

PhD Progress Update

James Hinns

April 2024

1 Overview

- **Intro**

This document is split into two major sections: Research; which covers current and future research I am undertaking as a part of this PhD, and PhD Study Programme; which covers my current and future progress towards the assigned goals for the study components.

- **Clear aim and contributions** This PhD aims to enhance the understanding and transparency of machine learning predictions through the introduction of novel methods and evaluations of eXplainable AI (XAI).

- **Progress towards Study Program - Required Credits: 30**

Competency	Name	Credits	Date
A	DLCV	6	June 23
C	Masters Supervision	1	Jan 2023
C	Masters Supervision	1	Jan 2024
D	XAIStories arXiv	2	Sep 2023
Total Credits			10

2 Research

2.1 Current

XAIStories

In today's critical domains, the predominance of black-box machine learning models amplifies the demand for Explainable AI (XAI). The widely used SHAP values, while quantifying feature importance, are often too intricate and lack human-friendly explanations. Furthermore, counterfactual (CF) explanations present 'what ifs' but leave users grappling with the 'why'. To bridge this gap, we introduce XAIStories. Leveraging Large Language Models, XAIStories provide narratives that shed light on AI predictions: SHAPstories do so based on SHAP explanations to explain a prediction score, while CFstories do so for CF explanations to explain a decision. Our results are striking: over 90% of the surveyed general audience finds the narrative generated by SHAPstories convincing. Data scientists primarily see the value of SHAPstories in communicating explanations to a general audience, with 92% of data scientists indicating that it will contribute to the ease and confidence of nonspecialists in understanding AI predictions. Additionally, 83% of data scientists indicate they are likely to use SHAPstories for this purpose. In image classification, CFstories are considered more or equally convincing as users own crafted stories by over 75% of lay user participants. CFstories also bring a tenfold speed gain in creating a narrative, and improves accuracy by over 20% compared to manually created narratives. The results thereby suggest that XAIStories may provide the missing link in truly explaining and understanding AI predictions.

Aim: To enhance the understanding behind AI predictions for lay-users by providing natural language explanations.

Under revision for Decision Support Systems.

CoF Tables

The rise of deep learning in image classification has brought unprecedented accuracy but also highlighted a key issue: the use of 'shortcuts' by models. Such shortcuts are easy-to-learn patterns from the training data that fail to generalise to new data. Examples include the use of a copyright watermark to recognise horses, snowy background to recognise huskies, or ink markings to detect malignant skin lesions. The explainable AI (XAI) community has suggested using instance-level explanations to detect shortcuts without external data, but this requires the examination of many explanations to confirm the presence of such shortcuts, making it a

labour-intensive process. To address these challenges, we introduce Counterfactual Frequency (CoF) tables, a novel approach that aggregates instance-based explanations into global insights, and exposes shortcuts. The aggregation implies the need for some semantic concepts to be used in the explanations, which we solve by labelling the segments of an image. We demonstrate the utility of CoF tables across several datasets, revealing the shortcuts learned from them.

Aim: To detect shortcuts in image classification models through the aggregation of local explanations.

Currently under review at KDD, likely resubmitting to NeurIPS.

2.2 Future

- **PCherry** Given a set of counterfactuals, model and data they explain, can we get a probability that there is cherry picking in their selection.
- **Something to do with Optimisation (Explanation Combination, further exploration, improvement of CoF tables).**
- **Personalised Explanation Improvement**
- **Fairness of Image Classification**
- **Fairness of Counterfactual Explanations Algorithms**

3 PhD Study Programme

3.1 Courses

- **DLCV Summer School**
- **Kenneths Optimisation**
 - Need to work on paper - outline idea
 - Could work in PCherry

3.2 Other Requirements

- **Masters Supervisions**

Student Name	Date of Completion
Camille Dams	31/01/23
Margot Willemse	29/01/24
Tibo Vanleke	May 2024

4 Other

- **Research Stays**
- **Cambridge Summer School and Probabilistic Machine Learning**
- **Reviewer for KDD**