DS3 NOVEMBER PAPER READING GROUP

Explainable Deep Learning: A Field Guide for the Uninitiated

Tan Lin

PAPER READING GROUP GOALS

1. Summarize ‘Explainable Deep Learning: A Field Guide for the Uninitiated’
2. Facilitate Discussion on the Role and Importance of Explainability
3. Provide Additional Resources for Learning

ASKING QUESTIONS

I highly encourage you to ask questions during the presentation!

Raise your hand when you have a question and I will call on you.

AGENDA

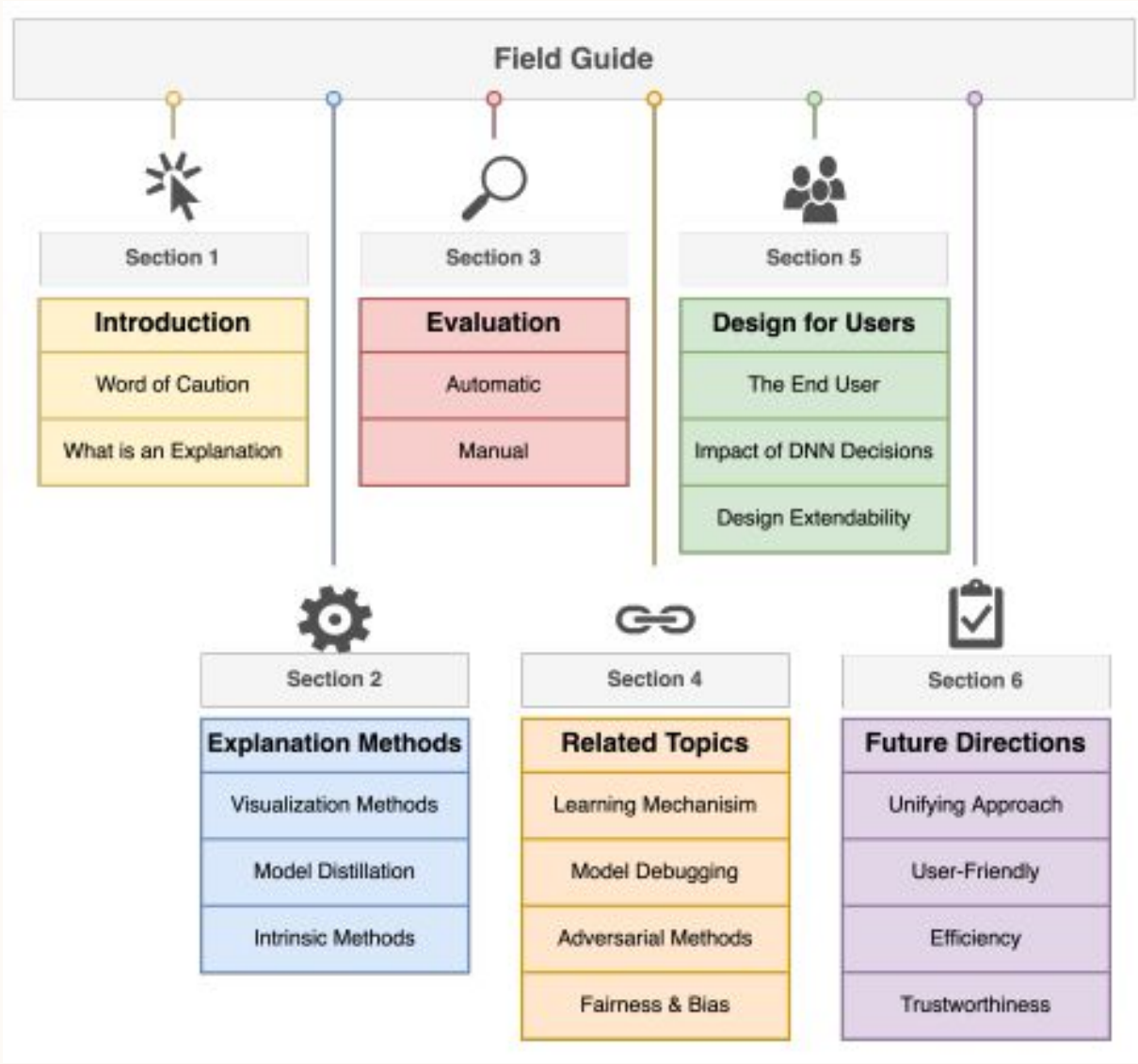
A. Paper Summary

1. Introduction
2. Explanation Methods
3. Evaluation
4. Related Topics
5. Designing Explanations for the User
6. Future Directions
7. Conclusion

B. Additional Discussion

C. Final Comments

A. PAPER SUMMARY



1. INTRODUCTION

MOTIVATION

- Deep neural network (DNN) artificial intelligence systems have become widespread across a wide range of disciplines.
- Explaining the workings of such networks can be difficult especially for those who are less experienced with DNNs.
- The paper is an introduction to the explainability of DNNs

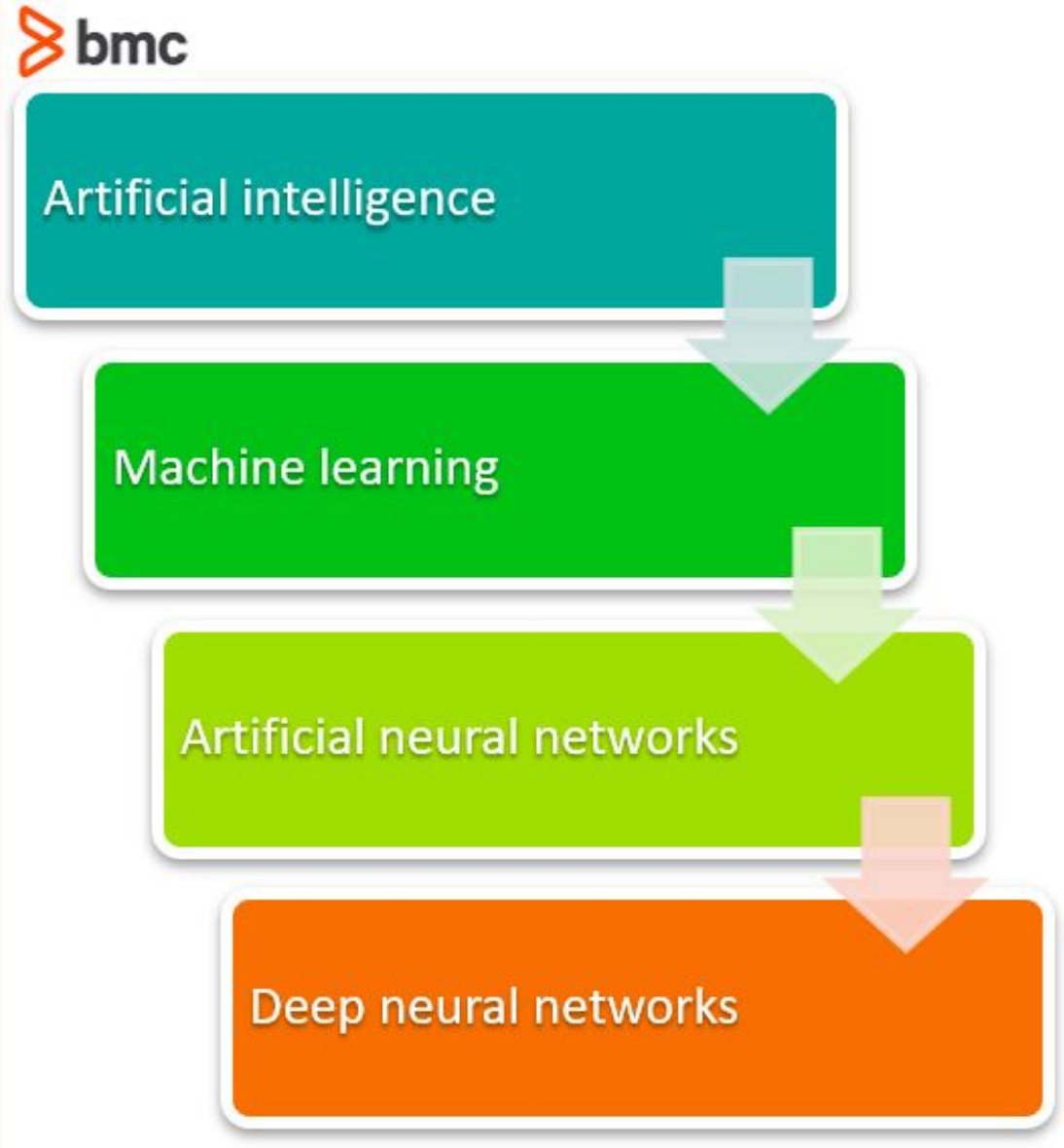
WHAT ARE DEEP LEARNING NETWORKS?

Artificial Intelligence refers to computer systems that can perform tasks that usually require human intelligence.

Machine Learning refers to the framework that automates statistical models and learns to improve their predictions.

Artificial Neural Networks refers to the use of hidden layers to store and evaluate the significance of the inputs on outputs.

Deep Neural Networks refer to the use of multiple hidden layers to deliver better predictions.



SUGGESTION VS PRESCRIPTION

- What are the pros and cons of suggestion/prescription?
- How can we trust an AIs decision?

TRUST AND JUSTIFICATION

- The paper refers to four different elements required for trust and justification:
  - Easily Interpretable
  - Relatable to the User
  - Connects the Context to the Choice
  - Reflects the Intermediate Steps

… AN EXPLANATION IS ANY INFORMATION THAT CAN HELP THE USER UNDERSTAND AND COMMUNICATE WHY A MODEL EXHIBITS SOME PATTERN OF DECISION-MAKING AND HOW INDIVIDUAL DECISIONS COME ABOUT.

TYPES OF EXPLAINABILITY

- Two Types of Explainability:
  - Model Training and Generalization
  - Model Predictions
    - Counterfactual
    - Contrastive

CONTRIBUTIONS

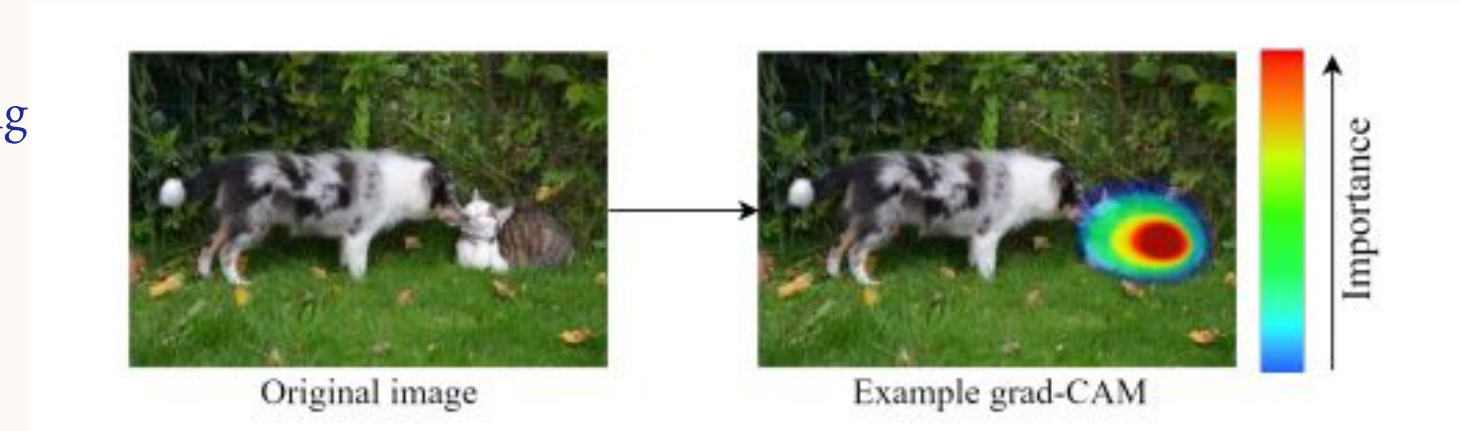
- This paper aims to:
  - Specifically Target the Explainability of Deep Neural Networks.
  - Support Readers in Understanding Explainable Deep Learning Architectures.
  - Introduce a New Method of Classifying Explainability Methods.
  - Connects Related Fields to Deepen Understanding.

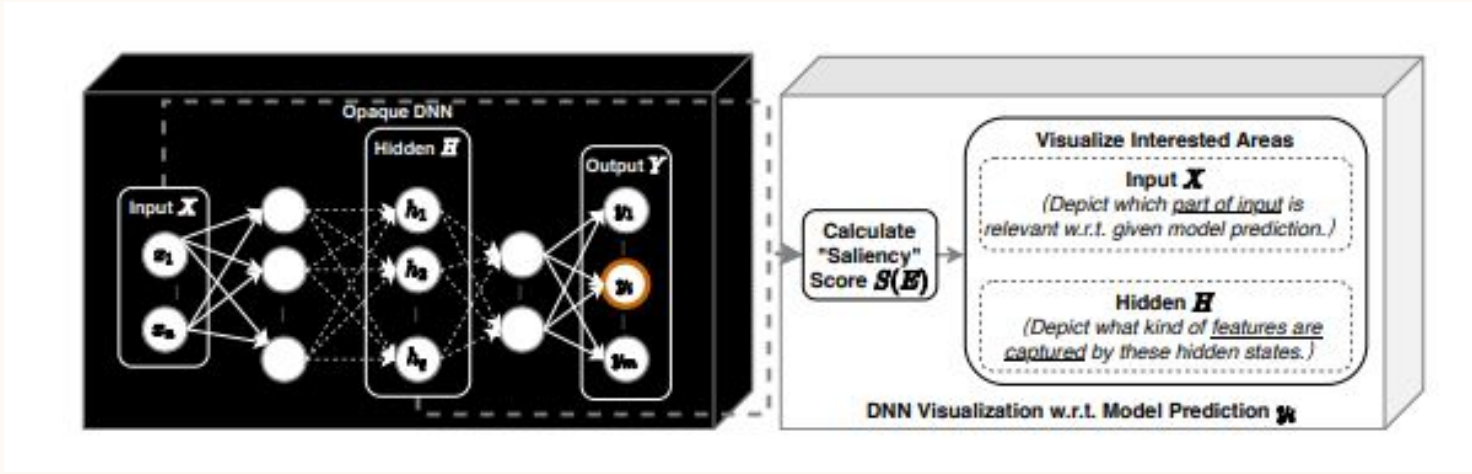
2. METHODS

How do we explain how a model works?

VISUALIZATION

- Backpropagation-based
  - Identifies **saliency** based on evaluating gradients during training
- Perturbation-based
  - Identifies **saliency** by comparing the network output of an input and a modified copy of the input.

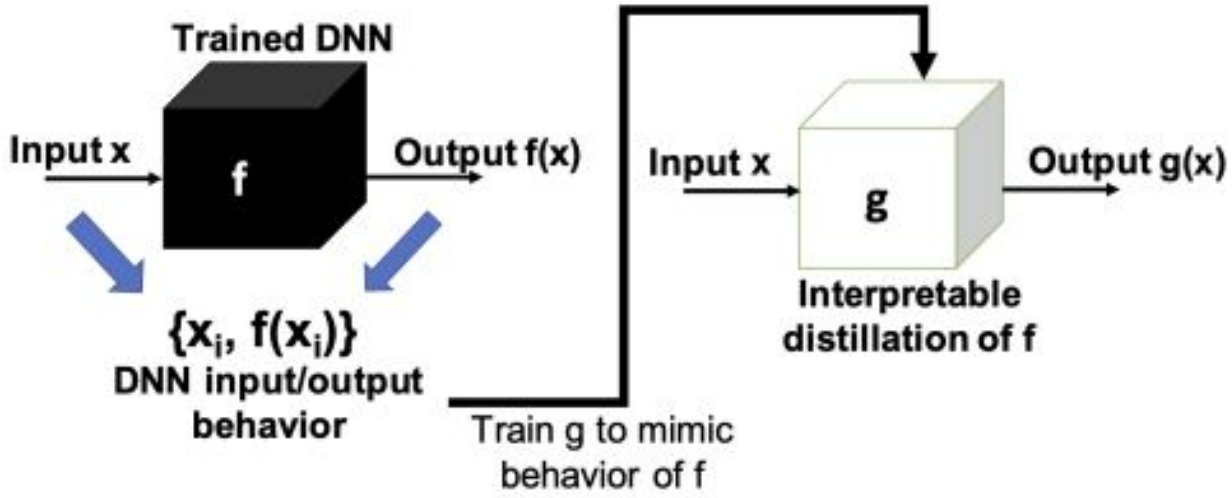




DISTILLATION

- Local Approximation
  - A simple model is used to approximate the input/output behavior of the DNN
- Model Translation
  - An alternative smaller model is used to mimic the behavior of the DNN

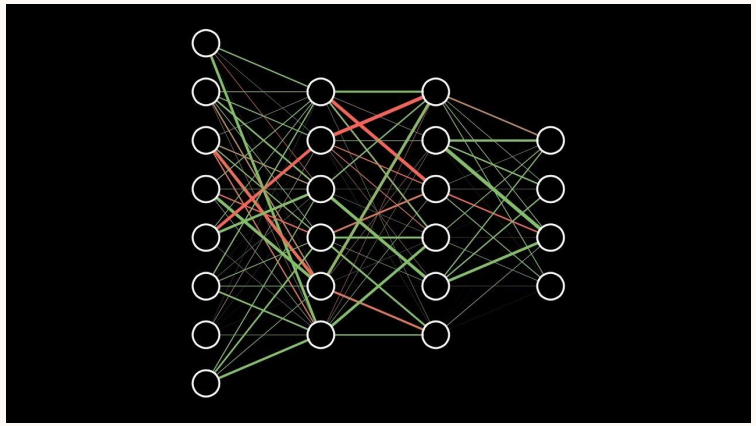
Table 2: Model distillation.

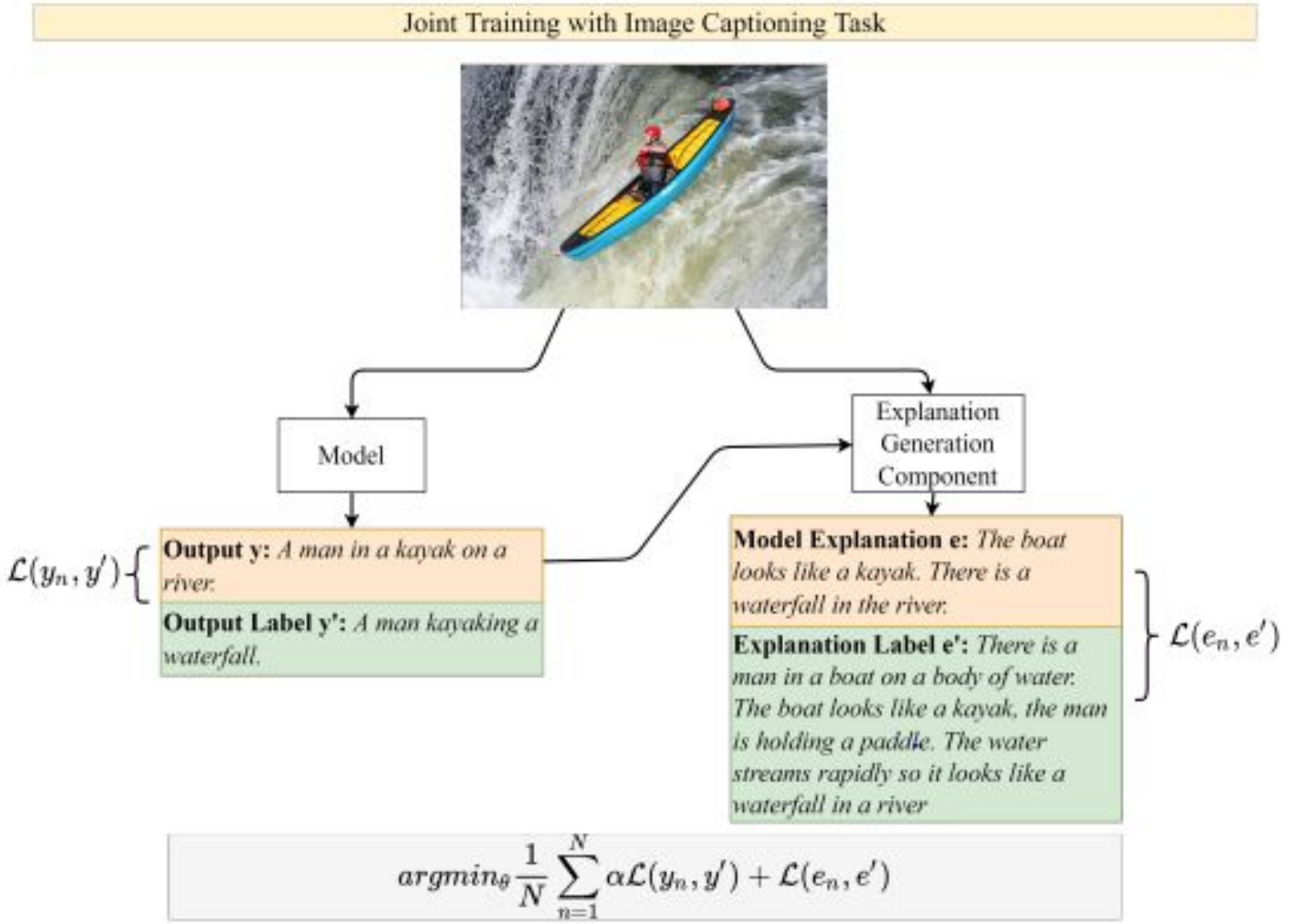


INTRINSIC

- Attention Mechanisms
  - It adds a mechanism that learns a conditional distribution over the inputs (weights).
- Joint Training
  - An additional model is trained alongside to provide explanations directly or indirectly.

1. Text Explanation
2. Explanation Association
3. Model Prototype




Joint Training with Image Captioning Task
Model
Explanation Generation Component
$\mathcal{L}(y_n, y')$
Output y: A man in a kayak on a river.
Output Label y': A man kayaking a waterfall.
Model Explanation e: The boat looks like a kayak. There is a waterfall in the river.
Explanation Label e': There is a man in a boat on a body of water. The boat looks like a kayak, the man is holding a paddle. The water streams rapidly so it looks like a waterfall in a river
$\mathcal{L}(e_n, e')$
$argmin_\theta \frac{1}{N} \sum_{n=1}^{N} \alpha \mathcal{L}(y_n, y') + \mathcal{L}(e_n, e')$

WHAT DO YOU THINK ARE SOME OF THE PROS AND CONS FOR EACH METHOD?

3. EVALUATION

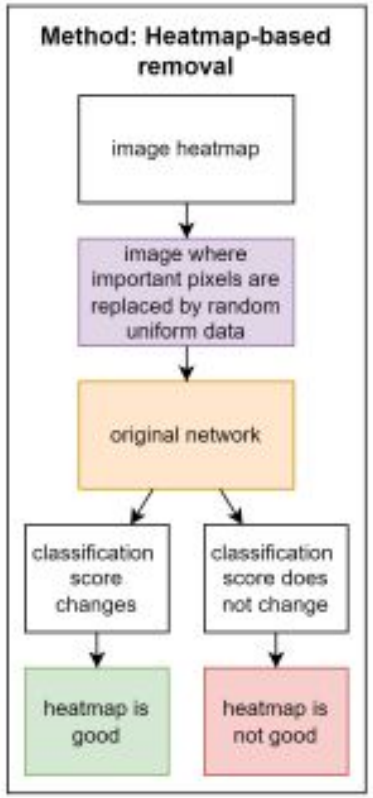
What makes a good evaluation?

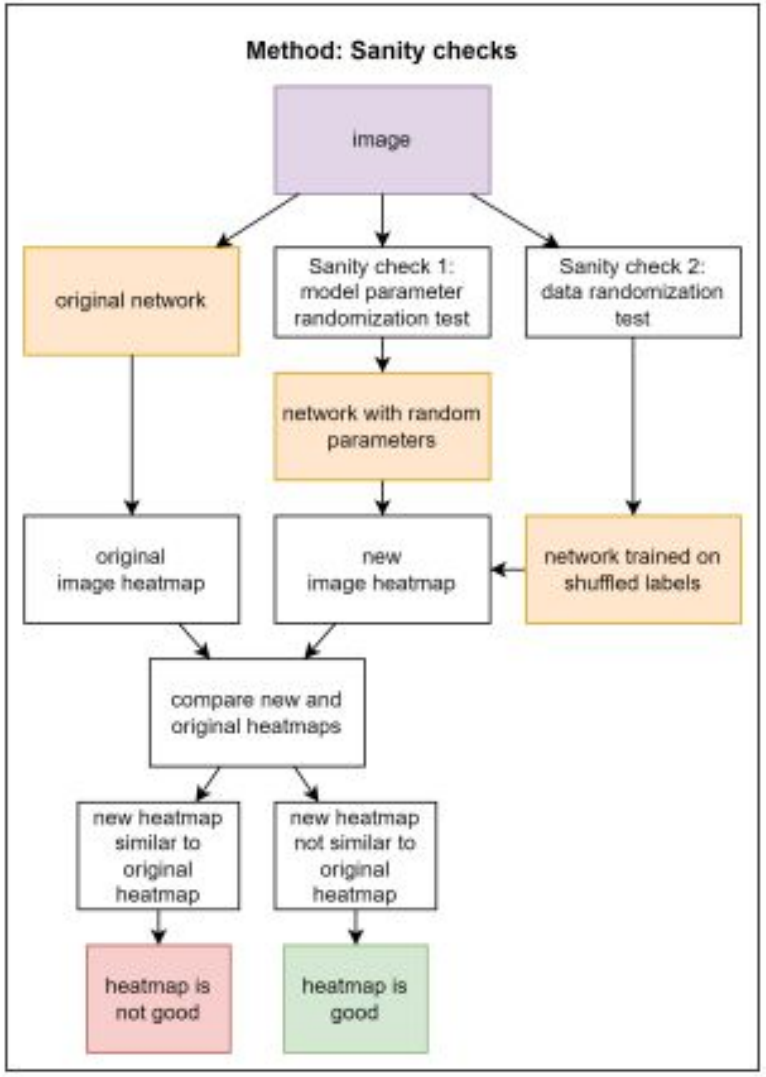
GOOD EXPLANATIONS

- What is important in a good explanation?
- **Fidelity**
- **Consistency**
- **Stability**
- **Comprehensibility**

BENCHMARKS

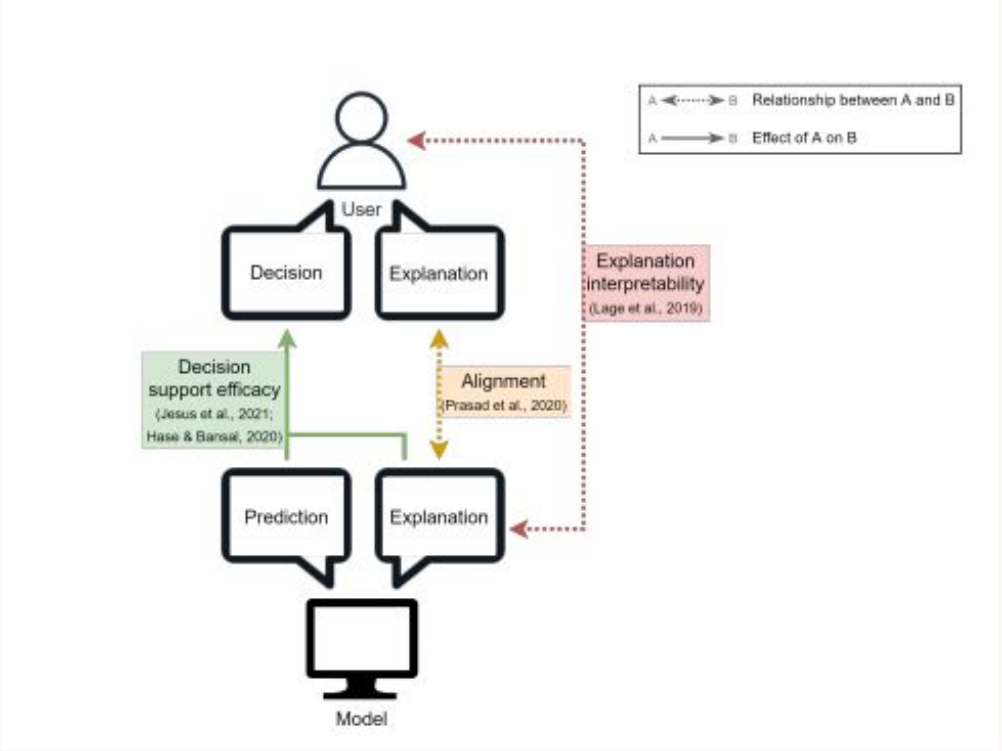
- Objective benchmarks are useful for comparing different models.
- Heatmaps are a common target for benchmarks.





HUMAN EVALUATION

- **Model Interpretability** is how easy it is for a human to predict the model output based on past predictions.
- Another way to judge explanations is by investigating how much model explanations align with human explanations.

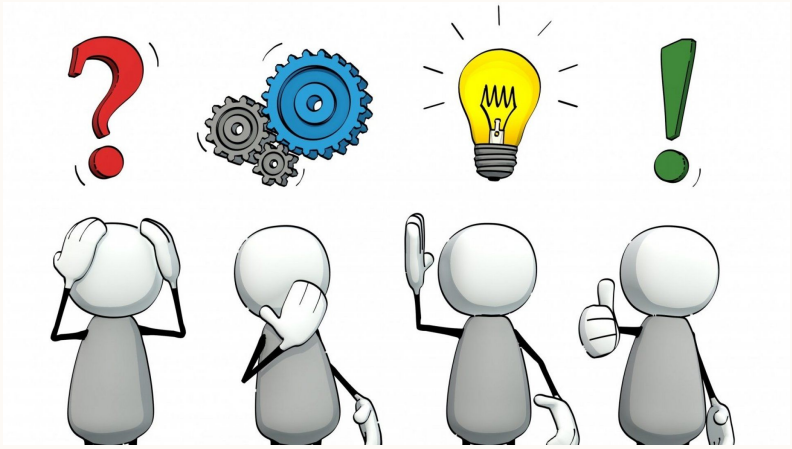


WHAT ARE SOME PROS AND CONS TO EACH METHOD OF EVALUATION?

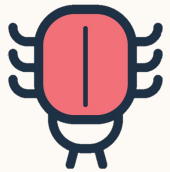
4. RELATED TOPICS

What other topics relate to explainability?

LEARNING MECHANISMS

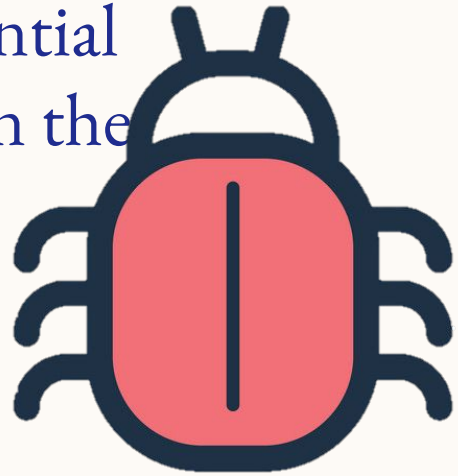


- Learning Mechanisms describes how models arrive at their conclusion.
- Some methods ascribe concepts to the model's learning process.
- Another way is to search for convergence in learned state.



MODEL DEBUGGING

- Refers to techniques focused on finding where errors in the model occur.
- An independent auxiliary model called a ‘probe’ is used to investigate the model to ensure that the model output is consistent with meta observations.
- ‘Neural stethoscopes’ quantifies the importance of specific influential factors and as well as promotes and suppresses information within the model.



ATTACK AND DEFENSE

- Adversarial attack refers to the use of artificial inputs designed to disturb the judgement of DNNs.
  - Black-box and White-box
  - Perturbations (What are a few different ways to perturb input?)
- Adversarial defense
  - Including Adversarial Attacks within Training Data
  - Filtering Perturbations

FAIRNESS AND BIAS

- Consider the difference between group fairness and individual fairness.
- Three Ways to Increase Fairness and Reduce Bias:
  - Pre-Processing
  - In-Processing
  - Post-Processing

5. DESIGNING EXPLANATIONS FOR THE USER

What do we have to consider when designing explanations?

END USER

Explanations require a case by case approach due to the variety of users.

What type of people use DNNs?

PRACTICALITY

Is it time critical, decision critical, or both?

What are some examples of each?

EXTENDIBILITY

It's expensive to make a unique explanation for every DNN model.

It's important for models to be modular and flexible enough to be adaptable to multiple architectures.

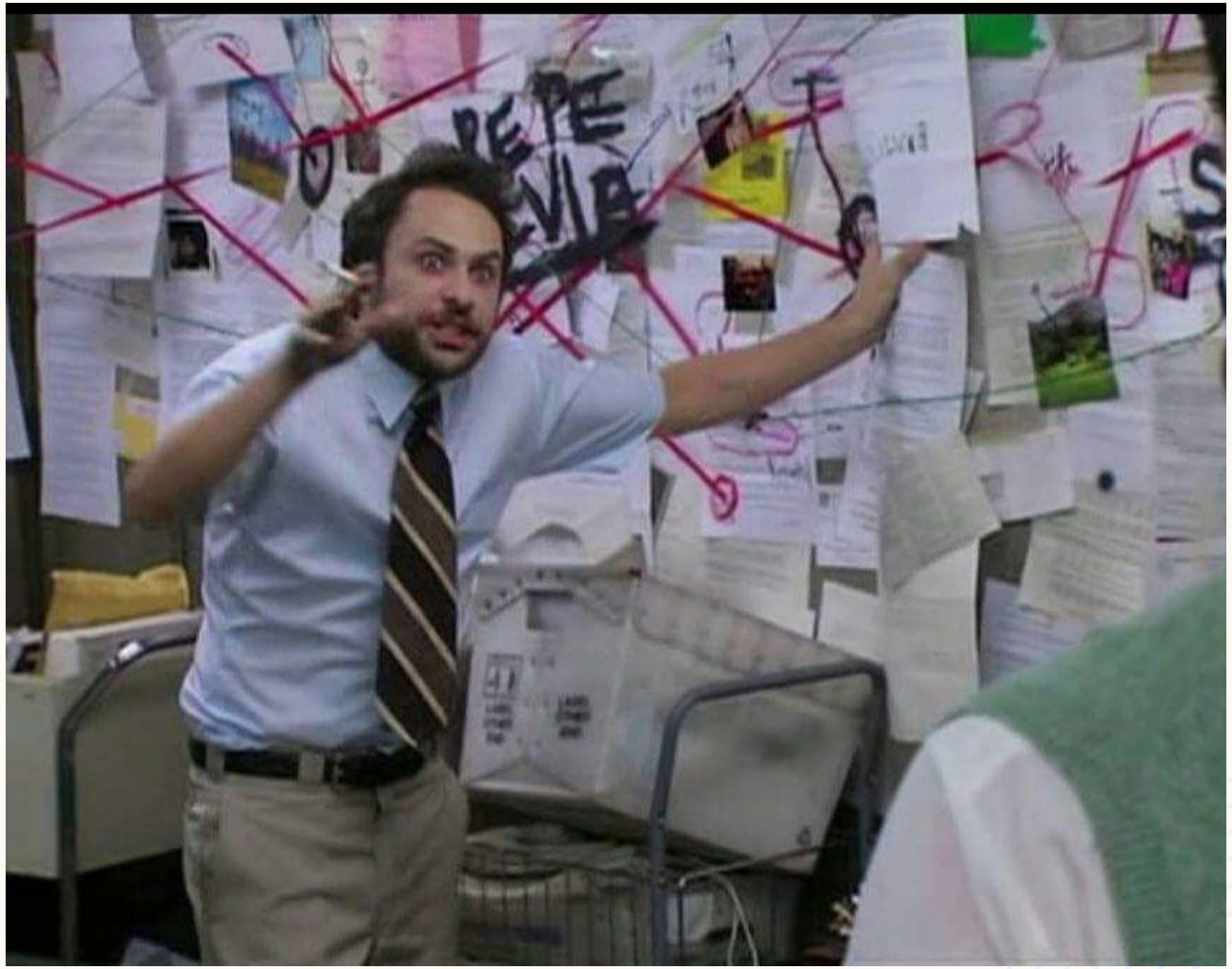
6. FUTURE DIRECTIONS

What else needs to be done to make DNNs more accessible?

UNIFYING EXPLAINABILITY

There have been many attempts to create a systematic framework for artificial intelligence or interpretable machine learning.

It's extremely difficult since many of the basic concepts of explainability are difficult to formalize.



USER-FRIENDLY EXPLANATIONS

- User-friendly explanations reduces the requirements for technical knowledge.
- An unfriendly explanations greatly restricts a models ability to be used in a wide range of situations.

EFFICIENCY

For time sensitive and decision sensitive situations, DNNs must be efficient enough to give the user enough time to react.

TRUSTWORTHINESS

As DNNs are used more widely, concerns about adversarial attacks and poisoned training data grow.

How can we trust DNNs if these issues are not addressed?

What are some possible consequences of adversarial attacks or poisoned training data sets?

7. CONCLUSIONS

DOES THE PAPER ACCOMPLISH ITS GOALS?

- Specifically Target the Explainability of Deep Neural Networks.
- Support Readers in Understanding Explainable Deep Learning Architectures.
- Introduce a New Method of Classifying Explainability Methods.
- Connects Related Fields to Deepen Understanding.

B. ADDITIONAL DISCUSSION

Remember those questions at the start?

YOUR QUESTIONS

1. HOW CAN WE TRUST AI TO MAKE DECISIONS THAT WE ARE RESPONSIBLE FOR?

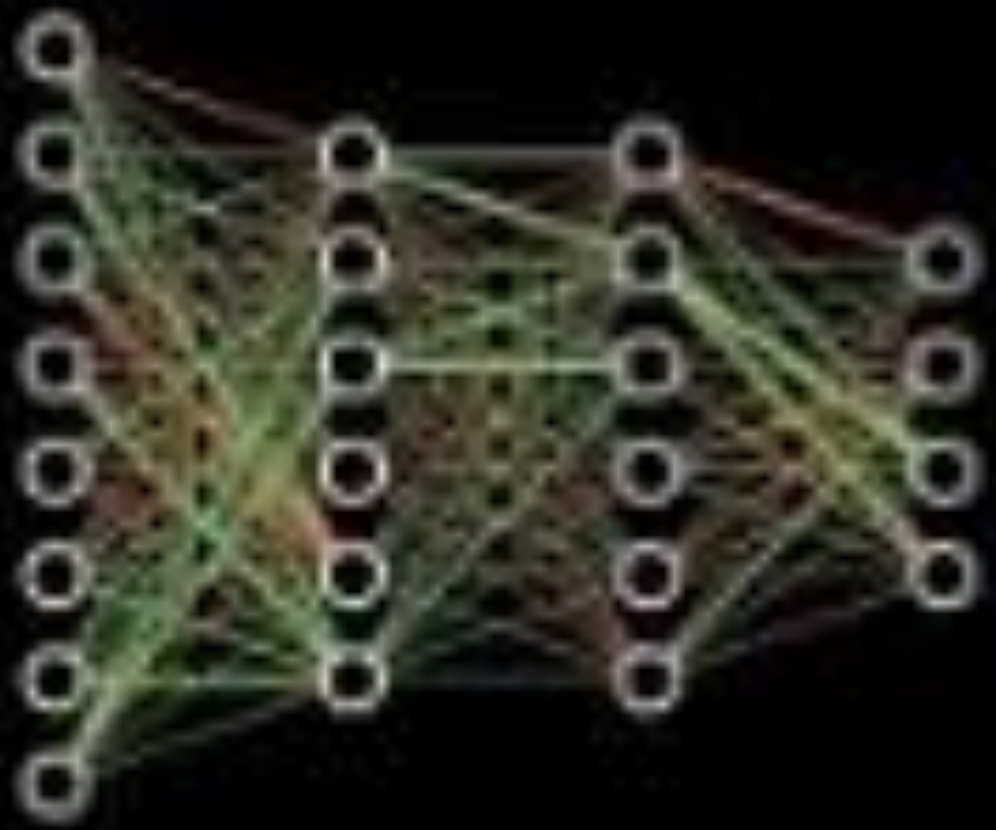
2. WHO IS ACCOUNTABLE FOR THE MISTAKES MADE BY ARTIFICIAL INTELLIGENCES?

3. WHAT IS A PROBLEM THAT YOU CARE ABOUT THAT CAN BE ADDRESSED BY ARTIFICIAL INTELLIGENCE?

4. WHAT IS EXPLAINABILITY'S ROLE IN REMAINING IN CONTROL OF OUR DEEP NEURAL NETWORKS?

C. FINAL COMMENTS

Neural Networks



From the ground up