Abstract  
  
**Introduction:**

Intro:

1. features attribution definition
2. Feature attribution variance tells us whether we have a consistent training process defined
3. inception definition, how may feature attribution in inception be different from other models
4. Test method summarize and result

Feature attribution is a method that decides the importance score of features, evaluating their contribution to making the correct answer given the data. Given the same dataset and a model architecture, and a set of hyperparameters, but with different initialized parameters, the function of specific neurons is surely different, so is their contribution to the correct prediction in the end. So is different branch of neural network, their functions varies from training process to training process; however, the amount of variation in their functions is not studies base on the best of our knowledge. This paper specific focus on the variation in feature attribution variation in branches of neural network. Specifically, multiple dataset is used to train with the neive inception model architecture, hundreds of times, with accuracy, loss, and feature attribution recorded, using different combinations of hyperparameters.   
  
Inception model is a model that uses different rbanches in each stack and each branch is meant to fit on some function, fore example for a kernel with size of 1\*1, it is more efficient when processing more details, and 5\*5 is good at processing the broader picture rathe than the details.

Motivation:

1. low variance gives us consistent resulting model, infer more consistent learning tendency in layers, provides better interpretability. The results thus give us insight in hyperparamas choices consistent and interpretable model
2. Provide a baseline for feature attribution variance in these models of small size and simple dataset
3. The potential of study the layer’s innate tendency of learning certain feature using the variance, for branch specialization - (later talkabout how variance is high for 3\*3 layer, implies that 11 55 layer tends to deal with lower level and higher level features but the 33 is in between to help them when needed to, thus have a much higher variance in fa), then explain why use inception

* Consistency and Interpretability: Highlighting that lower variance in feature attributions leads to more consistent models, which in turn enables better interpretability. This aspect is crucial for understanding the internal workings of deep learning models and for trustworthiness in applications where model decisions need to be explained.
* Hyperparameter Insights: The study provides valuable insights into how different hyperparameter choices influence the consistency and interpretability of the model. This can guide practitioners in making informed decisions about model configuration for optimal performance.
* Baseline Establishment: Offering a baseline for feature attribution variance in smaller models trained on simple datasets. This benchmark can be useful for future research and comparisons, especially in studies focusing on model behavior and layer specialization.
* Layer Learning Tendencies: Exploring the potential of using variance in feature attributions to study the innate tendencies of different layers in learning specific features. This approach is innovative in understanding how different layers contribute uniquely to the model's learning process.
* Role of Different Kernel Sizes: Discussing the observation that higher variance in 3x3 layers suggests a more flexible role in the network, compared to 1x1 and 5x5 layers which might be more specialized in dealing with lower and higher-level features, respectively. This insight could shed light on the hierarchical nature of feature processing in neural networks.

**Related work**

Branch specialization 3 paper, states that it’s important to explore

Deeplift ig details intro, its Example usage,

Improving performance of deep learning models with axiomatic attribution priors and expected gradients, LEARNING EXPLAINABLE MODELS USING ATTRIBUTION PRIORS, helps me to say that attribution can promise a consistent training process

Inception feature attribution papers, or visualization, or probing of the inner layers

Experiment

1. Details in setup
2. Preliminary tests conclusion, to show what to test here
3. Details setup in main experiments not talked about yet, such as dataset choice

Result

Batch size, lr, shuffle, others (graph, t test, paired t test, do a rank variance comparison as well)

11 33 55 fa var comparison

Result potential reason discussion

Talk about 11 33 55 fa value in ave a bit, and their potential explanation, briefly

Limitations

Conclusion(future directions)

Others data like acc recorded, can be used for studying predictability of feature attribution variance prediction, to predict the consistency in training process given a hyperarameter combo and the dataset complex encoding(using flops and acc)

Branch specialization, var in visualization or probing, use together with var in feature attribution