Survey of data generation and evaluation in Interpreting DNN pipelines

Presenter: Sanchit Sinha

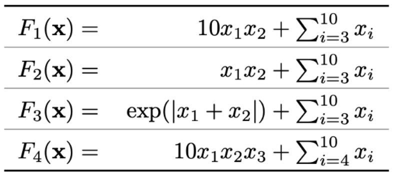
https://qdata.github.io/deep2Read/

24 April 2020

Approach-1a Generating data from probability distributions/functions

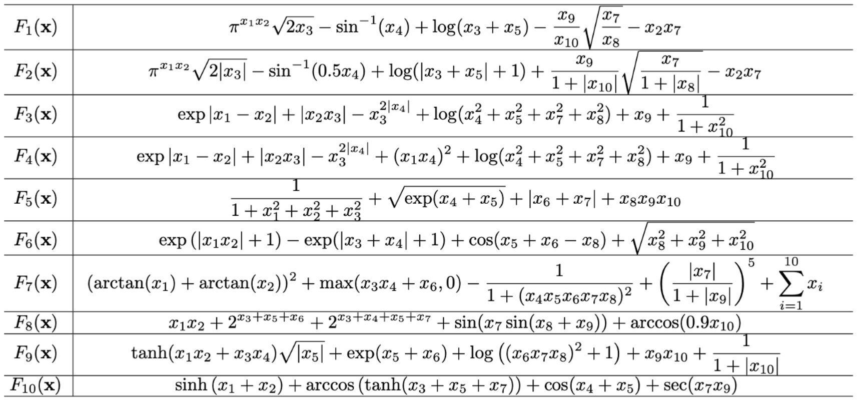
- We specifically model the data to find out what exactly we want to test
- “Can I trust you more? Model-agnostic Hierarchical Explanations” Tsang et al.
    - Objective is to study the interactions b/w the features.
    - Generated data from functions have interactions between the variables pre-defined in the generating functions
- Sorokina et al. Used a random function generator (**Hooker**)
- Detecting Statistical Interactions from Neural Network Weights, **Tsang** et al., ICLR 2018 again used functions to model interactions between features
- All these methods are modelling interactions between features

Table 2: Data generating functions with interactions



$F_1(\mathbf{x}) =$	$10x_1x_2 + \sum_{i=3}^{10} x_i$
$F_2(\mathbf{x}) =$	$x_1x_2 + \sum_{i=3}^{10} x_i$
$F_3(\mathbf{x}) =$	$\exp(\lvert x_1 + x_2\rvert) + \sum_{i=3}^{10} x_i$
$F_4(\mathbf{x}) =$	$10x_1x_2x_3 + \sum_{i=4}^{10} x_i$

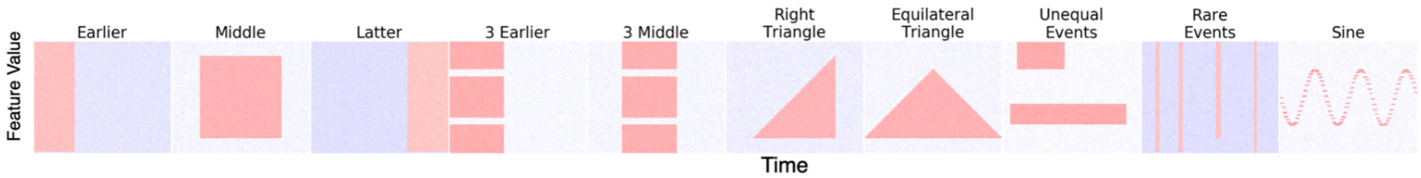
Table 1: Test suite of data-generating functions



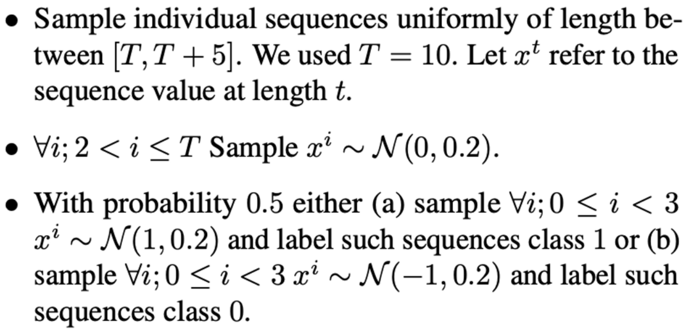
$F_1(\mathbf{x})$	$\pi^{x_1x_2}\sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3+x_5) - \frac{x_9}{x_{10}}\sqrt{\frac{x_7}{x_8}} - x_2x_7$
$F_2(\mathbf{x})$	$\pi^{x_1x_2}\sqrt{2\lvert x_3\rvert} - \sin^{-1}(0.5x_4) + \log(\lvert x_3+x_5\rvert+1) + \frac{x_9}{1+\lvert x_{10}\rvert}\sqrt{\frac{x_7}{1+\lvert x_8\rvert}} - x_2x_7$
$F_3(\mathbf{x})$	$\exp\lvert x_1-x_2\rvert + \lvert x_2x_3\rvert - x_3^{2\lvert x_4\rvert} + \log(x_4^2+x_5^2+x_7^2+x_8^2) + x_9 + \frac{1}{1+x_{10}^2}$
$F_4(\mathbf{x})$	$\exp\lvert x_1-x_2\rvert + \lvert x_2x_3\rvert - x_3^{2\lvert x_4\rvert} + (x_1x_4)^2 + \log(x_4^2+x_5^2+x_7^2+x_8^2) + x_9 + \frac{1}{1+x_{10}^2}$
$F_5(\mathbf{x})$	$\frac{1}{1+x_1^2+x_2^2+x_3^2} + \sqrt{\exp(x_4+x_5)} + \lvert x_6+x_7\rvert + x_8x_9x_{10}$
$F_6(\mathbf{x})$	$\exp(\lvert x_1x_2\rvert+1) - \exp(\lvert x_3+x_4\rvert+1) + \cos(x_5+x_6-x_8) + \sqrt{x_8^2+x_9^2+x_{10}^2}$
$F_7(\mathbf{x})$	$(\arctan(x_1)+\arctan(x_2))^2 + \max(x_3x_4+x_6, 0) - \frac{1}{1+(x_4x_5x_6x_7x_8)^2} + \left(\frac{\lvert x_7\rvert}{1+\lvert x_9\rvert}\right)^5 + \sum_{i=1}^{10} x_i$
$F_8(\mathbf{x})$	$x_1x_2 + 2^{x_3+x_5+x_6} + 2^{x_3+x_4+x_5+x_7} + \sin(x_7\sin(x_8+x_9)) + \arccos(0.9x_{10})$
$F_9(\mathbf{x})$	$\tanh(x_1x_2+x_3x_4)\sqrt{\lvert x_5\rvert} + \exp(x_5+x_6) + \log\left((x_6x_7x_8)^2+1\right) + x_9x_{10} + \frac{1}{1+\lvert x_{10}\rvert}$
$F_{10}(\mathbf{x})$	$\sinh(x_1+x_2) + \arccos(\tanh(x_3+x_5+x_7)) + \cos(x_4+x_5) + \sec(x_7x_9)$

Approach-1b Manually creating data and modelling attention*

- Generating data according to certain conditions which are desirable to the distribution
- This method allows to visually "see" the data and manually custom-craft the data distribution and features
  - Used in Input-Cell Attention, Ismail et al., 2019.
  - Think architecture first, Manuscript



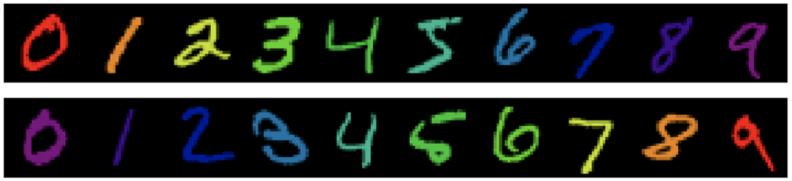
- Neural Network Attributions: A Causal Perspective, Chattopadhyay et al.



- Sample individual sequences uniformly of length between $[T, T+5]$. We used $T = 10$. Let x^t refer to the sequence value at length t.
- $\forall i; 2 < i \leq T$ Sample $x^i \sim \mathcal{N}(0, 0.2)$.
- With probability 0.5 either (a) sample $\forall i; 0 \leq i < 3$ $x^i \sim \mathcal{N}(1, 0.2)$ and label such sequences class 1 or (b) sample $\forall i; 0 \leq i < 3$ $x^i \sim \mathcal{N}(-1, 0.2)$ and label such sequences class 0.

Approach 2 - Modifying real datasets

- Altering real datasets
  - Adding color to Color MNIST - REPAIR, Li et al. and CDEP, Reiger et al. Adding color to MNIST - to learn if model looks at color or shape
  - DecoyMNIST - adding gray patch to the corner for images
  - Spurious signals (noise) in SST



- Creating new features from original datasets:
  - As described by using SLIC superpixel in images to graphs
  - Using word relations to create graphs

Evaluation Methods

Evaluation Methodologies (1)

- **Approach-1: Using qualitative assessments** - visuals like saliency maps in images or scores of different words in NLP tasks [Might include Human experiments]
- Papers using this approach:
  - Contextual Decomposition (CD), Murdoch et al., ICLR 2018
    - Picked phrases which showed ability to show negation
  - Hierarchical Interpretations of NN Predictions (ACD), Singh et al., ICLR 2019
    - Shown scores on images and sentences visually + human experiments
  - GradCAM, Selvaraju et al., ICCV 2017
    - Visual examination + error rates
  - Axiomatic Attribution for Deep Networks (IG), Sundararajan et al., ICML 2017
    - Visual examination + human perceived
  - "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Ribeiro et al., SIGKDD, 2016
    - Human experiments to identify the important regions
  - Explaining Explanations: Axiomatic Feature Interactions for Deep Networks, Janizek et al., 2020
    - Qualitative experiments on samples from datasets

- Papers with this approach
    - Neural Network Attributions: A Causal Perspective
        - Qualitative + simulated data
    - Can I trust you more? Model agnostic hierarchical explanations, Tsang et al, ICLR 2019
        - Randomly chosen phrases from the dataset, manual inspection
    - Interpretations are useful: penalizing explanations to align neural networks with prior knowledge, Reiger et al. ICLR 2020 (reject)
        - Penalizing interactions adding to loss function, training the network

Evaluation Methodologies (2)

- **Approach-2: Comparison** between different methods [Usually occurs along with papers in Approach-1]
- Papers using this approach:
  - Contextual Decomposition, Murdoch et al., ICLR 2018
    - Compared correlation b/w LR and CD, Gradient, LOO, Cell Decomp,IG
  - Sanity Checks for Saliency Maps, Adebayo et al., NIPS 2018
    - Correlation between various saliency methods and sanity checking their effectiveness by using experiments

Evaluation Methodologies (3)

- **Approach-3: Masking the most important attributions** (the ones found with highest scores) and calculating the drop in accuracy/performance [When proposing new attribution technique]
- Papers using this approach:
    - Data-Shapley : What is your data worth? Equitable Valuation of Data, Ghorbani et al., ICML 2019
        - Calculating score of a data point, removing highest, drop in accuracy calculated
    - L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data, Chen et al., ICLR 2019
        - Masking important features on text and images and calculating drop in Log odds ratio
    - Think Architecture First, Manuscript, 2020
        - Masking important features, drop in accuracy in LSTMs

Evaluation Methodologies (4)

- **Approach-4: Correlation between weights and outputs** (for methods which claim a particular technique provides good explanation) or dependence
- Papers which use this approach:
  - Attention is not Explanation, Jain et al. , ACL, 2019
    - Exploring correlation between attention weights and importance scores of grads/LOO showing no significant correlations

Evaluation Methodologies (5)

- **Approach-5:** Robustness
  - Interpretation of Neural Networks is Fragile*, Ghorbani et al., AAAI 2019
    - ?