State-of-the-art model-specific XAI techniques: Advantages, Limitations and Perspective

CSE3000 - RESEARCH PROJECT
RESPONSIBLE PROFESSOR: MAURO CONTI
SUPERVISOR: CHHAGAN LAL

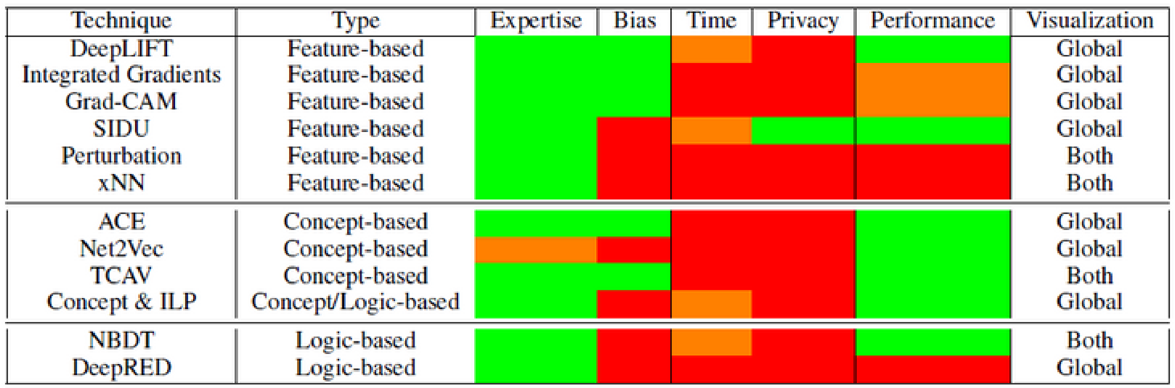
ARGHEM KHAN
m.a.khan-4@student.tudelft.nl

1. BACKGROUND

- IN THE RECENT PAST, MORE AND MORE CATASTROPHIC ISSUES OF AI SYSTEMS ARE BEING HIGHLIGHTED.
- THIS IS CAUSED BY THE BLACK-BOX NATURE OF AI SYSTEMS.
- THEREFORE QUESTIONS ARE BEING RAISED ON: TRANSPARENCY, BIAS, TRUST AND ETHICS OF AI SYSTEMS.
- BY NOT EXPLAINING OUR AI MODELS WELL ENOUGH, WE ARE AVOIDING ACCOUNTABILITY BUT ALSO PUTTING A LIMIT ON IMPROVEMENT.
- WE WILL ONLY FOCUS ON MODEL-SPECIFIC XAI: XAI TECHNIQUES THAT APPLY TO A SPECIFIC TYPE OF AI MODEL. MODEL AGNOSTIC TECHNIQUES ON THE OTHER HAND FOCUS ON XAI TECHNIQUES THAT WORK IN GENERAL

2. RESEARCH QUESTION

- THE MAIN OBJECTIVE OF THIS RESEARCH IS TO ANALYZE THE CURRENT MODEL-SPECIFIC XAI TECHNIQUES.
- HOW DO THE TECHNIQUES COMPARE TO EACH OTHER?
- WHICH REQUIREMENTS SHOULD A GOOD TECHNIQUE ADHERE TO? WHAT ARE THE LIMITATIONS OF CURRENT TECHNIQUES? CAN WE ADDRESS SOME OF THEM?
- FINALLY, WHAT IS THE SCOPE FOR FUTURE WORK?
- WHEN LOOKING AT MODEL-SPECIFC XAI, THIS RESEARCH ONLY FOCUSES ON XAI TECHNIQUES FOR DEEP LEARNING METHODS (NEURAL NETWORKS (NN)).



Technique	Type	Expertise	Bias	Time	Privacy	Performance	Visualization
DeepLIFT	Feature-based	🟩	🟩	🟧	🟥	🟩	Global
Integrated Gradients	Feature-based	🟩	🟩	🟥	🟥	🟧	Global
Grad-CAM	Feature-based	🟩	🟩	🟥	🟥	🟧	Global
SIDU	Feature-based	🟩	🟥	🟧	🟩	🟩	Global
Perturbation	Feature-based	🟩	🟥	🟥	🟥	🟥	Both
xNN	Feature-based	🟩	🟥	🟥	🟥	🟥	Both
ACE	Concept-based	🟩	🟩	🟥	🟥	🟩	Global
Net2Vec	Concept-based	🟧	🟥	🟥	🟥	🟩	Global
TCAV	Concept-based	🟩	🟩	🟥	🟥	🟩	Both
Concept & ILP	Concept/Logic-based	🟩	🟥	🟧	🟥	🟩	Global
NBDT	Logic-based	🟩	🟥	🟧	🟥	🟩	Both
DeepRED	Logic-based	🟩	🟥	🟥	🟥	🟥	Global

FIGURE 1: AN OVERVIEW OF HOW THE DIFFERENT TECHNIQUES PERFORM ON THE GENERAL REQUIREMENTS. GREEN = GOOD, ORANGE = AVERAGE, RED = BAD

3. COMPARISON

- THE TECHNIQUES CAN BE DIVIDED IN FEATURE-BASED, CONCEPT-BASED AND LOGIC-BASED (FIGURE 1)
- THERE ARE SOME GENERAL REQUIREMENTS.
- EXPERTISE: OVERALL EASY TO USE.
- BIAS: NOT A LOT OF WORK DONE. PROVEN TO BE A DIFFICULT TOPIC.
- TIME: OVERALL NOT EFFICIENT. SOME DIRECT COMPARISONS (E.G. INTEGRATED GRADIENTS VS DEEPLIFT)
- PRIVACY: NO REAL PRIVACY AWARENESS
- PERFORMANCE: FEATURE-BASED TECHNIQUES HAVE SOME GOOD POINTS. CONCEPT-BASED PERFORM WELL OVERALL. LOGIC-BASED COMPROMISES ON ACCURACY OF EXPLANATION

4. FUTURE WORK

- TRADE-OFF BETWEEN ACCURACY AND EXPLAINABILITY
- USE OF HYBRID TECHNIQUES. CONCEPT & ILP HAS PROVEN TO GAIN MORE HUMAN TRUST
- GUIDELINES FOR EVALUATING FEATURE-BASED TECHNIQUES. IMPORTANCE SCORES ARE BEING EXTRACTED BUT THERE IS NO WAY TO FIND OUT HOW ACCURATE IT IS
- CURRENT TECHNIQUES SHOULD EXPAND ON MORE DATA TYPES (AUDIO, TABULAR OR SEQUENTIAL)
- PRIVACY AWARENESS IS LACKING AND ALL TECHNIQUES SHOULD LOOK INTO THIS