MLP (Multi Layer Perceptron) – Standard neural architecture typically composed of fully connected network
- CNN (Convolutional Neural Network)
- ResNet (Residual Network) – skip connections
- Others: DenseNet, MobileNetV2, Π-nets, etc.

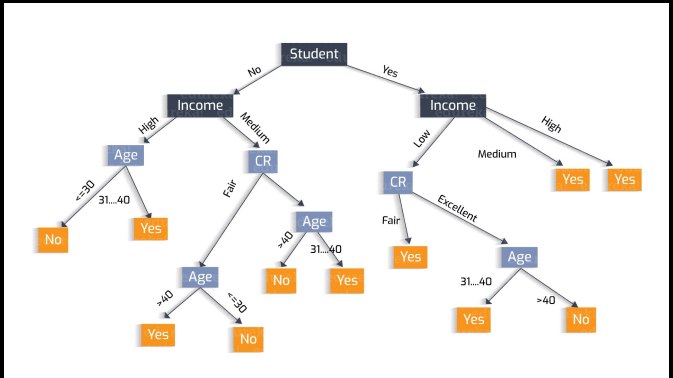


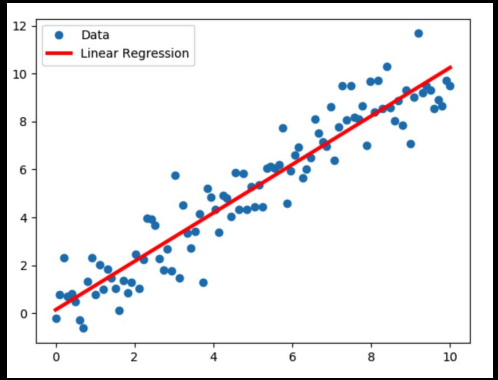


Background: Prior Approaches

2. Globally Interpretable Models

- Standard ML methods such as (1) logistic regression, (2) decision trees, (3) rule based models, etc.
- GAMs (Generalized Additive Models), NAMs (Neural Additive Models), EBMs (Explainable Boosted Machines)
- LassoNet (could be considered interpretable, but has restrictions)





Background: Prior Approaches

3. Local to Global Post-Hoc Methods

- Take local explanations and create global ones
- TreeSHAP – creates global SHAP explanation for tree-based models
- MAME (Model Agnostic MultiLevel Explanation) creates global LIME explanations

Background: Prior Approaches

4. Self Explaining Models

- Not globally interpretable, but provides local explanations without post-hoc mechanisms
- Frameworks might provide explanation for models given previously defined explanations for training

Background: Prior Approaches

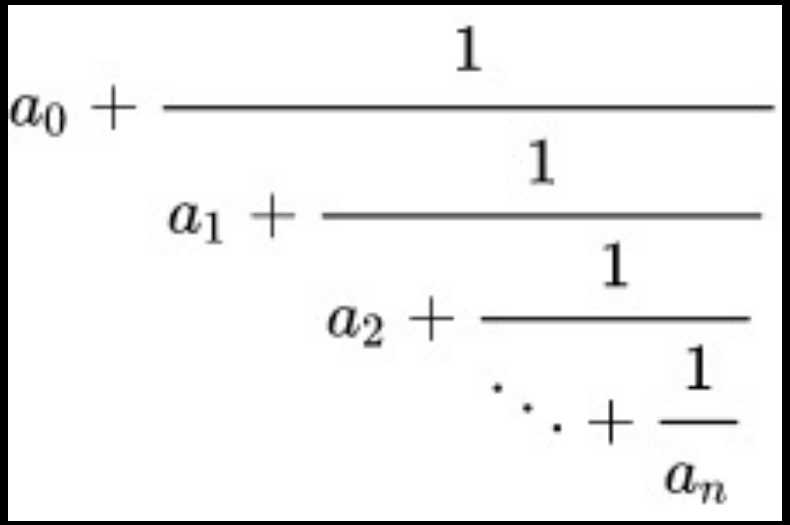
1. Black Box Neural Architectures and Variations
2. Globally Interpretable Models
3. Local to Global Post-Hoc Methods
4. Self Explaining Models

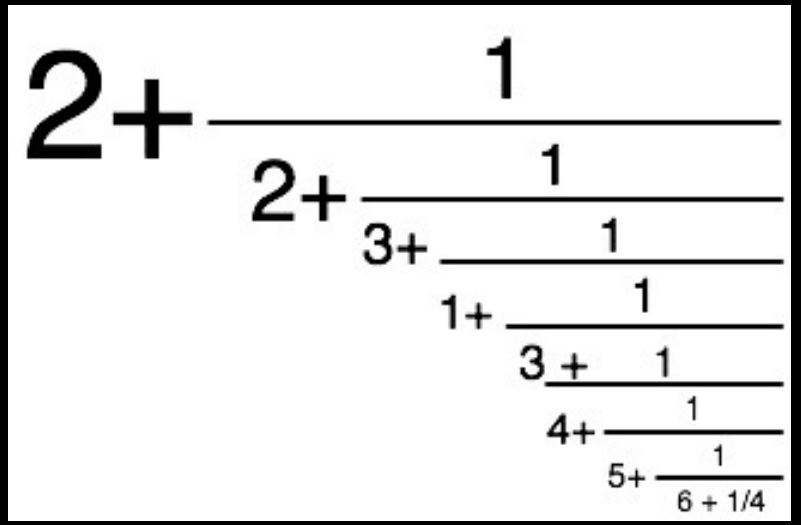
Background: What Are continued fractions?

Continued fractions are just fractions made of fractions.

Background: What Are continued fractions?

Continued fractions are just fractions made of fractions.





“ladder”

Background: What Are continued fractions?

Any number – rational or irrational – can be represented as a continued fraction.

$$\pi = \cfrac{4}{1+\cfrac{1^2}{3+\cfrac{2^2}{5+\cfrac{3^2}{7+\cfrac{4^2}{9+\cdots}}}}}$$

$$\phi = \frac{1+\sqrt{5}}{2} = 1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cdots}}}$$

$$e = 2+\cfrac{1}{1+\cfrac{1}{2+\cfrac{2}{3+\cfrac{3}{\ddots}}}}$$

Background: What Are continued fractions?

Continued Fractions can represent any number or analytic function.

$$\sum_{n=0}^{\infty} B_n x^n = \cfrac{1}{1 + \cfrac{x}{\frac{2}{1} - x + \cfrac{2x}{3 + \cfrac{2x}{\frac{2}{2} - 2x + \cfrac{3x}{5 + \cfrac{3x}{\frac{2}{3} - 3x + \cfrac{4x}{7 + \cfrac{4x}{\frac{2}{4} - 4x + \cfrac{5x}{9 + \ddots}}}}}}}}}$$

$$\tan(z) = \cfrac{z}{1 - \cfrac{z^2}{3 - \cfrac{z^2}{5 - \cfrac{z^2}{7 - \ddots}}}}.$$

$$e^z = \cfrac{1}{1 - \cfrac{z}{1 + z - \cfrac{\frac{1}{2}z}{1 + \frac{1}{2}z - \cfrac{\frac{1}{3}z}{1 + \frac{1}{3}z - \cfrac{\frac{1}{4}z}{1 + \frac{1}{4}z - \ddots}}}}}.$$

Background: What Are continued fractions?

In representing any real number with natural number parameters a_k and b_k, the "truncations" ("convergent") of continued fractions are the best possible rational approximation.

$$a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_{2\,+\,\ldots}}}$$

Background: What Are continued fractions?

- Represented as "ladder-like" sequence: $a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + \ldots}}$

- can represent any real number, analytic function, including:
  trigonometric functions, polynomials, exponential functions, power function, special functions (gamma, hypergeometric, Bessel functions)

- best rational approximation of numbers and functions

- Fast convergence of approximations to real numbers

What is a CoFrNet?

Continued Fraction Net

Based on the structure of a continued fraction

$$a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + \dots}}$$

Aims to embody many of CF's desirable properties – CFs can represent any real number and any analytic function (trig, polynomial, exponential, etc.)

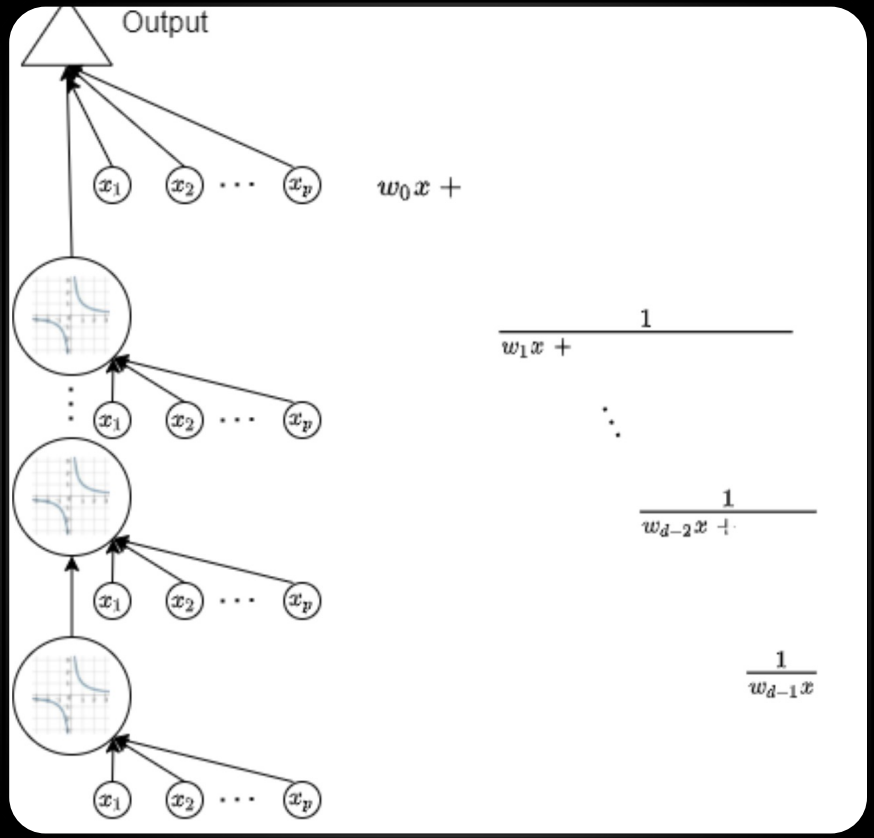
What is a CoFrNet?

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_{2\,+\,\ldots}}}$$

proposed “ladder-like” architecture replaces a_ks in CFs with linear functions - the input x multiplied by a weight vector w_k in each layer k

Reciprocal $\frac{1}{x}$ of the function is the nonlinearity in each layer (differs from commonly used ReLU, sigmoid, etc).

We show that linear functions are sufficient for universal approximation with a finite number of ladders

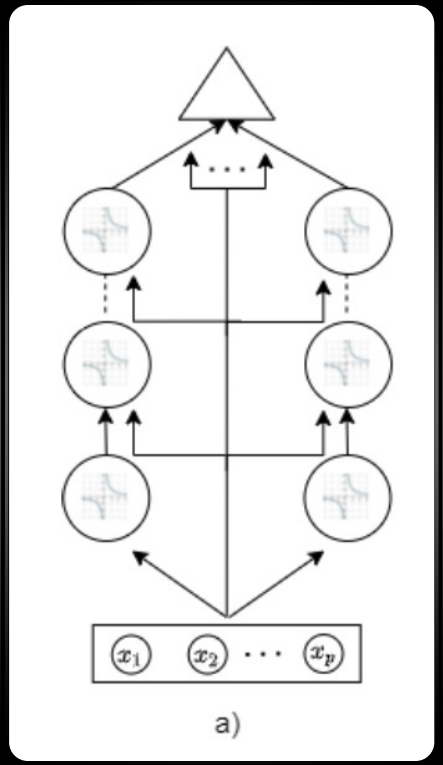


What is a CoFrNet?

We have three initial variations of this architecture that can serve as the basis of future experimentation

#1: Full-Fledged Variant, fully connected CoFrNet-F

– Each ladder receives whole input x at every stage



What is a CoFrNet?

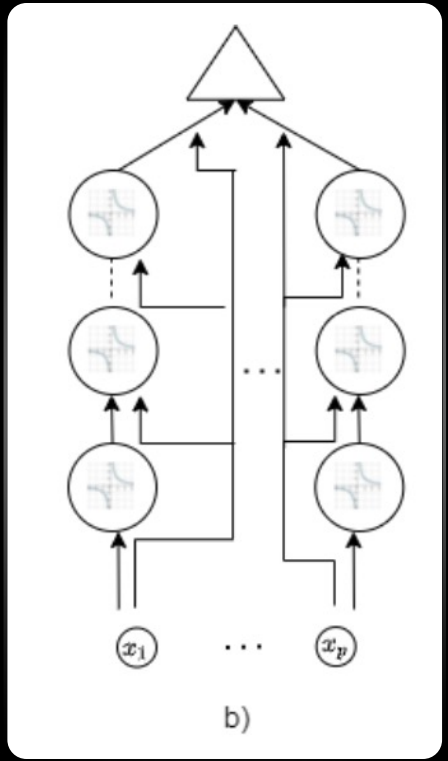
We have three initial variations of this architecture that can serve as the basis of future experimentation

#2: Diagonalized Variant - CoFrNet-D

– Each ladder receives only one of the input dimensions

– Additive model

– Each ladder can serve as representation for individual feature



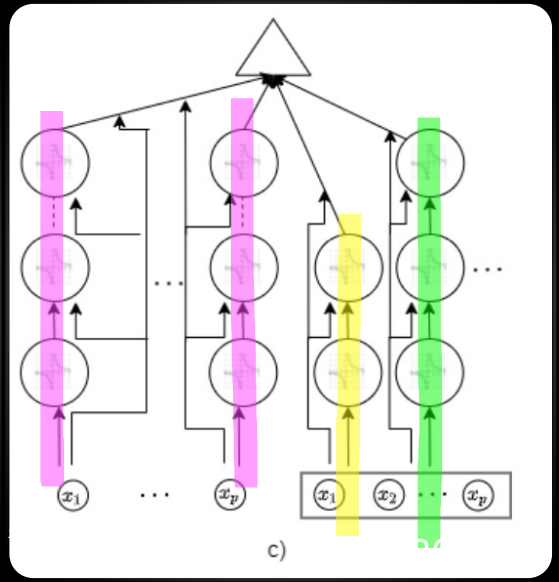
What is a CoFrNet?

We have three initial variations of this architecture that can serve as the basis of future experimentation

#3: Combination of Diagonalized and Fully Connected Variant - CoFrNet-DL

– Full ladders are of increasing depth

– Full ladders can each capture importance of respective order of interactions



Novel Proof of Universal Approximation

- We prove that a linear combination of continued fractions (with finite depth and linear layers) has the property of universal approximation.
- Novel Proof Strategy: with three steps:
  - i) showing polynomials of linear functionals are a unital subalgebra and separating on the domain
  - ii) applying the Stone-Weierstrass theorem to show that they are thus dense in the space of bounded continuous functions
  - iii) showing that they are a subset of the aforementioned class of function

Key Insights:

- Even “shallow” ladders can represent complicated functions if you ensemble them.
- We can represent sparse polynomials compactly.

Interpretability advantages

Exciting interpretability advantages can be shown:

1) Using Continuants
2) Using Power Series

Both exploit functional form of continued fractions.

These strategies can globally interpret all three variants (including the full variant).

Interpretability advantages – using continuants (Self-Explaining)

- A function $f(x; w)$ can be expressed as a ratio of polynomials
- We derive a compact expression for the gradient of $f(x; w)$ with respect to the inputs $x_j = 1 \ldots p$

Proposition 1. *The partial derivative of $f(x; w)$ with respect to x_j is given by*

$$\frac{\partial f(x;w)}{\partial x_j} = \sum_{k=0}^{d} (-1)^k \left(\frac{K_{d-k}(a_{k+1}, \ldots, a_d)}{K_d(a_1, \ldots, a_d)} \right)^2 w_{jk}.$$

- We show that computing the gradient of a ladder with respect to its inputs is useful for determining scores of feature importance
- This method is example based, gives only first order attributions

Interpretability advantages – Using power series (Global Interpretations)

- To determine both high order and first order interaction scores, we represent our ladder as a multivariate power series
  - Also supports a linear combination of ladders (for input size > 1)
- These coefficients provide feature importance scores for both individual features and higher order interactions , 1-to-1 mapping between CF terms and PS coefficients
- Original Continued Fraction of X33 “Ladder” in Waveform Dataset:

$$0.09536925*x+0+\cfrac{1}{-0.000634073*x+0+\cfrac{1}{0.052001227*x+0.20613371+\cfrac{1}{-0.0006788851*x+0+\cfrac{1}{0.058380947*x+0.082712844}}}}$$

- Power Series:

$$0.106349-0.00346107x+0.00007736x^2-1.97998*10^{-6}*x^3+4.85234*10^{-8}*x^4-1.2074*10^{-9}x^5+2.98914*10^{-11}*x^6$$

Done via Mathematica Here

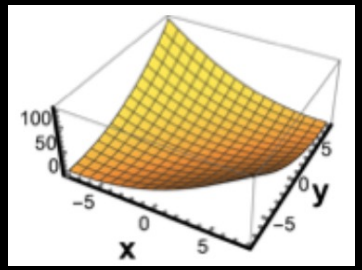
Results on synthetic datasets

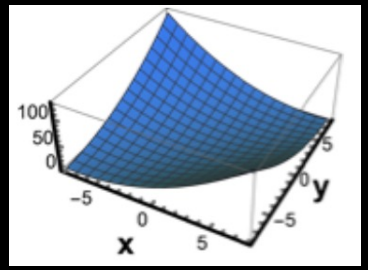
We used CoFrNet-F variant to: 1) model different synthetic functions and 2) compute feature attributions using the continuants and power series strategies

Results on synthetic datasets

We used CoFrNet-F variant to: 1) model different synthetic functions and 2) compute future attributions using the continuants and power series strategies

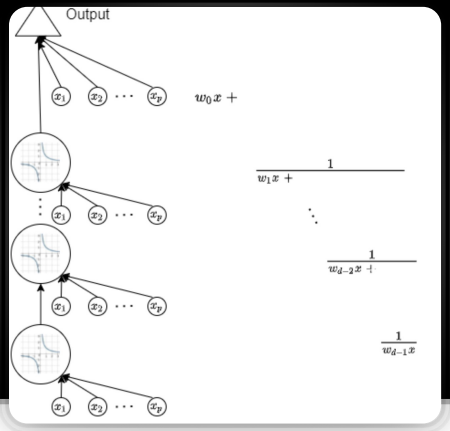
Matya's Function:





$OF: f(x,y) = .54x^2 + .54y^2 - xy$

$IPS: f(x,y) = .56x^2 + .44y^2 - xy$



Interpretation using Power Series (IPS):

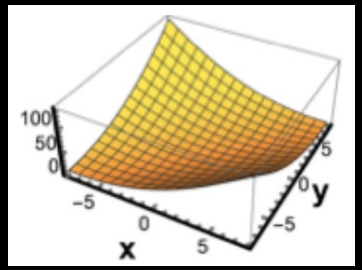
$$.4903 * \frac{1}{-1.06x + .967y + .554 + \frac{1}{-.88x + .78y - .954}} + .9594 \rightarrow. \qquad \rightarrow .567x^2 + .439y^2 - xy + (-.0103 - .071y - .06x)$$

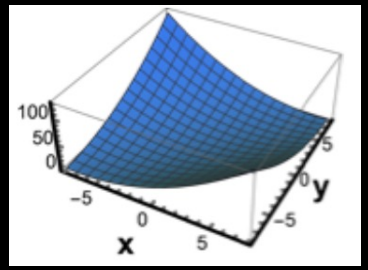
Expand, Normalized Taylor Series

Results on synthetic datasets

We used CoFrNet-F variant to: 1) model different synthetic functions and 2) compute future attributions using the continuants and power series strategies

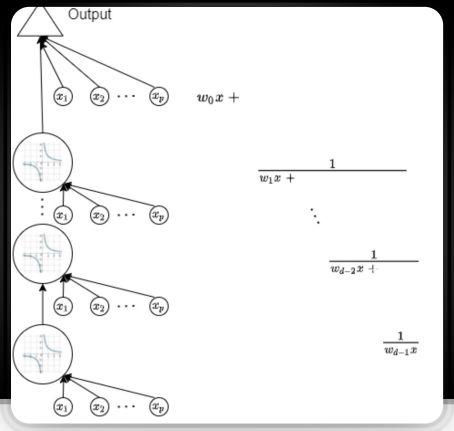
Matya's Function:





$OF\colon f(x,y) = .54x^2 + .54y^2 - xy$

$IPS\colon f(x,y) = .56x^2 + .44y^2 - xy$



Interpretation using Continuants (IC):

Proposition 1. *The partial derivative of* $f(x;w)$ *with respect to* x_j *is given by*

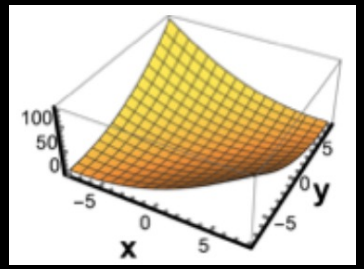
$$\frac{\partial f(x;w)}{\partial x_j} = \sum_{k=0}^{d} (-1)^k \left(\frac{K_{d-k}(a_{k+1}, \ldots, a_d)}{K_d(a_1, \ldots, a_d)} \right)^2 w_{jk}.$$

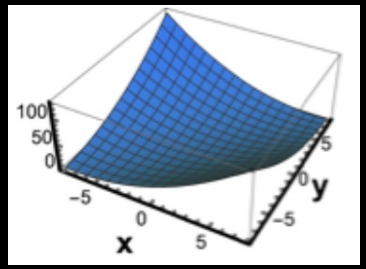
$IC\colon x \to 0.06, y \to -0.07$

Results on synthetic datasets

We used CoFrNet-F variant to: 1) model different synthetic functions and 2) compute future attributions using the continuants and power series strategies

Matya's Function:





$OF\colon f(x,y) = .54x^2 + .54y^2 - xy$

$IPS\colon f(x,y) = .56x^2 + .44y^2 - xy$

CoFrNet-F is able to accurately approximate Matya's with 7.31% error

Interpreting using continuants, we are able to calculate feature importance scores for x and y, which show they are of equal magnitude importance

Interpreting using power series, we are able to recover the function quite accurately

Results on Real Life datasets

We experimented with three modalities of datasets:

- Tabular (Credit Card, Magic, Waveform),
- Text (Sentiment Analysis and Quora Insincere Questions),
- Image (CIFAR10)

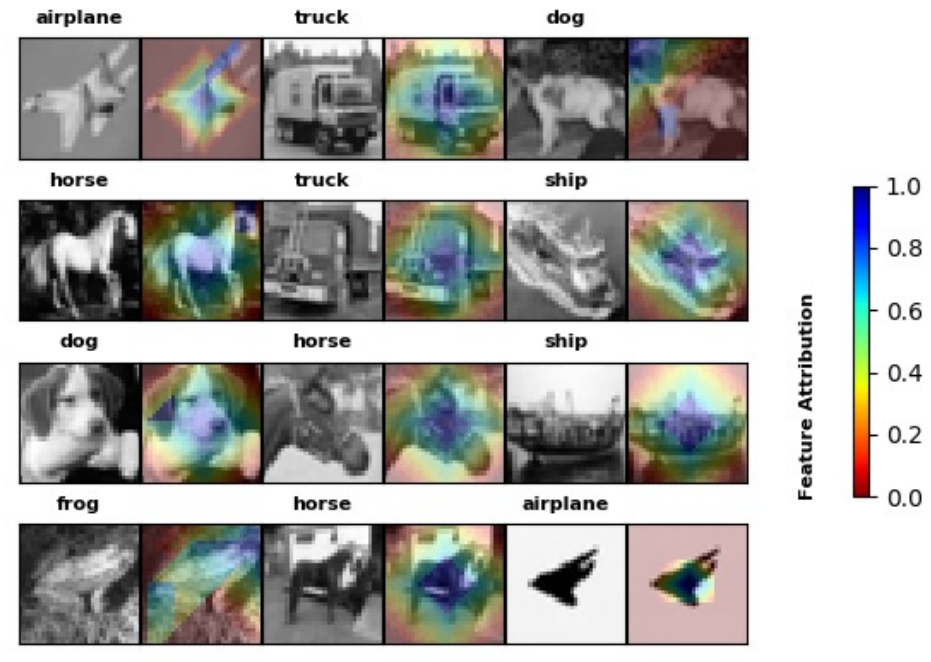
- Competitive with other interpretable models on tabular datasets, within 6% of SOTA black box models

- On Images/Text, we perform significantly better than interpretable models, and similar/better than MLP.

Methods	Interpretable	Waveform	Magic	Credit Card	CIFAR10	Sentiment	Quora
CoFrNet-DL	Yes	0.81	**0.84**	0.69	**0.87**	**0.84**	**0.88**
CoFrNet-D	Yes	0.69	0.76	0.66	0.38	0.80	0.75
GAM	Yes	**0.85**	**0.85**	**0.72**	DNC	0.51	DNC
NAM	Yes	**0.86**	0.81	0.69	0.38	0.50	0.49
EBM	Yes	**0.85**	**0.85**	**0.72**	0.40	0.59	0.49
CART	Yes	0.75	0.79	0.69	0.29	0.52	0.73
LassoNet	Yes	0.84	0.76	0.67	0.28	0.50	0.53
MLP	No	0.34	0.65	0.50	0.35	**0.83**	0.85
SOTA	No	0.86	0.86	0.75	0.99	0.96	0.94

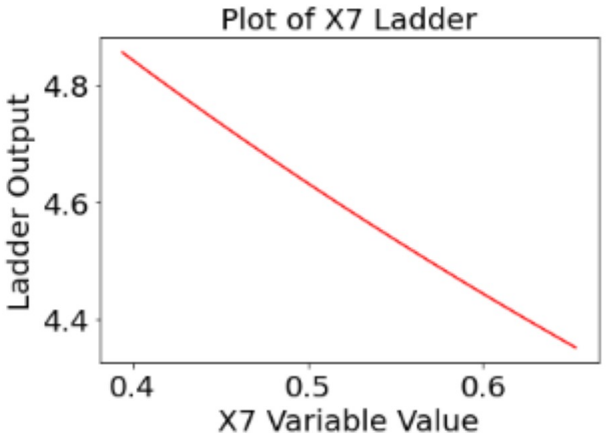
Shows that CoFrNets can compactly represent rich class of functions in high dimensional space where continued fractions converge quickly

Interpretability Results - Image

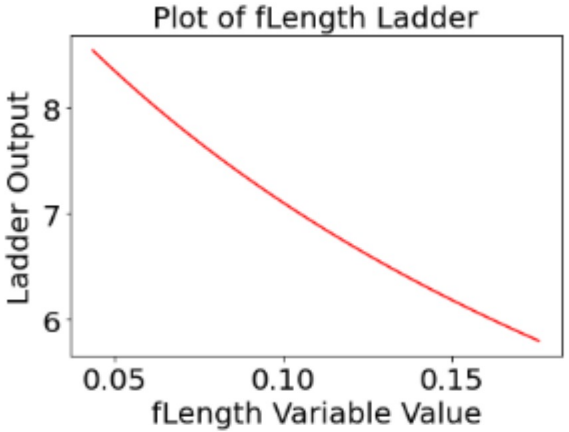


24 randomly chosen CIFAR10 test images (in grey scale) and to the immediate right of each their corresponding (normalized) attributions overlayed as a colormap over each of them using the IC strategy.

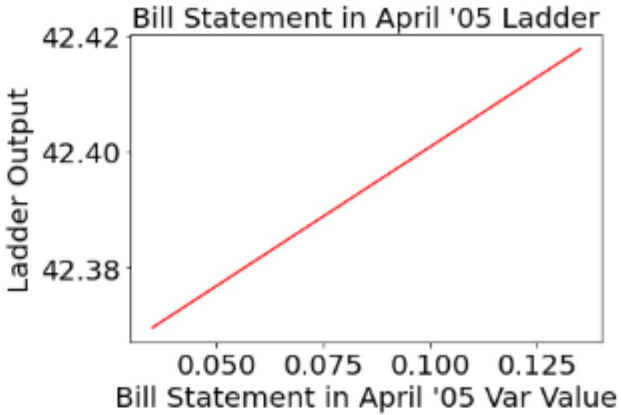
Interpretability Results - Tabular



(a) Waveform



(b) Magic



(c) Credit Card

Interpretability Results - Text

Word Clouds show the importance of different words in the IMDB Sentiment Analysis Dataset (left) and the Quora Insincere Questions Dataset (right).

Importances were determined by positions (indices) of words.

Summary of Advantages

- Creates architecture that empirically is either competitive or significantly better than other interpretable models and (sometimes) comparable even to the State of the Art
- Theoretically Proven 1) Interpretability and 2) Universal Approximation Ability
- Proposes novel architecture based /reliant on specific properties of Continued Fractions that can be expanded upon and experimented with
- Hypothesized favorable adversarial robustness properties
- Showed interpretability visualizations
- Novel Strategy to Prove Universal Approximation

Ongoing work

- There is a ton of very exciting work that we are continuing to pursue with this new architecture!
- We strongly believe that although we are already matching SOTA for many models, we can boost our accuracy with **new training strategies.**
- Proving Faster Inference times and exploring hardware implications
- Optical Computing for Energy Saving
- Real-Time inference may be possible for deep ladders (as same x is passed to every layer)
- Quantum Implementation in Qiskit
- Proving a Rademacher Complexity bound for CoFrNets

THANK YOU!