

ICCBR Doctoral Consortium 2022

Kerstin Bach^{1,*}, Stelios Kapetanakis²

¹*Norwegian University of Science and Technology (NTNU), Høgskoleringen 1, Trondheim, 7034, Norway*

²*The University of Brighton, Lewes Road, Brighton, BN2 4AT, UK*

Abstract

The thirteenth Doctoral Consortium (DC) was held on September 11-12 2022 in Nancy, France, as part of the 30th International Conference on Case-Based Reasoning.

Preface

The thirteenth ICCBR Doctoral Consortium (DC) was held in September 2022 in Nancy, France. After two years of virtual conferences, in 2022 ICCBR and the DC resumed as an on-site and hybrid event. Since ICCBR 2009, the DC has been an integral part of the annual Case-Based Reasoning conference inviting Ph.D. candidates to submit their research statements to be discussed with senior members of the community.

Ph.D. candidates who applied to the program submitted summaries of their doctoral research. In their research summaries, they detailed the problems they are addressing, outlined their proposed research plans, and described progress to date. We received 16 submissions this year and accepted 10 students to attend the DC. Accepted applicants were paired with mentors who helped them to refine their research summaries in light of reviewer feedback. The updated research summaries, which appear in this volume We are proud to carry on the tradition with a cohort of ten doctoral students from five different countries.

The DC activities kicked off on Sunday, September 11th for a two hour session during which mentees and mentor met to discuss the research statements and made final preparations for the presentation on Monday. Nine out of ten contributions were presented orally during Monday, September 12th. Each student gave a 20 min presentation of their work followed by a discussion led by each mentor.

In the research statement submissions, we could clearly see the trend of developing XAI methods. Craig Pirie, Michael Clemens, Greta Warren, Pedram Salimi, Malavika Suresh, Eoin Delaney all address XCBR in their research focusing on various input data. Innovative applications in sports and health have been presented by Ciara Feely and Paola Marin, while Mark van der Pas presented an approach for Case-Based Reasoning for Manufacturing Incident Handling. Zachary Wilkerson's work on using deep learning (DL) methods to learn features represents the work on pairing DL and CBR.

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France

*Corresponding author.

✉ kerstin.bach@ntnu.no (K. Bach)

✉ 0000-0002-4256-7676 (K. Bach)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

We are thankful to the AI Journal for their support of the DC. Together with support from the ICCBR 2022 organizers, we received funding that allowed us to waive the registration fees and cover the accommodation for DC participants. Furthermore, we would like to thank the 17 PC members who gave detailed and valueable feedback on the research statement.

Thank you to all of the students, mentors, and program committee members who worked so hard to make the DC a success.

Kerstin Bach and Stelios Kapetanakis

Nancy, France, September 2022

Program Committee

- David Aha, Naval Research Laboratory, USA
- Klaus-Dieter Althoff, DFKI / University of Hildesheim, Germany
- Isabelle Bichindaritz, SUNY Oswego, USA
- Sutanu Chakraborti, IIT Madras, India
- Sarah Jane Delany, Technological University Dublin, Ireland
- Mark Keane, University College Dublin, Ireland
- David Leake, Indiana University, USA
- Jean Lieber, Université de Lorraine, France
- Kyle Martin, Robert Gordon University Aberdeen, Scotland
- Stewart Massie, Robert Gordon University Aberdeen, Scotland
- Mirjam Minor, Goethe University Frankfurt, Germany
- Stefania Montani, Università del Piemonte Orientale, Italy
- Juan Antonio Recio Garcia, University Complutense of Madrid, Spain
- Antonio Sanchez, University Complutense of Madrid, Spain
- Barry Smyth, University College Dublin, Ireland
- Rosina Weber, Drexel University, USA
- David Wilson, University of North Carolina at Charlotte, USA

Table of Contents

The Three E's of Explainability in Collaborative Computational Co-Creativity: Emotionality, Effectiveness, and Explicitness	1
<i>Michael Clemens</i>	
Case-based Explanation for Black-Box Time Series and Image Models with Applications in Smart Agriculture	7
<i>Eoin Delaney</i>	
Explaining and Upsampling Anomalies in Time Series Data	13
<i>Craig Pirie</i>	
Addressing Trust and Mutability Issues in XAI utilising Case-Based Reasoning	19
<i>Pedram Salimi</i>	
CBR For Interpretable Response Selection In Conversational Modelling	25
<i>Malavika Suresh</i>	
Counterfactual Explanations for eXplainable AI (XAI)	31
<i>Greta Warren</i>	
Using Machine Learning Techniques to Support Marathon Runners	36
<i>Ciara Feely</i>	
The Use of Case-Based Reasoning for Personalizing Musculoskeletal Pain Treatment Recommendations	42
<i>Paola Marin</i>	
Developing a Decision Support System leveraging Distributed and Heterogeneous Sources: Case-Based Reasoning for Manufacturing Incident Handling	48
<i>Mark van der Pas</i>	
DL-CBR Hybridization for Feature Generation and Similarity Assessment	54
<i>Zachary Wilkerson</i>	

The Three E's of Explainability in Collaborative Computational Co-Creativity: Emotionality, Effectiveness, and Explicitness

Michael P. Clemens

University of Utah, 50 Central Campus Drive, Salt Lake City, 84112, USA

Abstract

While explainable computational creativity (XCC) seeks to create and sustain computational models of creativity that foster a collaboratively creative process through explainability, there remains no way of quantitatively measuring these models. Through this research, we propose *The Three E's of Explainability in Collaborative Computational Co-Creativity: Emotionality, Effectiveness, and Explicitness* to quantitatively assess the artists' experience within the system concerning this communication paradigm.

Keywords

computational creativity, explainable artificial intelligence,

1. Introduction

With the recent explosion of work using neural networks and deep learning for co-creative applications, researchers are calling for explainability within these computational models [1]. The work surrounding this effort is called Explainable Computational Creativity (XCC). Although there have been efforts to explore this area—e.g. the work by Zhu et al. 2018 for video games, Bryan-Kinns et al. 2022 for music—none to date have defined a framework for evaluating a computational model's *explainability* within collaborative co-creative applications. This research introduces the *Three E's of Explainability in Collaborative Computational Co-Creativity: Emotionality, Effectiveness, and Explicitness* and elaborates on each E related to its creative application.

Success in computational modeling has led to a surge of research that promises autonomous systems to learn, decide, and act on their own volition. Although these systems have produced tremendous results within their respective contexts, their effectiveness is often hindered by the lack of transparency from the model itself [4]. From this lack of transparency, humans are often reluctant to implement techniques that are not interpretable, tractable, and trustworthy [2].

XAI explores how computational models such as ensembles, neural networks, or deep-learning methods can be made more understandable to humans. The motivation behind this field of

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France

*Corresponding author.

✉ michael.clemens@utah.edu (M. P. Clemens)

🌐 http://mclem.in/ (M. P. Clemens)

🆔 0000-0002-4507-8421 (M. P. Clemens)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

research is to increase the usability and accessibility for non-experts to utilize these models intuitively in their respective domains [5].

XCC has been presented as a sub-field of XAI, emphasizing the construction of models that foster bi-directional communication between the user and the system. Within this same context—and grounded in HCI and creativity literature—Bryan-Kinns et al. (2022) argued that AI in the interactive arts can be systematically analyzed in three ways:

1. Per the role of AI - ranging from models that perform generative tasks without engaging with humans to AI models that serve as collaborators in creative partnerships.
2. Per the interaction with the AI - the more interactive and responsive an AI is, the more likely people would grasp what it is doing now and in the future.
3. Per the common ground with the AI - classify what a person might be able to infer about an AI's output state.

They argue—and we agree—that a creative AI's explainability depends on all these facets. For instance, more explanation seems necessary as a process becomes more collaborative, demanding more engagement and grounding. Further, increased contact with the agent supports individuals in learning about and inferring knowledge and understanding of the co-creative AI. We complement their case-study based work by arguing that explainability can *also* depend on the creative AI's status relative the creativity literature surrounding the Four P's [6].

2. Research Plan

The Three E's is a framework we created to evaluate a system's explainability, specifically in the context of co-creative applications. This framework does not seek to find the optimal solution for every creative domain within computational creativity (CC). Rather, this framework provides a tool for members of the CC community to use while building explainability into their computational models as it is needed. Although our Three E's should *not* be used as a way to justify whether the system is inherently explainable, this tool can be used to guide designers in curating the type of experience they want to create between the co-creative agent and the artist.

2.1. Research Objectives

The main research objective is to create a framework that can be used to quantitatively assess the explainability of a model employed within a co-creative application. Researchers have converged on evaluating a system's creative potential relative to *The Four P's: Person, Process, Product, and Press* [6]. That is, modern evaluations use at least one P to describe the work's type of impact on the field [7]. We therefore plan to discuss each P as it relates to explainability.

At the same time, we note that the need for explainability is (at least) culturally determined. Different cultures have idiosyncratic information needs and processes that demand unique information architectures [8]. In turn, the systems are impacted by the environment in which they are deployed. Hence, culture has a profound bi-directional influence on co-creative system design [9]. Our current measures of E's within P's have the caveat that we do so from a global Western cultural perspective and that other cultures may require different levels of emotionality, effectiveness, and explicitness more appropriate for their explainability needs.

2.2. Approach / Methodology

In this section, we explore each of the E's in detail and justify our rationale behind using them to deepen our understanding of explainability within collaborative computational co-creative applications.

Emotionality In collaborative systems, it's imperative to discern how affect will be demonstrated as it has a significant influence on how the interaction between the participants unfolds and, subsequently, what kind of sensemaking is derived from this interaction [10]. Participants' reactions are triggered by emotions induced by stimulus events, allowing them to adjust to an ongoing collaboration [11]. During a collaborative session, participants can be aware of their collaborators' feelings, which helps them verify their actions from their collaborator's perspective and use this awareness to continue with participatory sensemaking [12]. As Leite et al. (2013) have demonstrated, reifying these characteristics is essential: meaningful human-robot relationships are *shaped* by the robot's ability to communicate emotions.

Inspired by this line of work, we present *emotionality* as our first key dimension of explainability. We define *emotionality* as the agent's capacity to (a) understand the user's emotions and (b) offer feedback that predictably elicits targeted emotion(s). While working alongside co-creative agents, artists will ascribe certain beliefs and values to that agent [14]. To meet artists' expectations, the system's interaction framework should include the articulation of emotional input from the user and provide the appropriate feedback for the user to observe emotional output.

This dimension begs the question: "*How do you measure the emotionality of the explainability within a system?*" While it is not obvious how to measure this property of a model, we propose using a *high*, *medium*, and *low* scale for the amount of emotional feedback observed by the artist from the co-creative agent, as well as for the amount of emotional input the system affords the artist to articulate. We might imagine using this same scale to quantify *effectiveness* and *explicitness*. For example, high emotionality might be described as a co-creative agent receiving the articulation of emotional input by the user and presenting emotionally relevant feedback that is observable. Medium emotionality might be a co-creative agent that can receive emotional input from the user yet the system presentation lacks emotionally relevant feedback. An agent that cannot engage emotionally with the artist might have low emotionality.

Effectiveness Designing creative systems ultimately requires evaluating their underlying computational models of creativity [15]. When evaluating these systems, the term *effective* is used to express whether the system was successful in accomplishing its intended goal.

Although we as a community wish to have a framework that supports a standardized way of evaluating a creative system's effectiveness, we lack an agreed-upon metric that can be used across domains [16]. Further, the term *effective* itself is subject to interpretation. For example, Hartson et al. (2001) use it to denote the *thoroughness* and *validity* of a system, per quantitative *usability* evaluations—a system is effective (i.e. achieves its intended goal) to the degree it is thorough and valid.

In our Three E framework, we reformulate the term *effectiveness* to describe how well the user's mental model of the creative system corresponds to its exhibited behavior. In the design

sciences, Gero and Kannengiesser (2004) argue that designers perform an *evaluation* to identify whether the user behavior a designed artifact *should* elicit corresponds to the behavior that *actually manifests* from the designed artifact's use. Here, we generalize this notion to include all steps of the design process, not solely the artifact's evaluation. That is, *effectiveness* reflects the user's capacity to *predict* (and thereby *direct*) how the co-creative system will act based on their mental model of the system.

High effectiveness might describe a frictionless match between the agent's behavior and the expected behavior. Medium effectiveness might describe varying discontinuities between the system behavior and the user's mental model of the agent's behavior. Low effectiveness represents almost no match between how the system behaves and the expected behavior derived by the structure from the user.

Explicitness Interpretability and explainability have become conflated in the AI literature [19]. To situate our work, we rely on the use of these terms within XAI. Recently, Alvarez Melis and Jaakkola (2018) proposed that explanations should meet three general characteristics: explicitness, faithfulness, and stability. To them, explicitness addresses the question: “*Are the explanations immediate and understandable?*” Relatedly, Palacio et al. (2021) define explicitness as “*how understandable are the explanations*” relative to the ease of a person’s interpretation for given explanations.

Inspired by this line of work, we propose explicitness ought to assess both the explainability of the model and the justification for the model’s complexity. Explaining black-box models mathematically or computationally may not be appropriate for tackling explainability in CC applications [22]. Instead, justification and rationale for added complexity should be centered.

High explicitness might describe when non-specialists can readily understand the model’s explanations without the intervention of an expert user. Medium explicitness might denote when the explanations require domain-specific knowledge but are still understandable within that context. Low explicitness might describe a model’s explanations that are either completely absent or unintelligible by anyone other than a domain expert.

Computational Model Although the proposed evaluation framework focuses on assessing and defining the explainability of an artifact, we have yet to discuss what an explanation is and how it manifests itself via a computational creativity system. In this manner, we proposed a computational model including George Abowd’s [23] four agents in his *Framework for Discussing Interaction*: two explicit (*User* and *System*) and two implicit (*execution* and *evaluation*). *Execution* is articulation from the *User* of the problem and the *performance* metric from the *Input* to the *System* itself. *Evaluation* is the presentation from the system, creating the *Output* that is then observed by *User*.

Figure 1 demonstrates a computational model of an explanation in CC domains, including three stages, the *User*, the explainable user interface (XUI), and the Computational Creative Agent (CCA). The XUI is the interaction for the explanation between the CCA and the User. The User articulates the problem they want to explore (e.g., a *trigger*). The XUI will take this trigger, create a problem formulation based on the User’s goal, and articulate that need to the CCA. The CCA will interpret this Input as the User’s Goal State. The User’s Goal and Initial

State will be assessed in the Planning Stage, where the CCA will produce a plan to take the User from the Initial State to the Goal State through a series of explanations. The Output from the Planning Stage will be the Input to the Plan Synthesis, which will focus on the XUI. This step in the process will determine from the list of explanations provided by the planning stage which will be most effective for the User. The Output of the Plan Synthesis will be both an Output to the User and an Input to the User's Initial State. This pathing is due to the long-term memory design principle within XCC systems, where the system will update over time based on what the User has learned. As the system explains more of the process, these explanations will be added to the memory bank as explanations that have been used previously and the determination of whether they were influential on that User. Let us take a User who has been presented with an explanation yet continuously asks the same trigger question from the XUI to the CCA. The CCA should be able to determine that the chosen explanation used was not a sufficient explanation for the User's intended goal state.

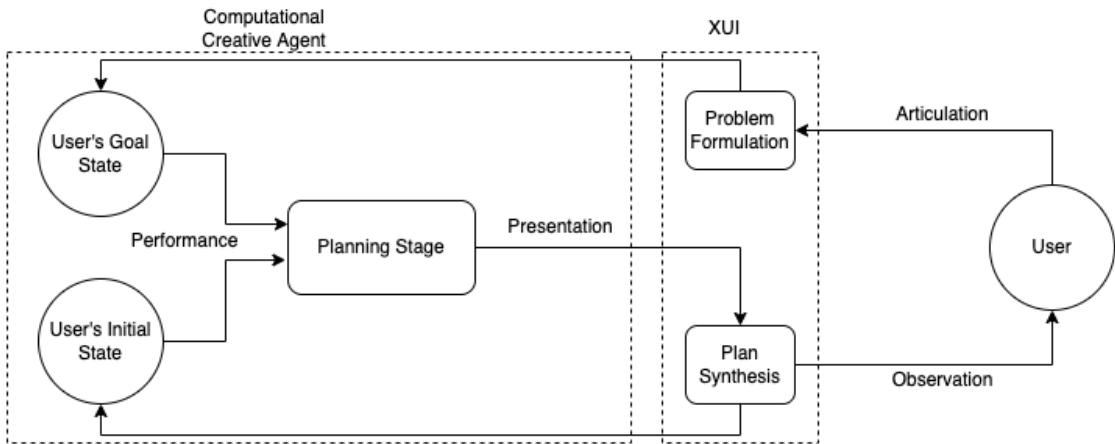


Figure 1: A computational model of an explanation using Abowd's Interaction Framework, concepts from XUI, and XCC.

3. Progress Summary

The first research project targeting this line of work used CBR within a co-creative agent to assist musicians in their aesthetic goals through a vocal audio plugin. Results showed that although participants were interested in using a co-creative agent throughout the production process, they acted against the vocal plugin parameter recommendations set by the agent. Participants showed frustration when the co-creative agent acted in a way that deviated from set expectations. From this research, we posit that explainability is an essential aspect of effective CBR models within co-creative agents.

Our next goal is to assess the various levels of explainability aforementioned in the context of the Four P's. This will involve at least four projects focusing on each P individually and evaluating the explainability based on user interactions.

References

- [1] M. T. Llano, M. d'Inverno, M. Yee-King, J. McCormack, A. Ilsar, A. Pease, S. Colton, Explainable computational creativity., in: ICCC, 2020, pp. 334–341.
- [2] J. Zhu, A. Liapis, S. Risi, R. Bidarra, G. M. Youngblood, Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation, in: IEEE CIG, 2018, pp. 1–8.
- [3] N. Bryan-Kinns, B. Banar, C. Ford, C. Reed, Y. Zhang, S. Colton, J. Armitage, et al., Exploring xai for the arts: Explaining latent space in generative music (2022).
- [4] A. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in explainable ai, arXiv preprint arXiv:1810.00184 (2018).
- [5] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (xai): Toward medical xai, *IEEE trans. on Neural Networks and Learning sys.* 32 (2020) 4793–4813.
- [6] M. Rhodes, An analysis of creativity, *The Phi Delta Kappan* 42 (1961) 305–310.
- [7] C. Lamb, D. G. Brown, C. L. Clarke, Evaluating computational creativity: An interdisciplinary tutorial, *ACM CSUR* 51 (2018) 1–34.
- [8] J.-m. Choe, The consideration of cultural differences in the design of information systems, *Information & Management* 41 (2004) 669–684.
- [9] T.-F. Kummer, J. M. Leimeister, M. Bick, On the importance of national culture for the design of information systems, *B & I Systems Engineering* 4 (2012) 317–330.
- [10] S. Abdellahi, M. L. Maher, S. Siddiqui, Arny: A co-creative system design based on emotional feedback., in: ICCC, 2020, pp. 81–84.
- [11] R. K. Sawyer, Group creativity: Music, theater, collaboration, *Psychology Pr.*, 2014.
- [12] U. X. Eligio, S. E. Ainsworth, C. K. Crook, Emotion understanding and performance during computer-supported collaboration, *Comp. in Human Beh.* 28 (2012) 2046–2054.
- [13] I. Leite, C. Martinho, A. Paiva, Social robots for long-term interaction: a survey, *Int'l J. of Social Robotics* 5 (2013) 291–308.
- [14] L. Henrickson, Tool vs. agent: attributing agency to nlgs, *Dig. Creativity* 29 (2018) 182–190.
- [15] A. Jordanous, A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative, *Cog. Comp.* 4 (2012) 246–279.
- [16] P. Karimi, K. Grace, M. L. Maher, N. Davis, Evaluating creativity in computational co-creative systems, arXiv preprint arXiv:1807.09886 (2018).
- [17] H. R. Hartson, T. S. Andre, R. C. Williges, Criteria for evaluating usability evaluation methods, *Int'l J. of HCI* 13 (2001) 373–410.
- [18] J. S. Gero, U. Kannengiesser, The situated function–behaviour–structure framework, *Design stud.* 25 (2004) 373–391.
- [19] T. Miller, P. Howe, L. Sonenberg, Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences, arXiv (2017).
- [20] D. Alvarez Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, *NeurIPS* 31 (2018).
- [21] S. Palacio, A. Lucieri, M. Munir, S. Ahmed, J. Hees, A. Dengel, Xai handbook: Towards a unified framework for explainable ai, in: IEEE/CVF, 2021, pp. 3766–3775.
- [22] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Mach. Intel.* 1 (2019) 206–215.
- [23] G. D. Abowd, Formal aspects of HCI, University of Oxford Oxford, 1991.

Case-based Explanation for Black-Box Time Series and Image Models with Applications in Smart Agriculture

Eoin Delaney^{1,2,3,*†}

¹School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland

²VistaMilk SFI Research Centre, Dublin, Ireland

³Insight Centre for Data Analytics, Dublin, Ireland

Abstract

Black-box models are frequently deployed for high stakes prediction tasks in a variety of domains (e.g., disease diagnosis and agricultural prediction). The predictions of these opaque systems are often plagued by a lack of transparency, motivating novel research in eXplainable AI (XAI) aiming to understand why a certain prediction was made. One increasingly promising form of explanation is counterfactual explanation where the aim is to elucidate how a prediction could change, given some change in the input space. While the majority of existing work has focused on producing counterfactual explanations for tabular data, significantly less focus has been placed on generating and evaluating counterfactual explanations for time series and image data. Explaining predictions for these data types, arguably, presents a whole new set of issues for XAI, due to the complex and multi-dimensional nature of the data. In this research, we examine how leveraging case-based reasoning (CBR) techniques such as Nearest-Unlike-Neighbors (NUNs) can aid the generation and evaluation of explanations in these domains. We also demonstrate the inadequacies of many traditional techniques that are used to evaluate explanations and highlight the promise of CBR and user studies in the evaluation of explanations.

Keywords

Explainable AI, Counterfactual, Time Series, Prefactual, XCBR, Smart-Agriculture, User Study

1. Introduction

In recent years, the predictive prowess of machine learning systems has been undermined by a worrying lack of interpretability, fairness, accountability and transparency [1, 2]. These challenges have resulted in major research efforts in Explainable AI (XAI) where the core objective is to offer insights into the predictions of black-box models that are commonly deployed in high stakes scenarios. One such scenario that is of particular interest to our research is in smart agriculture. Previous CBR research has already shown immense promise in both grass growth and grasshopper infestation prediction [3, 4]. While the majority of XAI research focus has been on tabular data, less attention has been attributed to time series data, introducing a new set of complex issues for XAI due to high data dimensionality and strong feature dependencies [5].

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France

*Corresponding author.

✉ eoin.delaney@insight-centre.org (E. Delaney)

🌐 <https://e-delaney.github.io/> (E. Delaney)

🆔 0000-1111-2222-3333 (E. Delaney)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

A variety of eXplainable CBR (XCBR) methods have shown immense promise for XAI (see [6] for a review). These XCBR techniques provide factual, example-based explanations (e.g., [7, 8]), feature-weighting explanations (CBR-LIME; [9]), and counterfactual explanations [10] with a focus typically on tabular and sometimes image data.

Counterfactual explanations aim to elucidate how a prediction could change if some input was different. There is growing evidence from psychology, philosophy and sociology indicating that they provide more human friendly and GDPR compliant explanations in comparison to other popular forms of explanations [11, 12, 10]. While there are over 100 techniques proposed to generate counterfactual explanations [13], very few of these methods focus on image data, and even fewer on time series data (see e.g., [14, 15] for closest works). In a similar fashion, it is unclear if the proposed properties of *good* counterfactual explanations for tabular data such as proximity, sparsity, and plausibility [12, 10] will extend to other data types. Moreover, evaluating these properties is non trivial and there is growing evidence to suggest that user studies are desperately needed in order to reliably evaluate explanations [16, 10].

2. Research Plan and Objectives

The overall goal of this research is to develop techniques that can be used to generate and evaluate explanations for time series and image data through leveraging case-based reasoning. Building on evidence from psychology, philosophy and social science [11, 12, 13], a core focus of this research is in the generation and evaluation of counterfactual explanations.

I have identified several research questions that underpin the goal of generating and evaluating counterfactual explanations for time series and image data;

- Can case-based reasoning be leveraged to generate *good* explanations for applied time series prediction tasks both in terms of (i) counterfactual explanations for classification and (ii) explanations for applied agricultural forecasting problems?
- What are the properties of good counterfactual explanations for time series and image data, and do they mirror the properties of good counterfactual explanations for tabular data (e.g., proximity, sparsity and plausibility)?
- Do explanations that are automatically generated by computational techniques align with explanations that are informative for human users?

In previous work, Keane and Smyth [10] designed a novel case based technique to generate counterfactual explanations for tabular data through leveraging existing counterfactual instances in the training data (i.e., nearest unlike neighbors (NUNs) [17]). So, exploring the role of NUNs in the generation of counterfactual explanations is a promising line of research in the context of time series and image data. The combination of CBR with Deep Learning feature weighting techniques (e.g., class activation mapping [18]) in a Twin-Systems framework [19] is another promising area of research for the development of counterfactual explanations for time series and image data. Feature weighting techniques are perhaps the most common XAI method in time series classification [5], and the availability of open source data on the UCR archive [20] readily facilitates the development and experimental comparison of XAI techniques.

In terms of time series forecasting, one untapped line of work from our review of the psychological literature is in *prefactual explanation*. Prefactual explanations describe conditional (if-then) propositions about, as yet not undertaken, actions and the corresponding outcomes that may (or may not) take place in the future [21, 22]. While counterfactuals focus on the past, prefactuals look to the future, capturing the idea of something that is not yet a fact, but could become a fact [21]. Such explanations could also be leveraged in other challenging domains such as reinforcement learning [23].

One applied area that is of particular interest to our research is in smart and sustainable agriculture. We have a data set from an industry partner containing information about milk yield from over 2000 commercial dairy herds. One of our goals is to accurately provide long term milk supply forecasts to farmers, supplementing the predictions with explanations that indicate different actions they could take to boost milk yield in future years. Related CBR work in goal-based recommendation has shown how different training plans can be recommended to runners to produce new personal best times [24], so relating this CBR research to producing prefactual explanations for farmers to improve their output is a promising line for novel research.

Finally, it is unclear if the properties of good explanations for tabular data will extend to time series and image data. For example, when generating explanations one popular technique is to minimize the distance between the query and the counterfactual [12]. However, this runs the risk of generating adversarial explanations that may not be noticeably different for users in relation to the query instance. In time series and image data, discriminative and semantically meaningful information is often contained in localized regions of the time series or image. So, it is clear that user evaluation and rigorous testing of explanation evaluation metrics are needed in this research.

3. Progress Summary

We developed a novel CBR technique, *Native-Guide*, to generate counterfactual explanations for time series classification tasks [5]. The technique leverages both in-sample counterfactual explanations (e.g., Nearest Unlike Neighbors [17, 10]) and feature weight vectors from techniques such as class-activation mapping [18] to create explanations. This work was presented at ICCBR'21 where it received a best-student paper award. More recently, we developed a novel forecasting technique and a method to provide prefactual explanations with applications in milk supply prediction [25]. Specifically, we highlighted how producing explanations through comparatively contrasting high performing exemplar herds and low performing herds (retrieved using class prototypes) could boost future on-farm performance - "*Your projected milk supply for next year is 250'000 litres. However, if you reduced the calving period (In a similar fashion to farmer Y), your projected supply would be 300'000 litres and your milk would likely have a higher protein content*". This work will appear in the main proceedings of ICCBR'22.

In terms of counterfactual evaluation, we conducted a literature review of over 100 papers and discussed five key deficits to rectify in the evaluation of counterfactual XAI techniques. This review paper was presented at IJCAI'21 [13]. We noted the over-reliance on computational proxy measures for proximity, sparsity and plausibility without any conclusive evidence from user studies. In work presented at the ICML Workshop on Algorithmic Recourse we identified

the utility of case-based evaluation methods in determining how well a counterfactual fit the data distribution, and highlighted that optimizing for proximity often generated adversarial explanations that would not be noticeably different than the query for a human user [26]. Results in our work in time series classification also demonstrated similar results [5]. So, a natural avenue for current and future work is to focus on evaluating counterfactual explanations through conducting user studies.

Currently we are focusing on addressing some of the central issues presented in our IJCAI review paper and we are conducting large scale user studies to evaluate explanations and critically assess the suitability of computational evaluation techniques. In our latest experiments human users ($N=42$) created counterfactuals through correcting misclassifications of a convolutional neural network on the MNIST and Google Quickdraw data sets using a drawing tool. This data represents the first ground truth explanation data set for counterfactual visual explanations. By comparing explanations generated by humans with those that are generated automatically by computational techniques, we aim to provide novel insights into (i) the properties of good explanations according to humans and (ii) the unreliability of many popular evaluation metrics (e.g., L_1 and L_2 distances for proximity). Contrary to popular belief, our initial results indicate that people do not minimally edit instances when creating counterfactual visual explanations. Instead they modify a larger, and often semantically meaningful region when creating an explanation, often pushing the explanation towards a class prototype. So, leveraging psychologically grounded models of similarity such as Tversky's contrast model of similarity [27] in counterfactual generation and evaluation may result in more informative explanations and is an interesting avenue for future work.

References

- [1] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, *AI Magazine* 40 (2019) 44–58.
- [2] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI), *IEEE Access* 6 (2018) 52138–52160.
- [3] E. M. Kenny, E. Ruelle, A. Geoghegan, L. Shalloo, M. O’Leary, M. O’Donovan, M. T. Keane, Predicting grass growth for sustainable dairy farming: A cbr system using bayesian case-exclusion and post-hoc, personalized explanation-by-example (xai), in: International Conference on Case-Based Reasoning, Springer, 2019, pp. 172–187.
- [4] J. Hastings, K. Branting, J. Lockwood, Carma: A case-based rangeland management adviser, *AI Magazine* 23 (2002) 49–49.
- [5] E. Delaney, D. Greene, M. T. Keane, Instance-based counterfactual explanations for time series classification, in: International Conference on Case-Based Reasoning, Springer, 2021, pp. 32–47.
- [6] J. M. Schoenborn, R. O. Weber, D. W. Aha, J. Cassens, K.-D. Althoff, Explainable case-based reasoning: A survey, in: AAAI-21 Workshop Proceedings, 2021.
- [7] M. T. Keane, E. M. Kenny, How case-based reasoning explains neural networks: A theoretical analysis of xai using post-hoc explanation-by-example from a survey of ann-cbr twin-systems, in: Proc. ICCBR’19, Springer, 2019, pp. 155–171.

- [8] F. Sørmo, J. Cassens, A. Aamodt, Explanation in case-based reasoning—perspectives and goals, *Artificial Intelligence Review* 24 (2005) 109–143.
- [9] J. A. Recio-García, B. Díaz-Agudo, V. Pino-Castilla, CBR-LIME: A Case-Based Reasoning Approach to Provide Specific Local Interpretable Model-Agnostic Explanations, in: ICCBR, Springer, 2020, pp. 179–194.
- [10] M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), in: Proc. ICCBR'20, Springer, 2020, pp. 163–178.
- [11] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* 267 (2019) 1–38.
- [12] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the gdpr, *Harv.J.Law Tech.* 31 (2017) 841.
- [13] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques, in: IJCAI-21, 2021.
- [14] E. Ates, B. Aksar, V. J. Leung, A. K. Coskun, Counterfactual explanations for machine learning on multivariate time series data, *arXiv preprint arXiv:2008.10781* (2020).
- [15] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, Counterfactual visual explanations, in: ICML, PMLR, 2019, pp. 2376–2384.
- [16] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, *arXiv preprint arXiv:1702.08608* (2017).
- [17] C. Nugent, D. Doyle, P. Cunningham, Gaining insight through case-based explanation, *Journal of Intelligent Information Systems* 32 (2009) 267–295.
- [18] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE CVPR, 2016, pp. 2921–2929.
- [19] E. M. Kenny, M. T. Keane, Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ann-cbr twins for xai, in: IJCAI-19, 2019, pp. 2708–2715.
- [20] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, E. Keogh, The UCR time series archive, *IEEE/CAA Journal of Automatica Sinica* 6 (2019) 1293–1305.
- [21] K. Epstude, A. Scholl, N. J. Roese, Prefactual thoughts: Mental simulations about what might happen, *Review of General Psychology* 20 (2016) 48–56.
- [22] R. M. Byrne, S. M. Egan, Counterfactual and prefactual conditionals., *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 58 (2004) 113.
- [23] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong, Interpretable machine learning: Fundamental principles and 10 grand challenges, *Statistics Surveys* 16 (2022) 1–85.
- [24] B. Smyth, P. Cunningham, A novel recommender system for helping marathoners to achieve a new personal-best, in: Proceedings of the eleventh ACM conference on recommender systems, 2017, pp. 116–120.
- [25] E. Delaney, D. Greene, L. Shalloo, M. Lynch, M. T. Keane, Forecasting for sustainable dairy produce: Enhanced long-term, milk-supply forecasting using k-nn for data augmentation, with prefactual explanations for xai., in: To appear in ICCBR'22, 2022.

- [26] E. Delaney, D. Greene, M. T. Keane, Uncertainty estimation and out-of-distribution detection for counterfactual explanations: Pitfalls and solutions, arXiv preprint arXiv:2107.09734 (2021).
- [27] A. Tversky, Features of similarity., *Psychological review* 84 (1977) 327.

Explaining and Upsampling Anomalies in Time-Series Sensor Data

Craig Pirie^{1,2,*†}

¹*Robert Gordon University, Garthdee House, Garthdee Road, Garthdee, Aberdeen, AB10 7AQ, Scotland, UK*

Abstract

My research aims to improve anomaly detection methods in multi-sensor data by extending current re-sampling and explanation methods to work in a time-series setting. While there is a plethora of literature surrounding XAI for tabular data, the same cannot be said for the multivariate time-series settings. It is also known that selecting an optimal baseline for attribution methods such as integrated gradients remains an open research question. Accordingly, I am interested to explore the role of Case-Based Reasoning (CBR) in three ways: 1) to represent time series data from multiple sensors to enable effective anomaly detection; 2) to create explanation experiences (explanation-baseline pair) that can support the identification of suitable baselines to improve attribution discovery with integrated gradients for multivariate time-series settings; and 3) to represent the disagreements between past explanations in a case-base to better inform strategies for solving disagreement between explainers in the future. A common theme across my research is the need to explore how inherent relationships between sensors (causal or other ad-hoc inter-dependencies) can be captured and represented to improve anomaly detection and the follow-on explanation phases.

Keywords

anomaly detection, time-series, negative sampling, integrated gradients

1. Introduction

The advent of the Internet of Things (IoT) has empowered connected technology with new capabilities¹ [6]. One application of IoT is for the smart-monitoring of sensors to detect and predict failures. This is called anomaly detection and is an important problem across many domains. Its purpose is to identify patterns in data that lie outwith the realms of expectation [1]. It has many applications, including in fraud detection [10], medical analysis [3], industrial settings [17] and in sensor networks [8]. In the context of sensors, anomaly detection refers to the problem of identifying faulty or damaged sensors. This can rely on *univariate* (one sensor) or *multivariate* (many sensors) data. It is hypothesised that in real-world industrial settings, multivariate data is advantageous for anomaly detection. This is because it allows for the capturing of the hidden inter-dependencies between sensors which may improve the

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France

*Corresponding author.

✉ c.pirie11@rgu.ac.uk (C. Pirie)

🌐 <https://www.linkedin.com/in/craig-pirie-aberdeen/> (C. Pirie)

>ID 0000-0002-6799-0497 (C. Pirie)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹IoT refers to a network consisting of sensory, communication, networking and information-processing technologies [6].

performance of anomaly detection systems. For example, a key multi-variate time-series dataset that will be used for this project includes the smart-buildings dataset². It contains features such as **zone_air_cooling_setpoint** and **zone_air_temperature_sensor** that when evaluated on their own may be difficult to identify an anomaly. However, when considered in relation to one-another, the anomaly becomes more obvious (you would expect the two values not to be wildly dissimilar).

CBR has been used successfully with time-series problems [9] where there is a need to represent one or more data streams to enable decision support. Existing representation techniques for time-series data often make use of feature extraction (e.g. moving average) or feature transformation (e.g. DCT) methods. Common to these problems are the lack of human-annotated data or limited access to instances or a serious class imbalance as is expected with anomaly detection. By definition, anomalous instances are rare. Ergo, it is natural that the number of recorded normal instances far outweighs that of the abnormal class (see Figure 1). This can cause bias and be problematic for deep learning methods. Usually this leads to a pre-processing step to bring balance to the data prior to the learning phase. This can involve up-sampling the negative class at the cost of introducing artificial data into the set, or down-sampling the positive class at the cost of discarding valuable data. Class imbalance in time-series settings is particularly difficult due to the temporal connection between instances. The time dimension must be considered when sampling to allow the learning of time-series models. ‘Negative Sampling’ [16] is one method of up-sampling the anomalous class. It is based on the ‘Concentration Phenomenon’ and artificially inflates the feature space by applying a $\pm\delta$, to each of the features of samples in the normal space. However, data points are shuffled through time as the application of the δ is a random process and the time-dimension is not carefully considered. I wish to extend this method to work in the time-series setting. One approach I wish to take is to learn a sequence embedding of the feature space on which I will perform the re-sampling on prior to decoding. This way the temporal information is maintained and time-series methods may be applied.

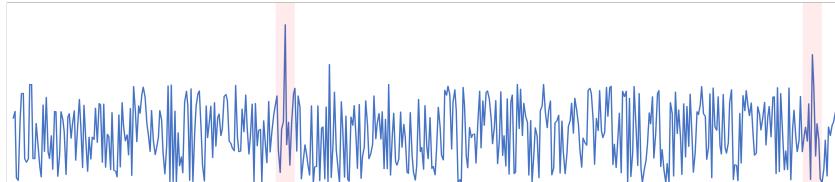


Figure 1: The figure shows a typical time-series representation in anomaly detection. Areas shaded in red are anomalies. This highlights the scarcity of abnormal instances compared to normal instances.

Before anomaly detection models can be deployed in production, we must learn to trust them. However, with the black-box nature of many deep neural architectures, this can be challenging. Confidence can be improved with Explainable AI (XAI). Through XAI, we can gain better insights around the ‘decision-making process’ of a deep-learning model. When we know why a decision is being made, we find it easier to trust that the decision was made on the correct grounds. Many strategies for explaining time-series re-purpose existing XAI methods (such as LIME [11] or SHAP [7]) that were originally used to explain models in other areas such as in

²The smart buildings dataset can be found here.

computer vision or natural language processing [12] and there is a lack of methods specifically designed for time-series. More so, they are often difficult to interpret and unsophisticated [14]. It has also been found that some saliency-based methods, that were designed for tabular data, fail when applied to sequence data [4, 2]. This can lead to anomalies being explained individually at a particular time-stamp, rather than collectively over a segment of a time-series graph. Increasingly, Case-based Reasoning (CBR) has seen application to facilitate XAI [15], giving rise to XCBR. Delaney *et al.* present a counterfactual-based XCBR approach to explaining time-series anomalies [2]. This works well for uni-variate data but interpretability suffers as the number of features increases. Several surveys have been conducted on the state-of-the-art in time-series explanation and there is a clear consensus that there is a need for more sophisticated methods that are dedicated to time-series applications [12, 13, 2].

Integrated gradients [19] is a popular XAI technique that does not modify the original network to provide explanations. It determines feature importance by evaluating the gradient between the input and output along uniform steps through the feature space. For this, it requires a baseline (often an all-zero embedding such as a black image). Alas, selection of a good baseline is problematic due to the *missingness* problem [18]. “Missingness” is a concept that is well-defined in game theory. It originally referred to the concept of determining how much value a group of participants added to a game by evaluating the value of the game after gradually adding more participants. This is the idea behind the baseline — to model the absence of the feature we are trying to evaluate. The problem arises when the same data resides in both the baseline *and* the sample because it becomes impossible to gradually introduce to determine its importance. In other words, $x_i - x'_i$ (the difference between baseline and output) will always be 0, meaning it can never be deemed important (see Figure 2 for an example). Therefore, it is essential that the baseline shares minimal information with the output image. There have been multiple strategies used to try accomplish this but none are perfect and the topic remains an open research area [18].

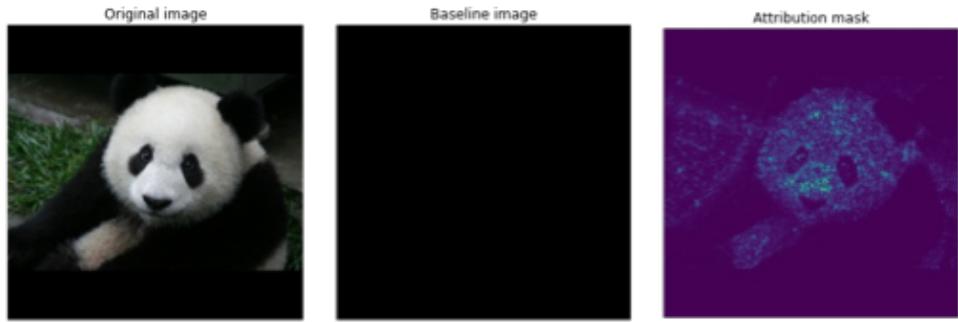


Figure 2: The original image (left) is of a giant panda and the baseline image (center) is a constant black image. The attribution mask (right) shows the pixels deemed important by integrated gradients by highlighting them. The figure demonstrates the missingness problem as the black features of the panda are deemed unimportant despite being a prominent characteristic.

To further ensure trust, often an ensemble approach is taken to explain models, meaning that explanations from multiple different explainers are used. This introduces the ‘Disagreement Problem’ [5] where it is common for many different explainers to produce wildly different

explanations for the same decision. Ironically, this can further the distrust in systems — exactly the opposite of its purpose. There is a need for tackling this problem so that multiple explainers can be made to agree on an explanation. I will explore if this problem can be tackled by learning from past experiences, in a case-based manner. Through this, it is hoped that past experiences in solving explanation disputes can be used to solve disagreements in the future for similar problems. The case-base would consist of a model, its output and the different explanations as the query; and the strategy used to solve the disagreement as the solution.

Lastly, I wish to take a similar approach to learn the relational information in the feature space. Often, sensors in industrial domains do not work in solitude. There may be obvious, physical inter-dependencies between sensors or they may be hidden. For example, as humidity is proportionally affected by temperature, spikes in temperature readings not seen in humidity readings may indicate a broken temperature sensor (or vice-versa). By learning an embedding of the relational information I can capture these inter-dependencies when sampling in the encoded space. Consequently, it is hoped this additional information will support the classification ability of anomaly detection systems.

2. Research Plan

2.1. Research Objectives

The main aim of this project is to improve anomaly detection in industrial sensor data and its explanations by extending current methods to work with time-series data. Through this I will be examining how to represent sequence data (such as time-series) to enable re-sampling methods needed to improve coverage for anomaly detection. This will involve experiments with feature embeddings that capture temporal and relational information between sensors. I will also explore explainer aggregation strategies to address the Disagreement Problem in the time-series context. As such, the following research questions are defined:

- **RQ1:** How can sequence data (such as time-series) be represented to enable negative sampling methods that capture the temporal and relational information between sensors needed to improve anomaly detection?
- **RQ2:** Can local methods such as a case-based approach to selecting a baseline for Integrated Gradients improve the quality of its explanations?
- **RQ3:** What explainer aggregation strategies can be used to address the explainer Disagreement Problem in the context of time-series data for anomaly detection?

2.2. Approach / Methodology

I wish to conduct experiments with applying Negative Sampling on an embedding of the feature space that captures the temporal information and inter-dependencies between sensors. This may involve the use of auto-encoders or sequence-to-sequence auto-encoders to learn the embedding. It is hoped that incorporating this data will improve the performance anomaly classifiers.

Currently I propose to expand on Integrated Gradients for the time-series setting by deploying a case-based approach to selecting baselines. To do this, a query (time-series window, image

etc.) and baseline pair will be stored in a case base. Similar images contain similar information, therefore, in theory, should react well to similar baselines. To use an image-based example, horses are very similar animals to zebras, so it is hoped that a baseline that provides good explanations for horses, can do so for zebras. For time-series, similar trends or sub-sequences will use similar baselines. This should circumvent the ‘missingness problem’ but may require an adaptation step to further ensure the missingness characteristic is upheld.

Finally, I will also explore the efficacy of case-based reasoning to solve the ‘Disagreement Problem’ in ensemble explainers. The aim is to establish a case-base of ‘disagreements’ and solutions (past strategies that were used to solve the disagreement).

3. Progress Summary

The research is still in its embryonic stages and we are still tweaking the research proposal. We have identified some key papers which we are reviewing in depth to better understand the problem and validate the need for our research. Accompanying this, some exploratory data analysis is being conducted in parallel with the hopes to replicate the results in studies found in the literature. At present, the main focus is evaluating negative sampling as a method to correct class imbalance and determining its suitability for time-series data.

References

- [1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. “Anomaly detection: A survey”. In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.
- [2] Eoin Delaney, Derek Greene, and Mark T Keane. “Instance-based counterfactual explanations for time series classification”. In: *International Conference on Case-Based Reasoning*. Springer. 2021, pp. 32–47.
- [3] Milos Hauskrecht et al. “Evidence-based anomaly detection in clinical domains”. In: *AMIA Annual Symposium Proceedings*. Vol. 2007. American Medical Informatics Association. 2007, p. 319.
- [4] Aya Abdelsalam Ismail et al. “Benchmarking deep learning interpretability in time series predictions”. In: *Advances in neural information processing systems* 33 (2020), pp. 6441–6452.
- [5] Satyapriya Krishna et al. “The Disagreement Problem in Explainable Machine Learning: A Practitioner’s Perspective”. In: *arXiv preprint arXiv:2202.01602* (2022).
- [6] Shancang Li, Li Da Xu, and Shanshan Zhao. “The internet of things: a survey”. In: *Information systems frontiers* 17.2 (2015), pp. 243–259.
- [7] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30 (2017).
- [8] Luis Martí et al. “Anomaly detection based on sensor data in petroleum industry applications”. In: *Sensors* 15.2 (2015), pp. 2774–2797.
- [9] Stewart Massie et al. “Monitoring Health in Smart Homes using Simple Sensors”. In: () .

- [10] Tahereh Pourhabibi et al. “Fraud detection: A systematic literature review of graph-based anomaly detection approaches”. In: *Decision Support Systems* 133 (2020), p. 113303.
- [11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “” Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.
- [12] Thomas Rojat et al. “Explainable artificial intelligence (xai) on timeseries data: A survey”. In: *arXiv preprint arXiv:2104.00950* (2021).
- [13] Udo Schlegel et al. “An empirical study of explainable AI techniques on deep learning models for time series tasks”. In: *arXiv preprint arXiv:2012.04344* (2020).
- [14] Udo Schlegel et al. “Towards A Rigorous Evaluation Of XAI Methods On Time Series”. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2019, pp. 4197–4201. doi: 10.1109/ICCVW.2019.00516.
- [15] Jakob M Schoenborn et al. “Explainable case-based reasoning: a survey”. In: *AAAI-21 Workshop Proceedings*. 2021.
- [16] John Sipple. “Interpretable, multidimensional, multimodal anomaly detection with negative sampling for detection of device failure”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9016–9025.
- [17] Ljiljana Stojanovic et al. “Big-data-driven anomaly detection in industry (4.0): An approach and a case study”. In: *2016 IEEE international conference on big data (big data)*. IEEE. 2016, pp. 1647–1652.
- [18] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. “Visualizing the Impact of Feature Attribution Baselines”. In: *Distill* (2020). <https://distill.pub/2020/attribution-baselines>. doi: 10.23915/distill.00022.
- [19] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic Attribution for Deep Networks”. In: *CoRR* abs/1703.01365 (2017). arXiv: 1703.01365. URL: <http://arxiv.org/abs/1703.01365>.

Addressing Trust and Mutability Issues in XAI utilising Case Based Reasoning

Pedram Salimi^{1,*}

¹*Robert Gordon University, Garthdee House, Garthdee Rd, Garthdee, AB10 7AQ, Aberdeen, United Kingdom*

Abstract

Explainable AI (XAI) research is required to ensure that explanations are human readable and understandable. The present XAI approaches are useful for observing and comprehending some of the most important underlying properties of any Black-box AI model. However, when it comes to pushing them into production, certain critical concerns may arise: (1) How can end-users rely on the output of an XAI platform and trust the system? (2) How can end-users customise the platform's output depending on their own preferences In this project, we will explore how to address these concerns by utilising Cased-based Reasoning. Accordingly, we propose to exploit the neighbourhood to improve end-user trust by offering similar cases and confidence scores and using different retrieval strategies to address end-user preferences. Additionally, this project will also look at how to leverage Conversational AI and Natural Language Generation approaches to improve the interactive and engaging user experience with example-based XAI systems.

Keywords

Explainable AI, Cased-based Reasoning, Conversational AI, Natural Language Generation

1. Introduction

Due to recent breakthroughs in Artificial intelligence (AI) such as deep learning approaches, AI models are getting more accurate and powerful while also becoming more complicated [1]. However, because of their complexity, comprehending how these models work and making judgments has proven difficult. Earlier AI systems were build on approaches that are fundamentally explainable (i.e. white-box) where Rule-based methods, Decision Trees, Hidden Markov Models, and Logistic Regressions are some examples. Thanks to recent breakthroughs, novel AI techniques such as deep learning are more accurate and powerful than traditional approaches; however, they are also more complicated [1, 2]. Due to complexity of the models (i.e. black-box), comprehending how they work and make decision is difficult. Consequently they reduce model explainability.

Accordingly, there is an armory of Explainable AI (XAI) methods developed in recent literature to explain black-box AI models and the decisions they make. Example-based explanations assist people in developing mental models of the machine learning method and the data on which the machine learning method was trained[3]. Literature shows that humans tend to provide contrastive explanations when explaining their decisions to one another. Accordingly, explaining an AI decision using counterfactual examples can be most understandable to humans

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France

✉ p.salimi@rgu.ac.uk (P. Salimi)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

because they both have the same conceptual model as human explanations [4]. Therefore, In this project, we are focusing on example-based XAI and we propose to use CBR approaches to address two existing challenges in this domain.

Example-based XAI systems, similar to other methods have been good at explaining the current user problem. Also, they are able to guide the users to solve their problems. However, there are several limits to these XAI methods. In this project we are focusing on two of these limitations. One of them is that current approaches are static, which means they provide explanations based on the user query but cannot react to user modifying the query based on their own personal preferences [5]. There is also a lack of trust between the user and the XAI system that is yet to be addressed. User studies by [4] and [6] are few those who highlight this issue of trust in XAI system with respect to applications in speech recognition, forest coverage prediction and recidivism.

Case-Based Reasoning (CBR) is a methodology that emulate how humans reason from precedent and examples and it has a central role in XAI systems [7]. CBR has been the underpinning of many example-based XAI methods providing explanations ranging from factual to counterfactuals [8]. Accordingly, we ask the following research questions:

- **RQ1:** How can we approach the issue of trust in an interactive example-based XAI system using the CBR system?
- **RQ2:** How can a case-based approach assist us in dealing with mutability of features when generating counterfactual explanations?

2. Background

In this section, we will first study example-based XAI approaches before briefly discussing the CBR methodology.

2.1. Example-based Explanations

Example-based approaches are classified into three categories: factual, semi-factual, and counterfactual.

Factual Explanations provide information about why a certain outcome was received based on query features [9]. Using nearest neighbours is an example-based approach for finding factual explanations. For example, in loan application a factual explanation using nearest neighbors could be “Your loan got rejected because there is another person whose situation is quite similar to you and had their loan declined”. Explanations-by-example is a factual explainer algorithm where nearest-neighbours are found using Critical Classification Regions in images[10].

Semi-factual Explanations present the maximum distance an instance may go without changing the black-box outcome. A semi-factual explanation for a reject loan application would be, “Even if the installment amount is increased, loan would be still rejected”. PIECE is a case-based method for generating semi-factual explanations which uses a convolutional model to detect important features and to generate semi-factual explanations [11].

Counterfactual Explanations define a causal, synthetic or past event with the smallest change in feature values that causes the prediction to shift to a desired outcome. A counterfactual explanation for a rejected loan application would be, “If the loan amount is reduced, loan would have been accepted”. Some of the state-of-the-art counterfactual methods are as follows:

- **NICE** approach is divided into two steps. First, the nearest unlike neighbour (NUN) is retrieved, which leads to the finding of non-overlapping features against the query. Then the algorithm iteratively attempts to determine the optimal counterfactual using a reward function that consider properties like [12].
- **DisCERN** is a case-based counterfactual explanation method. Here, counterfactuals are created by substituting feature values of the query from the NUN until an outcome change is detected. Features to substitute are selected based on feature attributions. [13].

2.2. Case-based Reasoning

Most XAI methods, including ones discussed above, fail to establish trust with the end-user as discussed in the introduction. Their one-shot nature (instead of being interactive) also fail to incorporate user preferences. To address these challenges, in this project, we explore techniques from Case-based Reasoning (CBR). A CBR methodology consists of four stages: retrieve, reuse, revise, and retain [14]. The first stage involves providing an input that describes the present user query and retrieving similar cases in the case base by employing similarity metrics. The second stage utilises retrieved cases and use adaption knowledge to present the user with a solution to their query. The user may accept or reject the solution for a variety of reasons, for example, user could be unable to accept the entire proposed solution, based on their own preferences. In the case of rejection, the following step is to revise. In most cases, the revise stage includes incorporating feedback acquired from testing the suggested solution. During the final step, the new case may be retained in the case base for future use.

3. Approach / Methodology

In this section we are going to explore the research questions identified using CBR techniques.

3.1. RQ1: How to address trust in an example-based XAI system using CBR?

This RQ explore how to establish trust between an XAI system and a user once they are presented with an explanation. Specifically we want to identify what additional information or explanation will help the user to better believe the recommendations provided by an XAI system in terms of reliability of the XAI system. Following CBR approaches may assist us in addressing the trust issues:

- **Nearest Neighbors** may be to retrieve the nearest neighbours of the provided solution which previously were successful examples
- **Coverage** and population density of instances which are similar to the provided counterfactual in a case base[15].

For evaluating the impact of proposed methods on trust, a user study is proposed. This work is informed by XAI evaluation methods like the Hoffman Trust scale when seeking feedback from users [16].

3.2. RQ2: How to address feature mutability preferences using CBR?

The interactive system should allow the user to modify the criteria on which an explanation is generated by considering mutability of features. In another words, mutability in an XAI system is about to giving the control to the user in terms of the degree of complexity and difficulty of what they can change. To address this we explore following techniques from CBR:

- **Collaborative Filtering** It is a mechanism for proposing alternative solutions to a user based on similarities (similarity assessment [17]) in the user's prior behaviour and that of other users.
- **Adaptation** When the best partial-matching case from the case repository does not perfectly match the new case, the previous solution must be altered to fit the new case solution more accurately. There are several adaptation methods such as null adaptation, structural adaptation, or a combination of methods[18].

Figure 1 depicts an example interaction between a user and AI that demonstrates how such a system might deal with trust and mutability issues.

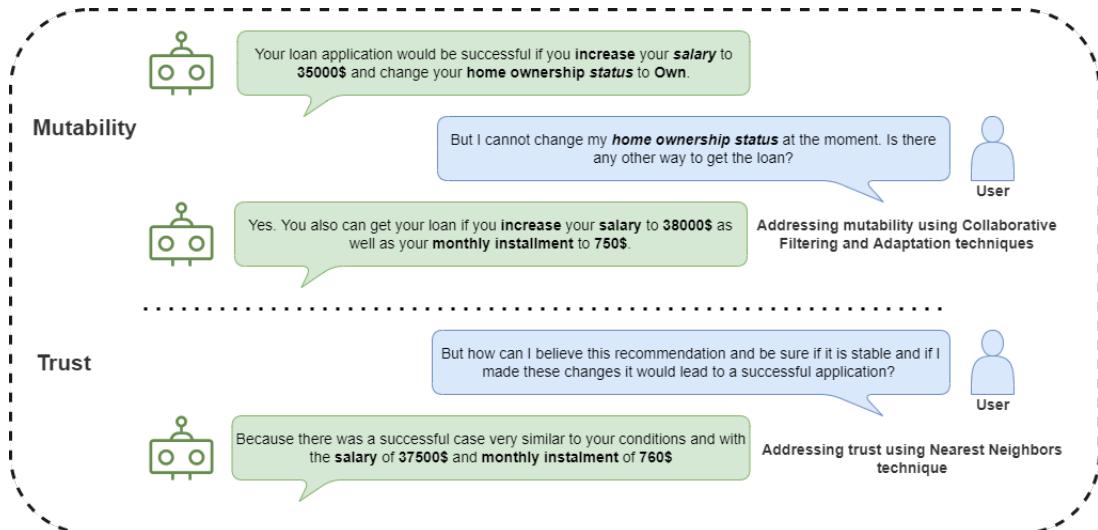


Figure 1: Employing Nearest Neighbors and Collaborative Filtering in order to deal with trust and mutability issues correspondingly.

4. Progress Summary

Recent work has explored several ways for data-to-text generation in terms of numerical reasoning, with the objective of mapping counterfactuals to a natural language representation

which would facilitate a more engaging interaction with the end user. We have designed two distinct template-based text generation algorithms, one with features grouped based on attribution change and the other without. This is illustrated in Figure 2 with two presentations of the counterfactual with and without the feature grouping template for a loan application.

	Loan Amount	Recoveries	Installment	Interest Rate	Home Ownership
Query	2200	1985.75	1200	10.78	OWN
Counterfactual	3700	1310	1700	4.78	RENT
Without Grouping	The loan application would be successful if you increase your loan amount by 1500\$, decrease your recoveries by 676.75\$, increase your installment by 500\$, decrease the interest rate by 6%, and change your home ownership status to rent in the exact order of priorities				
Grouping	The loan application would be successful if you increase your loan amount by 1500\$, your installment by 500\$, and decrease your recoveries by 676.75\$, interest rate by 6%, and change your home ownership status to rent in the exact order of priorities				

Figure 2: Different template based text generation based on feature grouping. This classification is part of the user research to determine which one is more plausible.

Our immediate next task is to design a user study to assess such generation templates and to understand to what extent it could impact end-user engagement and trustworthiness. A questionnaire will be prepared to gather feedback on several counterfactual explanation scenarios. In our user study we are going to consider three framing concept in order to prevent potential biases in our user study[19]. For example, we are going to employ NASA Task Load Index questionnaire. But we are going to modify them with respect to positive framing concept in a manner that instead of asking how much the user got frustrated during the task, we are going to ask how much the task was easy to do. Results of the user study will inform us to identify best template generation strategies in terms of the quality of generated textual explanation and also provide insights for addressing the project's research questions.

Having input from a DC mentor to help improve the user study design will be very valuable; as would directions for integrating case-based strategies for improving user trust.

References

- [1] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller, Explainable AI: interpreting, explaining and visualizing deep learning, volume 11700, Springer Nature, 2019.
- [2] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable ai for natural language processing, arXiv preprint arXiv:2010.00711 (2020).
- [3] C. Molnar, Interpretable machine learning, Lulu. com, 2020.
- [4] X. Wang, M. Yin, Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making, in: 26th International Conference on Intelligent User Interfaces, 2021, pp. 318–328.
- [5] K. Sokol, P. A. Flach, Counterfactual explanations of machine learning predictions: opportunities and challenges for ai safety, SafeAI@ AAAI (2019).
- [6] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, E. André, "do you trust me?" increasing user-trust by integrating virtual agents in explainable ai interaction design, in: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, 2019, pp. 7–9.

- [7] M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), in: International Conference on Case-Based Reasoning, Springer, 2020, pp. 163–178.
- [8] M. T. Keane, E. M. Kenny, M. Temraz, D. Greene, B. Smyth, Twin systems for deepcbr: A menagerie of deep learning and case-based reasoning pairings for explanation and data augmentation, arXiv preprint arXiv:2104.14461 (2021).
- [9] I. Stepin, J. M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.
- [10] E. M. Kenny, E. D. Delaney, M. T. Keane, Advancing nearest neighbor explanation-by-example with critical classification regions (2021).
- [11] E. M. Kenny, M. T. Keane, On generating plausible counterfactual and semi-factual explanations for deep learning, AAAI-21 (2021) 11575–11585.
- [12] D. Brughmans, D. Martens, Nice: an algorithm for nearest instance counterfactual explanations, arXiv preprint arXiv:2104.07411 (2021).
- [13] N. Wiratunga, A. Wijekoon, I. Nkisi-Orji, K. Martin, C. Palihawadana, D. Corsar, Discern: Discovering counterfactual explanations using relevance features from neighbourhoods, in: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2021, pp. 1466–1473.
- [14] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, AI communications 7 (1994) 39–59.
- [15] A. Lawanna, J. Daengdej, Methods for case maintenance in case-based reasoning, International Journal of Computer and Information Engineering 4 (2010) 82–90.
- [16] R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, ArXiv abs/1812.04608 (2018).
- [17] R. Burke, A case-based reasoning approach to collaborative filtering, 2000. doi:10.1007/3-540-44527-7_32.
- [18] S. Craw, N. Wiratunga, R. C. Rowe, Learning adaptation knowledge to improve case-based reasoning, Artificial intelligence 170 (2006) 1175–1192.
- [19] S. Schoch, D. Yang, Y. Ji, “this is a problem, don’t you agree?” framing and bias in human evaluation for natural language generation, in: Proceedings of the 1st Workshop on Evaluating NLG Evaluation, 2020, pp. 10–16.

CBR For Interpretable Response Selection In Conversational Modelling

Malavika Suresh^{1,*†}

¹*Robert Gordon University, Aberdeen, United Kingdom*

Abstract

Current state-of-the-art dialogue systems are increasingly complex. When used in applications such as motivational interviewing, the lack of interpretability is a concern. CBR offers to bridge this gap by using the most similar past cases to decide the outcome for a new problem, which then serves as a natural as well as accurate explanation of the outcome. This research proposes to extend the Abstract Argumentation CBR (AA-CBR) framework for predicting the next response type in an ongoing conversation by reusing the knowledge of previous conversations to achieve a desirable outcome for a new conversation context.

Keywords

Case Based Reasoning, Conversational Modelling, Motivational Interviewing, Abstract Argumentation

1. Introduction

There is recent research interest in automating motivational interviewing² (MI) conversations due to the effectiveness of MI and lack of trained MI interviewers [1]. While some work has shown that large-scale pre-trained language models can be useful to train MI interviewers by predicting responses [2], the model predictions are not explained. In such case, the decision to accept or reject a model's proposed response falls on the trainee. For example, without an explanation of possible outcomes, it is possible that the trainee may reject a model's suitable proposal for a less suited response that they prefer. This is a concern as it is already a difficulty for human MI practitioners to suppress the instinct to respond with premature advice [3]. Equally possible is that an unsuitable model prediction may be accepted by the trainee, which can be avoided when an explanation is available. Additionally, the type of response (i.e dialogue strategy) needs to be adapted based on the individual as the same strategy may result in different outcomes with different individuals. Thus, the interpretable CBR approach of using past cases to decide the outcome for a new case could be better suited for this problem.

CBR approaches to dialogue management have been studied in prior work [4] [5], where each utterance in a dialogue is considered as a case. In contrast, in this work we consider the whole dialogue as a case and propose the use of Abstract Argumentation for CBR (AA-CBR) for selecting next response type. As the framework represents past cases as a tree of arguments attacking and defending each other (i.e an argument graph, as in Fig 1), it can provide natural interactive explanations of the predicted outcomes [6].

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France

✉ m.suresh@rgu.ac.uk (M. Suresh)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

²MI is a form of therapy that encourages people to realize on their own the need for change in attitude/belief

The expected research contributions are - (i) case representation of an MI conversation, (ii) adaptation and extension of AA-CBR for response type selection in MI conversations. These contributions can be extended to any task-oriented conversational model that requires interpretable response generation based on personalized context.

1.1. Background

This section illustrates an example to briefly summarize the AA-CBR framework originally proposed by [7]. AA-CBR is based on the argumentation framework [8] which is basically a set of arguments and a binary *attack* relationship between them which defines whether ArgumentA attacks ArgumentB. AA-CBR represents each case as a set of factors. When adding a new

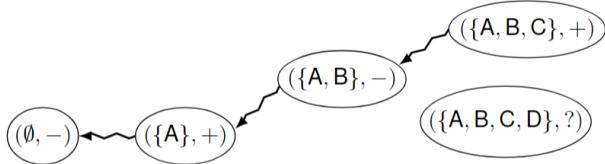


Figure 1: Example of an argument graph using the AA-CBR framework, taken from [7]: Alphabetical characters represent factors; Each node is a case in the case base and may consist of one or more factors and either a positive or negative outcome; The null node represents the default outcome in the absence of any factors; Arrow represents an attack relationship between two cases;

factor to the case changes the case outcome, it is considered an *attack*. Argumentation rules define the attack relationship between cases to then represent the case base as an argument graph. For instance in Fig 1, $\{A, B, C\}$ which is a more specific case (i.e more factors) attacks $\{A, B\}$ which is a less specific case with a different outcome. By inference, $\{A, B, C\}$ defends $\{A\}$ and attacks the default case. Note that $\{A, B, C\}$ does not directly attack the default case because $\{A\}$ already attacks the default case and $\{A\}$ is more concise than $\{A, B, C\}$. The default case defines the assumed outcome for any new case unless the default case is *sufficiently attacked* by other cases in the case base. A *sufficient attack* against the default case occurs if all the unattacked nodes of the graph attack the default case (i.e., none of the unattacked nodes defend the default case).

For a new case, the outcome is decided by first determining which, if any, of the historical cases the new case attacks and subsequently inferring from the argument graph whether the default outcome is attacked or defended. If the default outcome is defended, then the outcome for the new case is the default outcome (in this example, negative). Argumentation rules define that a new case attacks a past case if the past case factors are not contained in the new case³. Here, the new case $\{A, B, C, D\}$ does not attack any case because all historical cases are a subset of the factors of the new case. Since $\{A, B, C\}$ is the closest *unattacked* case in the graph and by inference it attacks the default case, the outcome for the new case is positive.

Argumentation rules are also used to decide the outcome when multiple similar cases with differing outcomes exist in the case base. For instance, in the above example, if the case base

³This ensures that factors which are not present in the new case (and thus deemed irrelevant) do not contribute to the outcome

included a historical case such as $\{\{A,D\}, -\}$, both $\{\{A,D\}, -\}$ and $\{\{A,B,C\}, +\}$ would be similar cases. By inference, the default outcome is now defended by at least one of the unattacked nodes (since $\{A,D\}$ attacks $\{A\}$) and is chosen as the outcome for the new case.

2. Research Plan

This section lists the research objectives, defines associated terminologies and describes the approaches considered. Fig 2 depicts an overview of the components in the research.

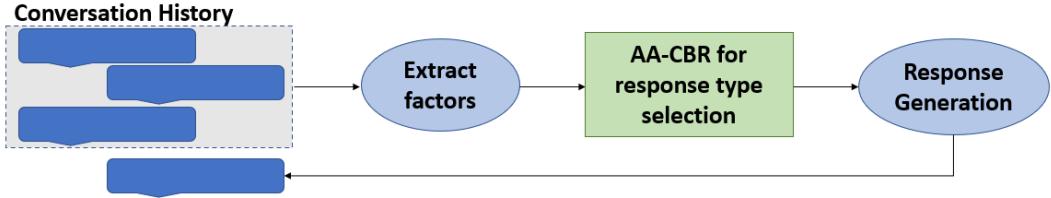


Figure 2: Overview of the research: CBR components in rectangle; Neural network components in oval

The research aims to build an interpretable conversational model for MI. An AA-CBR based approach is proposed to introduce interpretability when deciding the next response type. The following research questions will be investigated through the listed objectives:

1. How can an MI conversation be represented as a case of factors?
 - Identify a set of dialogue factors (i.e case attributes) to represent a given conversation history, which forms the problem component of the case.
 - Label each conversation as successful (good) or unsuccessful (bad), which forms the case outcome.
 - Identify a set of counsellor response types, which forms the solution for the case.
2. How can AA-CBR be applied for case retrieval?
 - Identify challenges in applying AA-CBR for MI conversations
 - Extend AA-CBR for MI conversations
 - Apply the extended AA-CBR framework and evaluate using a sample dataset

Definitions:

Case: A case comprises the entire available conversation history, represented as a set of dialogue factors and depicted as a node in the argument graph.

Dialogue factors: Dialogue factors can capture both relevant content such as the topic of the conversation as well as contextual features such as speaker sentiment and resistance or willingness to change (called MI talk-type). These will be annotated against each utterance.

Outcome: For MI, a good conversation outcome is either an explicit user expression of satisfaction at the end of a conversation or an implicit change in user perspective.

2.1. Approach / Methodology

Construct case base: First, the right set of dialogue factors that capture the separation between good and bad outcomes of an MI conversation need to be identified. Broadly, there are a few types of dialogue factors that may be considered - (i) frame of mind factors (eg. sentiment, LIWC markers [9], MI talk-type [10]) which are indicative of psychological state (ii) conversation topic (eg. addiction, weight-loss) (iii) linguistic factors (eg. utterance length, use of questions). AA-CBR assumes factors to be independent of each other while here, some of them such as sentiment and MI talk-type labels may be related and presence of one factor may entail the other. Such relationships will also need to be investigated. The final set of dialogue factors chosen will form the vocabulary knowledge container of the case base.

The quality of the AA-CBR framework will depend on the quality of the extracted factors and outcome labels. While public datasets such as [10] provide annotations for some factors, other factors may need to be either freshly annotated or predicted. Given the difficulty in obtaining expert annotations, this work will potentially adopt domain-transfer of well-studied models for classifying non-MI factors such as sentiment [11] since words indicating sentiment polarity are reasonably generalizable. Transparency of predictions will be enabled with explanation methods such as feature relevance scores [12]. The models will be trained for use in the continual learning setting [13] so that new data can be used to improve the model throughout its life. The overall approach towards case base construction is summarized as follows:

- Identify and define the dialogue factors to be used
- Annotate cases with factors (manually where expertise is available, or with existing models in literature, or by training a classifier on a few annotated cases)

Extend AA-CBR for response type selection: This work proposes to consider dialogue factors from available conversation history at each turn of a conversation as factors in the AA-CBR framework. Thus for a new case, the case representation would evolve with the unfolding of the conversation over time. It is worth noting that in reality not all possible factors may become available and new factors previously unseen in the case base may be added, which are both supported by the AA-CBR framework, making it suitable for this use case.

When retrieving a solution for a new case, some possible solutions (i.e, response types) are reflective, neutral or advice [10] and which of these is best suited will depend on the client's current frame of mind and other personality traits. For each of these possible solutions, similar cases can be retrieved and argumentation used to decide the outcome of choosing that solution. Accordingly, the response type(s) that leads to a positive predicted outcome can then be chosen for the next response from the counsellor. Thus, the AA-CBR framework argues why a particular response type should be chosen over other response types for a new client, based on the previous outcomes seen in the case base. Fig 3) depicts an example, where the default outcome for the response type of giving advice is taken to be positive. However, giving advice when the speaker sentiment is anger results in a negative outcome, and this node attacks the default outcome node. For the new conversation, the argument graph will result in proposing a positive outcome for choosing to advice, using the same logic as described in section 1.1.

It is likely that the AA-CBR framework may not be directly suitable for conversations. For instance, a crucial assumption in AA-CBR is that a particular combination of factors always

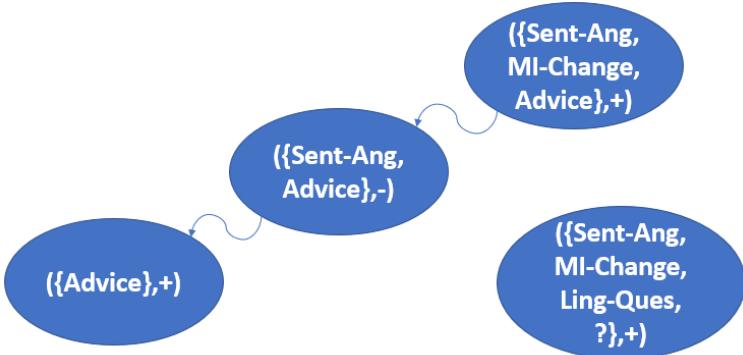


Figure 3: A simple example of the case structure and application of AA-CBR to MI: Each node represents a case; MI-Change is an expression of willingness to change; Linguistic-Question is the linguistic act of asking a question; Advice is the counsellor response type of providing advice

leads to the same outcome. However for conversations, this cannot be guaranteed in reality and outcomes may be probabilistic. Also, considering the temporal aspect, the same factors may appear at different times in the conversation and their ordering may result in different outcomes (eg. anger at the beginning vs end of the conversation are different). Therefore, the research will explore an extension of the AA-CBR framework to address such challenges.

Generate responses and evaluate: For a given conversation context, the next response-type as determined using the AA-CBR framework will be used as conditioning input to a suitable natural language generation model. The generated response will then be evaluated for:

- How well the response aligns to the given input response-type: by comparing the semantic similarity between outputs with and without the input conditioning.
- Whether the response-type conditioning can match the baseline performance using non-interpretable generative models as in [2], while providing interpretability.

3. Conclusion

This research proposes the use of AA-CBR for an interpretable modelling of motivational interviewing. The idea is to determine the response type at each counsellor turn in a new conversation by comparing it to similar conversations in the case base. Specifically, the case structure, i.e the representation of each conversation as a set of dialogue factors and the case structure evolution as the conversation progresses will be investigated. Further, approaches for extending the AA-CBR framework to allow for probabilistic attack relationships between cases and multi-outcome case representations will be explored and will form the major contribution of the research. The proposed work is currently in the initial stages and other research directions such as case adaptation in AA-CBR may also be explored in the future.

References

- [1] L. Tavabi, T. Tran, K. Stefanov, B. Borsari, J. Woolley, S. Scherer, M. Soleymani, Analysis of behavior classification in motivational interviewing, in: Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, 2021.
- [2] S. Shen, C. Welch, R. Mihalcea, V. Pérez-Rosas, Counseling-style reflection generation using generative pretrained transformers with augmented context, in: Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2020.
- [3] K. Resnicow, F. McMaster, Motivational interviewing: Moving from why to how with autonomy support, *The international journal of behavioral nutrition and physical activity* (2012).
- [4] N. Inui, T. Ebe, B. Indurkhy, Y. Kotani, A case-based natural language dialogue system using dialogue act, in: IEEE International Conference on Systems, Man and Cybernetics, 2001.
- [5] K. Eliasson, An integrated discourse model for a case-based reasoning dialogue system, SAIS-SSL event on Artificial Intelligence and Learning Systems (2005).
- [6] K. Ćyras, K. Satoh, F. Toni, Explanation for case-based reasoning via abstract argumentation, in: Computational Models of Argument, 2016.
- [7] K. Ćyras, K. Satoh, F. Toni, Abstract argumentation for case-based reasoning, in: Fifteenth international conference on the principles of knowledge representation and reasoning, 2016.
- [8] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence* (1995).
- [9] T. Althoff, K. Clark, J. Leskovec, Large-scale analysis of counseling conversations: An application of natural language processing to mental health, *Transactions of the Association for Computational Linguistics* (2016).
- [10] Z. Wu, S. Baloccu, V. Kumar, R. Helaoui, E. Reiter, D. R. Recupero, D. Riboni, Anno-mi: A dataset of expert-annotated counselling dialogues, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.
- [11] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, *Proceedings of the AAAI Conference on Artificial Intelligence* (2019).
- [12] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, in: Proceedings of the 10th International Joint Conference on Natural Language Processing, 2020.
- [13] M. Biesialska, K. Biesialska, M. R. Costa-jussà, Continual lifelong learning in natural language processing: A survey, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020.

Counterfactual Explanations for eXplainable AI (XAI)

Greta Warren^{1,2,*†}

¹School of Computer Science, University College Dublin, Dublin, Ireland

²Insight SFI Centre for Data Analytics, University College Dublin, Dublin, Ireland

Abstract

Counterfactual explanation has become a popular and promising method of explaining black-box AI systems and their decisions in recent years. However, a lack of rigorous psychological research means that little is known about what constitutes a ‘good’ counterfactual explanation, or how they facilitate user understanding of the underlying system. My doctoral research aims to examine how these sorts of explanations are understood and evaluated by users, identify desirable characteristics of counterfactual explanations, and investigate how current state-of-the-art counterfactual explanation techniques satisfy these criteria. These insights will guide the development of a novel explanation method designed to meet the psychological requirements of users. In order to address these research questions, to date I have conducted three large-scale, well-controlled user studies using materials drawn from an existing case-base. These studies have yielded novel findings about the impact of counterfactual explanation on users objective understanding and subjective judgments of an AI system. Based on these results, we have proposed an extension of a case-based counterfactual method that produces psychologically-valid explanations, which is to our knowledge, the first method designed with this specific criterion in mind.

Keywords

XAI, counterfactual, contrastive, CBR

1. Introduction

Explaining opaque AI systems and their decisions using contrastive counterfactual examples has gained considerable traction in recent years (see [1, 2] for reviews). To this end, concepts from case-based reasoning (CBR) such as Nearest Unlike Neighbours (NUNs [3]) have inspired such approaches to explanation-by-example, by providing information about how an alternative system decision could have been made, had some aspect of the input data been different [4]. For example, after rejection for a bank loan, a counterfactual explanation may inform the applicant: “had your salary been €10,000 higher, your application would have been approved”. Counterfactual explanations have been proposed to appeal to important characteristics of human explanation and causal reasoning [5, 6], as well as offering potential for recourse [2]. However, although there has been a surge in the number of methods proposed for generating counterfactual explanations computationally, there is limited evidence to show that the outputs of these methods meet the psychological criteria of a ‘good’ explanation, while a lack of controlled user studies to evaluate their impact on user understanding and perceptions of the

ICCBR DC’22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France

*Corresponding author.

✉ greta.warren@ucdconnect.ie (G. Warren)

✉ 0000-0002-3804-2287 (G. Warren)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

system jeopardises their real-world utility. Furthermore, many existing studies rely on users' subjective satisfaction, trust, or fairness judgments, which may not necessarily reflect the depth of their understanding of the system's causal mechanisms [7].

My doctoral research seeks to address these issues by examining how counterfactual explanations are evaluated by human users using both objective and subjective measures, and identifying psychological desiderata of these sorts of explanations. This is achieved by conducting large-scaled, controlled user studies with materials drawn from an existing case-base. These insights will guide an analysis of existing computational methods to assess how well they meet these psychological criteria, as well as the design of a novel, psychologically-grounded case-based approach to counterfactual explanation.

2. Research Plan

2.1. Research Objectives

Counterfactual explanations have received significant attention in recent years as a means of elucidating decisions made by black box AI systems to users. Over 100 methods have been proposed to generate counterfactual explanations [1], and are commonly compared to the state of the art with reference to proximity [8], sparsity [9], and plausibility [2]. However, it is striking that so few of these methods are evaluated with respect to the primary stakeholders (i.e., end-users [1]). Moreover, these quantitative metrics are based on researchers' intuitions about what constitutes a 'good' explanation, however, it is unclear how (and if) they map to longstanding psychological and philosophical definitions of explanatory power [5, 10]. Indeed, although there is a rich body of literature surrounding human explanation [10], and counterfactual reasoning [11], relatively little is known about counterfactual explanations beyond the context of XAI and how they are understood.

The core objectives of my research are to examine how counterfactual explanations impact users' understanding and perceptions of an AI system, and identify the optimal characteristics of these explanations, in order to guide the design of a novel, user-centric counterfactual method that produces psychologically-valid explanations. Specifically, I investigate how counterfactual explanations of AI predictions improve users' objective accuracy in a prediction task, and subjective judgments of explanation satisfaction and trust in the system. In addition, I examine how focusing on certain feature-types appears to increase user accuracy, and hence, help users more readily understand the AI system. These insights will guide both the development of a counterfactual explanation method that meets users' psychological requirements, as well as shed new light on counterfactual explanation in human cognition. The key research questions I have identified are:

- What are the optimal characteristics of a counterfactual explanation from a psychological perspective?
- Which counterfactual methods produce the best explanations in terms of computational metrics (e.g., sparsity, proximity, plausibility)?
- How can a counterfactual method produce explanations that meet users' psychological criteria of explanations?

2.2. Approach / Methodology

User Studies. In order to investigate the effects of counterfactual explanations on users' understanding and evaluations of an AI system, we conducted a series of user studies designed to assess the impact of counterfactual explanation on users' task accuracy and subjective judgments. We compare these effects to those of causal explanations and a control condition (in which participants receive only descriptions of the system's decisions). Participants in the studies were presented with materials in the form of case-instances, each consisting of five features used to predict blood alcohol content: gender (male/female), weight (in kg), amount of alcohol consumed by the person (in units), duration of drinking period (in minutes), and stomach-fullness (full/empty). Users were shown the output of a simulated AI system presented as an application, designed to predict whether someone is over the legal blood alcohol content limit to drive. Materials were selected from a case-base of instances of normally-distributed values of the feature-set. In the training phase of the experiments, participants were shown examples of tabular data for different individuals, and asked to make a judgment about whether each individual was under or over the limit on each screen. After giving their response, feedback was given on the next page, along with an explanation, the content of which was dependent on the experimental condition (see Figure 1 for a sample of the material used in the counterfactual condition). Upon completing the training phase, participants began the testing phase, in which they were shown more example instances referring to individuals and again asked to judge if each individual was over or under the legal limit to drive. For each instance, participants were asked to consider a specific feature in making their prediction; for instance, "Given this person's WEIGHT, please make a judgment about their blood alcohol level." After submitting their response, no feedback or explanation was given. In addition to measuring task accuracy, participants were also asked to provide judgments of explanation satisfaction and trust, measured using the DARPA Explanation Satisfaction and Trust scales [12] respectively, allowing us to evaluate explanation quality using both objective and subjective measures, which may not necessarily correspond with one another.



Figure 1: Feedback for Incorrect Answer in the Counterfactual condition of the study

Towards a Psychologically-valid Counterfactual Method. A key result from the user studies discussed above was that users were significantly more accurate when making predic-

tions about categorical features (stomach fullness and gender) than continuous features (units, weight and drinking duration). This finding is supported by evidence from the counterfactual reasoning literature that people do not spontaneously change continuous variables when generating counterfactuals for past events [13]. In light of this, we conducted an analysis of NUNs with categorical feature differences in a number of popular UCI datasets, observing that they are exceedingly rare. Hence, we developed a variation of Keane and Smyth's [9] case-based counterfactual method, which applies post-hoc transformations to the feature differences in order to produce counterfactual explanations more intuitively understandable to end-users (see [14] for more detail).

3. Progress Summary

To date, I have conducted three large-scale, well-controlled user studies (total $N = 474$) which have revealed novel insights into how counterfactual and causal explanations are understood and perceived by users. While counterfactual explanations are judged as more satisfying and trustworthy than causal explanations, they appear to be only slightly more effective in improving objective performance in a prediction task. This disconnect between objective and subjective measures suggests that it is critical to examine how explanations aid user understanding rather than merely improve subjective perceptions. Furthermore, users appear to understand the impact of categorical features on the system's decision more readily than that of continuous features, a distinction that current computational methods do not account for. Findings from the first user study were presented at the *Cognitive Aspects of Knowledge Representation* workshop at *IJCAI'22* [15], with preliminary results presented at *CogSci'21*. Findings from the complete series of user studies are currently being prepared for submission to a top-tier conference.

Based on the finding that counterfactuals that change categorical features are more readily understood than those focusing on continuous features, we developed a counterfactual method that accounts for this feature-type distinction. An analysis of common UCI datasets suggests that sparse counterfactuals with categorical feature-changes are relatively rare, and so our method adapts Keane and Smyth's [9] case-based technique to transform feature-differences into categorical versions, without significant decrement to performance in terms of coverage and proximity of the counterfactuals produced. To our knowledge, this is the first counterfactual method designed to meet identified psychological requirements for explanation by users, and will be presented at *ICCBR'22* [14].

The main focus of my research at present is the design of a second series of psychological experiments examining the role of simplicity (or sparsity) in counterfactual explanation, and how it impacts user understanding and subjective judgments. In tandem, I am working on the implementation and evaluation of popular counterfactual computational methods, in order to identify those methods which are most successful (i.e. have the best average performance) in generating counterfactuals that meet given criteria of an explanation over a set of representative problems. These criteria include conventional metrics (e.g., proximity, sparsity, plausibility) as well as novel properties derived from user testing (such as whether a counterfactual makes continuous or categorical feature-changes). The final phase of my Ph.D. research will involve synthesising the insights from these two strands of work in order to develop a novel, psychologically-grounded

method for counterfactual explanation.

References

- [1] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques, IJCAI-21 (2021).
- [2] A. H. Karimi, B. Schölkopf, G. Barthe, I. Valera, A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects, volume 1, Association for Computing Machinery, 2020.
- [3] C. Nugent, D. Doyle, P. Cunningham, Gaining insight through case-based explanation, *Journal of Intelligent Information Systems* 32 (2009) 267–295.
- [4] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, *IEEE transactions on visualization and computer graphics* 26 (2019) 56–65.
- [5] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* 267 (2019) 1–38. doi:10.1016/j.artint.2018.07.007.
- [6] R. M. Byrne, Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning, volume 2019-Augus, 2019, pp. 6276–6282. doi:10.24963/ijcai.2019/876.
- [7] Z. Buçinca, P. Lin, K. Z. Gajos, E. L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems, 2020, pp. 454–464. doi:10.1145/3377325.3377498.
- [8] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, *Harvard Journal of Law & Technology* 31 (2018).
- [9] M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), Springer, Cham, 2020, pp. 163–178.
- [10] F. C. Keil, Explanation and understanding, *Annual Review of Psychology* 57 (2006) 227–254. doi:10.1146/annurev.psych.57.102904.190100.
- [11] R. M. Byrne, Counterfactual thought, *Annual review of psychology* 67 (2016) 135–157.
- [12] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for Explainable AI: Challenges and Prospects, Technical Report December, 2018.
- [13] D. Kahneman, A. Tversky, The simulation heuristic, in: D. Kahneman, P. Slovic, A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York, 1982, pp. 201–8.
- [14] G. Warren, B. Smyth, M. T. Keane, “better” counterfactuals, ones people can understand: Psychologically-plausible case-based counterfactuals using categorical features for explainable ai (xai), in: To appear in ICCBR'22, 2022.
- [15] G. Warren, M. T. Keane, R. M. Byrne, Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in xai, in: IJCAI-22 Workshop on Cognitive Aspects of Knowledge Representation, 2022.

Using Machine Learning Techniques to Support Marathon Runners

Ciara Feely

ML Labs, University College Dublin, Dublin, Ireland
Ciara.Feely@ucdconnect.ie

Abstract. This document describes the research plan for the project “Using Recommender Systems Techniques to Support Endurance Athletes, in particular Marathon Runners”. My PhD began in September 2019 and I am due to complete it in 2023, which means the next 12 months will be focused on completing the remaining research and preparing a thesis for submission. Building on work previously presented at ICCBR, this project aims to develop supports for runners including performance prediction, understanding of the risk of sustaining training disruptions, and recommending tailored training plans by applying machine learning techniques to the noisy, inconsistent time series data that is routinely collected by wearable sensors. The progress consists of extracting weekly training features from 15 million sessions for 300,000 unique marathon runners. Several publications involving race-time prediction, training plan recommendation, and training disruptions have been produced, so far working on a reduced dataset. Future work planned will include further feature engineering, applying the models to the larger dataset and providing a more detailed evaluation of training recommendations and explanations with a user study.

Keywords: CBR for health and exercise; marathon running

1 Research Problems

The purpose of this PhD is to develop a recommender system that generates tailored training programmes that enable marathon runners to achieve their performance goals using machine learning with the raw activity data that is routinely collected by wearable sensors. A recommender system has several components. First we must elicit a user’s preferences to know how a given recommendation will benefit them. Here that translates to converting the raw activity data that comes in the form of time-series tracked at regular intervals for distance, speed, and elevation, into a suitable level of abstraction – daily, weekly, and monthly features, allowing for several machine learning models to be compared and contrasted to predict marathon performance based on training. Second we need to understand the cost of the recommendations, and for marathon running this would be how under or overtraining or training inconsistently could cause a runner to sustain an injury and miss their race or simply not meet their full

potential. Finally we need a way to represent suitable options to the user i.e. to present runners with various training programmes and explaining the benefit and potential cost that following or not following the programmes has on their performance. The research questions cover each of these.

1. Can we predict marathon performance from activity data?
2. Can we model training consistency and the risk of sustaining training disruptions?
3. Can machine learning be used to recommend tailored training programmes to marathon runners?

1.1 Feature Engineering and Performance Prediction

To plan their training, marathon runners need an accurate estimate of their fitness and future marathon performance. The sports science community has generated dozens of different techniques for estimating fitness and predicting performance on race-day based on key physiological fitness indicators, training programme summaries and race histories, however there is no single approach that works well for runners of all backgrounds [1]. Previously, in Smyth and Cunningham [2, 3] the authors used a CBR model to predict future personal best time based on previous race-times with recreational runners. While this was accurate, it was not a suitable method for novice runners who do not have previous marathon experience. What is lacking is a marathon performance prediction model that can produce predictions for elite and recreational runners alike, adaptable to each point in training. GPS and heartrate data routinely captured by wearable sensors allows for more detailed features such as fastest 10km and average heartrate to be calculate per individual training sessions, and this PhD will involve the creation of a variety of representations of such data to facilitate machine learning tasks, for example extracting daily, weekly or monthly features. Many of these features will not be explicit in the data, for example features that capture a runner's recovery time, and it is also possible that different representations and models will be required for different runners for example males versus females, or recreational versus elite runners.

1.2 Understanding Training Consistency and Disruption Risk

When developing a recommender system that generates training programmes to improve a marathoner's performance, the potential cost for a runner adopting the recommendation is that the programme might be too difficult, causing them to sustain some sort of injury, or too easy causing them to train and perform sub-optimally. Running is an extremely taxing sport, with an estimated 2.5 times a runner's bodyweight hitting the ground with each stride. Novice runners and longer distance runners such as marathoners are at the greatest risk of sustaining an injury [4]. While a number of different causes of injury have been studied – from training load features such as total weekly distance, to physiological variables related to running gait, to personality traits – there has been no concrete

cause found other than that having a history of running related injuries makes you more likely to have a subsequent injury. *Consistency* in training is a key marker that the runner's training plan is at the appropriate level – striking the balance between intense and achievable such that they will be able to optimise their performance. A *training disruption* on the other hand is a marker that the training could be too intense potentially leading to some running related injury, or that the runner has become demotivated. Regardless of the reason for their inconsistency in training runners will need to understand how this has impacted their performance and how they can get their training back on track following some disruption.

1.3 Recommending and Explaining Training Plans

Training for a marathon requires at least 12-16 weeks of intense training with sessions of a variety of goals to build strength, endurance, and speed. While elite marathon runners might have access to coaches to help them plan their training programmes, recreational runners rely on one-size-fits-all training programmes gleaned from an internet search. These programmes use goal marathon pace, but otherwise these programmes are not tailored to the runner's current status, history, training preferences, lifestyle etc. Additionally as training progresses the training programmes do not adapt to become perhaps more ambitious if a runner is over-performing, or to become lighter if a runner has experienced some training set-back. This research question surrounds developing a recommender system capable of offering tailored training programmes to runners, that are adaptable as training progresses. Explaining the training suggestions and presenting them to runners suitably is imperative for trust and adoption of the system.

2 Research Plan

This research project commenced in October 2019, and the intended finish date is August 2023. In what follows I will give an overview of the work to date and proposed future work for each of the main research questions.

2.1 Feature Engineering and Performance Prediction

For my PhD I have access to an anonymised set of all sessions uploaded into popular mobile fitness application Strava between 2014-2017, via a data-sharing agreement. I have since identified over 500,000 unique runners who have completed marathon distances, and extracted training features for the 16 week programme leading up to race-day. Data cleaning and censoring has resulted in a set of 15 million training activities for over 500,000 marathon programmes from just under 300,000 unique runners.

Work to date consisted of building case-based reasoning models to predict marathon performance at any point in training using training features [5,6], and also incorporating previous race-times [7]. The current best performing model

achieves error rates of approximately 5-8% depending on the point in training. The planned work is to further explore features for performance prediction in particular to better summarise training sessions for example the variability of heartrate or pace might indicate an interval session. The incorporation of heartrate data could also give an indication of the level of intensity of individual training sessions. Another key step is to investigate methods beyond case-based reasoning for performance prediction – so far some baseline linear models and decision tree models have been tested but were found to be less accurate than the case-based reasoning models. An exploration of more black-box models is planned to understand the performance-explainability trade-off in this domain.

2.2 Understanding Training Consistency and Disruptions

One difficulty in working with raw activity data is that we lack information from runners – so we don't know what a runner's goal is, whether a runner is finding the training too challenging or too easy, whether they have become injured or what their race or injury history is. To combat this, to date the work on understanding the cost of running recommendations has focused on training disruptions – lengthy periods without any training activities – as a proxy for injury. We presented a model predicting training disruptions with 20,000 cases that lacked accuracy – approximately 60% accurate, however the use of a case-based reasoning model and counterfactuals allowed for a fully explainable output to runners to facilitate their understanding of why their risk of sustaining a disruption was high based on training patterns of other runners [8]. Presently I am conducting a large scale data analysis of training disruptions – their frequency, impact on marathon performance and how training leading up to and returning from training presents – which will be presented in a sports analytics journal. Future work will involve reapplying the previous prediction model with the larger dataset and improved feature engineering to hopefully improve the error rates, and additionally to combine this model with the performance-based training recommendations to help runners understand how pushing for a more ambitious goal may have some risks to consider.

2.3 Recommending and Explaining Training Plans

Thus far training recommendations have been incorporated into the marathon performance prediction model [5, 6] by filtering the case-base by goal-time and using the future weeks of training of the most similar runner to the query runner based on training to this current point. I used a similar approach to generate training recommendations based on reducing injury risk [8] and here counterfactuals were incorporated to help runners understand what specifically in their current training has contributed to them having a greater risk of experiencing a training disruption.

The planned future work is to generate more detailed training explanations based on performance goals, and to evaluate these further. Counterfactual explanations allow runners to reflect on things they could have done differently

along the training programme, and prefactual explanations will offer runners a sense of what adaptions they can make to their training programme now to reach their goal. Up until now evaluation has been an offline proof-of-concept style evaluation, and a user-study is planned with runners to ascertain whether the training explanations and recommendations are sensible to runners. Additionally, I need to consider how this can be presented to runners. While the case-based recommendation system allows us to easily retrieve the suitable training features, however, presenting these features will not necessarily translate into runners knowing what training to complete. Thus a user study with mock-ups of training plans is planned to understand how to present this to runners.

2.4 Timeline

The aim is to spend the remainder of 2022 finishing up the research projects outlined above, and to then move onto writing the thesis ahead of submission in August 2023. An initial literature review has been written up, as well as a first draft of chapters describing the dataset and work to date.

2.5 Expected Contributions

The main contribution of this work is to develop a system capable of recommending an adaptable programme of training activities to help recreational marathon runners achieve their performance goals safely and effectively. The complexity comes from recommending a series of activities, compared to a one-shot recommended item, as well as from representing the noisy, inconsistent, and unlabelled sensor data available. I believe that the work will have benefits for the running community, and running related research, and that the representations and models developed in this PhD will be applicable to other endurance sports such as races of other distances, cycling, and swimming, and any other recommender systems domain that involve recommending an ordered series of items.

2.6 Conclusions

This PhD aims to develop supports for marathon runners as they train. A dataset of over 500,000 marathon training programmes has been extracted, and modelling for race-time prediction and injury risk based on training programmes completed and presented at conference proceedings. Future work involves finalising the research into feature engineering, and providing counterfactual explanations of training recommendations, as well as writing up the thesis. Being awarded the opportunity to participate at the ICCBR DC would enable my engagement with senior members and also students of the CBR community and receive useful guidance as I finalise my research plans in advance of my final year.

2.7 Acknowledgements

This work is supported by Science Foundation Ireland Centre for Research Training in Machine Learning (18/CRT/6183).

References

1. A. Keogh, B. Smyth, B. Caulfield, A. Lawlor, J. Berndsen, and C. Doherty, "Prediction equations for marathon performance: A systematic review," *International Journal of Sports Physiology and Performance*, vol. 14, no. 9, pp. 1159–1169, 2019.
2. B. Smyth and P. Cunningham, "Running with cases: A CBR approach to running your best marathon," in *Case-Based Reasoning Research and Development - 25th International Conference, ICCBR 2017, Trondheim, Norway, June 26-28, 2017, Proceedings*, pp. 360–374, 2017.
3. B. Smyth and P. Cunningham, "An analysis of case representations for marathon race prediction and planning," in *Case-Based Reasoning Research and Development - 26th International Conference, ICCBR 2018, Stockholm, Sweden, July 9-12, 2018, Proceedings*, pp. 369–384, 2018.
4. B. Kluitenberg, M. van Middelkoop, R. Diercks, and H. van der Worp, "What are the Differences in Injury Proportions Between Different Populations of Runners? A Systematic Review and Meta-Analysis," *Sports Medicine (Auckland, N.Z.)*, vol. 45, pp. 1143–1161, Aug. 2015.
5. C. Feely, B. Caulfield, A. Lawlor, and B. Smyth, "Using case-based reasoning to predict marathon performance and recommend tailored training plans," in *Case-Based Reasoning Research and Development*, pp. 67–81, Springer International Publishing, 2020.
6. C. Feely, B. Caulfield, A. Lawlor, and B. Smyth, "Providing explainable race-time predictions and training plan recommendations to marathon runners," in *Fourteenth ACM Conference on Recommender Systems, RecSys '20, (New York, NY, USA)*, p. 539–544, Association for Computing Machinery, 2020.
7. C. Feely, B. Caulfield, A. Lawlor, and B. Smyth, "An extended case-based approach to race-time prediction for recreational marathon runners," in *Case-Based Reasoning Research and Development - 30th International Conference, ICCBR 2022, Nancy, France September 12-16, 2022, Proceedings*, 2022.
8. C. Feely, B. Caulfield, A. Lawlor, and B. Smyth, "A case-based reasoning approach to predicting and explaining running related injuries," in *International Conference on Case-Based Reasoning*, pp. 79–93, Springer, 2021.

The Use of Case-Based Reasoning for Personalizing Musculoskeletal Pain Treatment Recommendations

Paola Marín-Veites*

¹Norwegian University of Science and Technology, Høgskoleringen 1, Trondheim, 7034, Norway

Abstract

This Ph.D. research proposal presents an overview of the project SupportPrim, a Case-Based-Reasoning (CBR) application for the management of musculoskeletal pain complaints, and its research goals. SupportPrim seeks to become an intelligent decision support system that facilitates co-decision making between clinicians and patients by using machine learning methods. Through its clinician dashboard a treatment plan can be review and tailor to the patient specific needs, moving from the one-size-fits-all mentality to personalized healthcare. The main goals of SupportPrim also include to extend and adapt the decision support system for other primary care settings

Keywords

XCBR, Explainable AI, Visualizations, Decision-Support Systems

1. Problem

Musculoskeletal pain has been described as an epidemic. Approximately 10% of the general population report a chronic musculoskeletal pain complaint in the western world [1]. Musculoskeletal pain is a major reason for consultation in primary care putting a high burden on health services, it also brings serious consequences, such as loss of productivity at work and distress of patients and their families[2]. Current management of musculoskeletal pain is inconsistent across countries and settings, and treatment decisions depend largely on clinician's expertise or opinion. A high number of patients have non-specific symptoms with large variations between individuals. This heterogeneity does not fit well with evidence from clinical trials and clinical guidelines that typically proclaim one-size-fits-all treatment recommendations. This can lead to inadequate patient management and higher costs of resources.

The implementation of tailored treatments for patients can improve treatment planning and ideally yield to result in better patient outcomes and better use of resources. One solution towards this goal is creating intelligent healthcare systems by using explainable and transparent AI methods, such as using Case-Based Reasoning (CBR).

This PhD work is part of a collaborative research project between the Department of Computer Science (IDI) at NTNU and the Department of Public Health and Nursing (ISM) at NTNU. The

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France

*Corresponding author.

✉ paola.m.veites@ntnu.no (P. Marín-Veites)

🌐 <https://www.ntnu.edu/employees/paola.m.veites> (P. Marín-Veites)

🆔 0000-0002-7720-2044 (P. Marín-Veites)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

main goal is to improve research prototypes of a CBR system applied in the SupportPrim project, the target is to improve management of musculoskeletal pain disorders (MSD) in primary care and provide decision support for clinical practice. A general overview of the project in this PhD work is as follows: a group of physicians, general practitioners (GP), collect relevant information from several patients under their care (data acquisition) through previously answered questionnaires and assessment by the GP. This data is then fed to the existing CBR system to expand and adapt the query database for GPs. The system looks for the most similar cases using CBR for suggested treatments and these results are displayed on a clinical dashboard for the GP to assess and create a personalized treatment (patient-clinician co-decision) tailored to the patient's needs. Currently, the CBR system searches for most similar cases based on the set of relevant attributes defined by the project's domain experts. The similarity is modeled on the local and global similarity principle but does not *learn* yet from the cases it obtains. An important goal for this PhD project is to advance the current CBR approach to improve today's application. One major task is to include learning strategies for the system to retain new cases using the provided patient outcomes. For patient management, an important element to be developed is creating visualizations to facilitate the co-decision making of the end users. The ultimate aim is to integrate CBR methods to develop an explainable, intelligent decision support system.

2. Research Plan

For the SupportPrim CBR system, we want to focus on developing a system that reflects its multidisciplinary team. The domain experts are an important part in the development phase, the current CBR system has a global similarity function with weighted attributes defined by them. Recent experimentation suggest there is room for improvement. Finding the right experimentation setting to incorporate their knowledge with computer science will lead to obtain the best possible outcome. In the first phase of this research, we focused on how to improve the development phase of a CBR system. As this is a core starting point in any application, in our first paper[3], we focused on creating visualizations for domain experts, so they understand better how their data is working within the CBR system and assess its performance, such as the attributes' contribution to the global similarity measure and the retrieval results. During this work, we observed that modifications need to be made, e.g. re-evaluating the attributes' weights, so our next experiments will focus on further improvements in the development phase of the CBR system. We will create visualizations that allow to observe the correlations within the attributes and assess their weights influence in the overall results.

At this first phase, we focused on understanding the current system, how it was built and how it works to create a baseline that we can improve. The second phase of the research are the improvements of the SupportPrim CBR system, resulting from the changes that need to be addressed from the first version. The new version will incorporate modifications in the global similarity function, assess the case base representation and redefine weights of the attributes. For the extended version, we aim to incorporate a learning capability. The process can be divided in three general steps. In step one, through the clinician dashboard, the physician can see the most similar patients for the new case. In step two, the dashboard shows the new case

characteristics and allows the physician to annotate the treatment information that will be followed, data on this step is stored and it's where the co-decision making between clinician and patient is done. In step three, successful patients treatments will be retained as new cases in the case base automatically. Recio et al,[4] mention features like a system that learns as the therapy evolves, data-driven configuration that besides patients' data also includes the experts input in an initial configuration and in the reuse of cases to make them more suitable. To achieve this objective different clustering methods will be tested to find the appropriate setting. For the third phase of the research we want to focus on the explainability aspect of the CBR system. This task involves creating visualizations not only for the domain experts but also for the end users. Explanation types will be defined as the research evolves, at the present time we are considering the use of counterfactuals, the literature review on XCBR, model agnostic explanations and visualizations. The expected result of this research is an improved, fully functional CBR system that:

- Personalizes treatment recommendations
- Automatically creates intuitive summaries for physicians
- Generates explanations for treatment recommendations for physicians to understand the system's results

2.1. Research Objectives

2.1.1. Investigate knowledge acquisition techniques to adjust the knowledge containers over time

To provide accurate and more personalized information for each treatment recommendation, the learning approach should take into consideration the different angles and knowledge discovery, e.g. treatment evolution, data representation, physician's inputs. Domain experts and clinicians will help with the integration of the CBR output into clinical context to ensure that it is not only functional in theory but also in practice.

2.1.2. Learning strategies for case-base evolution

The CBR system will keep collecting new cases (data points) from incoming patients from the general practitioners and physiotherapists. Every time a new treatment is created, it will be added to the case base (dataset) for reference. We will explore learning strategies for retaining new cases. These include different patients' factors, such as their clinical data and/or the data recorded from previous sessions.

2.1.3. Extend the existing CBR tool with explainability capabilities

The current tool for the project is myCBR, a Java-based development framework. It is designed to expose modelling functionality, as described by Bach et al. [5], creating concepts and similarity functions that run through a HTTP REST API and can be used with all programming languages that supports Rest API and parsing JSON objects. An assessment of the current CBR system will be done to make improvements from the existing functions that can be transfer to different

clinical settings, where the main goal is to incorporate explainability functionality. Factors like data visualization and clinicians' adoption of intelligent support systems need to be addressed in the integration of the system in clinical practice for it to be successful [6].

2.2. Methodology

2.2.1. Apply Clustering Techniques and CBR Methods combined with domain knowledge

Oliveira et al. proposed a methodology for the CBR modeling process that “facilitates the allocation of expertise between the application domain and the CBR technology”[7]. Their approach will be useful for analysing and redefining the new SupportPrim CBR system as a whole. Starting from the modeling by studying their approach of static, contextual and dynamic attributes, to studying the CBR system variables relevant for its management process and actions to perform in the retrieval results and in storing new data.

Clustering techniques will be explored to evaluate the data. SupportPrim groups patients with similar characteristics in classes (phenotypes). This grouping only happens at the beginning before starting treatment, K-Means can be used to assess if a new clustering later on, with the patient evolution might further help in pinpointing their treatment needs. As highlighted by Bichindaritz et al.[8] “CBR is also known for its knowledge containers - vocabulary, similarity measure, case base, and adaptation. The case base in and of itself is often a major focus of knowledge discovery in CBR, with its cases, structures, and organization”. Bichindaritz et al., mention several functionalities well defined for knowledge discovery, learning new trends and association of data, for clustering particularly, they mention hierarchical clustering or density-based algorithms, which could be adequate to explore for the SupportPrim project, for possibly assessing a re-grouping of the patients depending on the treatment evolution. We expect that pattern recognition of the SupportPrim data can help to investigate if there are other existing patterns that can be integrated in the CBR configuration to make the reuse of cases treatment more suitable for recommendations, a ranking of cases with clustered case-based organization. Lamy et al[9], mention several algorithms in their CBR system for cancer detection, such as KNN, high-dimensional multivariate data visualization and Artificial Feeding Birds (AFB) metaheuristic for adaptable optimization algorithms that propose an interesting approach and that could be helpful for this research, for SupportPrim, their methods could lead to creating better visualizations of the data, as it can be presented in both qualitative and quantitative form, while contributing to the explainability element as well. Mahdi and Seifi [10] suggest a Bayesian network for classifying diseases to support effective medical treatment using experts' knowledge. Their classification methods are based on data and domain experts knowledge and both are considered in the cases, for SupportPrim, their methods with feature reduction and clustering might improve the CBR performance. Other works will be reviewed to improve the existing CBR system.

2.2.2. Create user friendly visualizations and results explainability

Currently, the SupportPrim clinician's dashboard displays the patient relevant data and stores the physician's examination and treatment plan. We want to update the dashboard so that it reflects

the summary of relevant data and it's user friendly and intuitive to facilitate the clinicians work as the CBR system evolves. As Kenny et al. [11] mention, adoption barriers can be addressed by the explanation capabilities designed to improve adoption, such as adequate predictions and providing "personalised explanation-by-example". For this task we are considering creating counterfactual or model agnostic explanations, including unsuccessful cases can also help in creating these explanations. Cunningham et al. [12] outline their experiment setting on a case-based explanation system, in their work, subjects score the explanations. The case-based explanation system showed to perform better than having no explanation and better than rule-based systems. Visualizations in the development phase are an important element as well, as they allow to assess and verify that the implemented CBR system works as intended. We will work on creating tools to explore CBR system's for domain experts

3. Progress Summary

As this project already had a system prototype, one of the first tasks was to revise and understand the system's programming and its existing functions. Understanding the type of data and its meaning was important to get familiar with the case base representation. Currently, a review of the state of the art is being done, we narrowed the topics to four main ones of interest: cbr explainability, visualizations, model agnostic explanations and counterfactuals. This task is expected to be finished by the end of August 2022. We have worked on the first paper soon to be published related to understanding our current CBR system through visualizations for our domain experts, to have the baseline to improve from, the visualizations created will be useful in the next development phase of the new CBR system. The next step is working on extending and improving the CBR system and visualizations doing experiments for a second paper, we will explore the use of autoencoders in a CBR system taking into consideration the input from domain experts and visualizations on attributes correlations within the system. Later on, depending on experimentation, we will explore the methods mentioned in the Methodology to incorporate explainability and find the right setting with the clinicians to achieve a human-centered AI.

References

- [1] J. T. Gran, The epidemiology of chronic generalized musculoskeletal pain, *Best Practice & Research Clinical Rheumatology* 17 (2003) 547–561.
- [2] C. J. Main, A. C. de C Williams, Musculoskeletal pain, *Bmj* 325 (2002) 534–537.
- [3] P. Marín-Veites, K. Bach, Explaining cbr systems through retrieval and similarity measure visualizations: A case study, in: M. Keane, N. Wiratunga (Eds.), *Case-Based Reasoning Research and Development*, Springer, Cham, 2022.
- [4] J. A. Recio-García, B. Díaz-Agudo, A. Kazemi, J. L. Jorro, A data-driven predictive system using case-based reasoning for the configuration of device-assisted back pain therapy, *Journal of Experimental & Theoretical Artificial Intelligence* 33 (2021) 617–635.
- [5] K. Bach, B. M. Mathisen, A. Jaiswal, Demonstrating the mycbr rest api., in: ICCBR Workshops, 2019, pp. 144–155.

- [6] Q. Yang, A. Steinfeld, J. Zimmerman, Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–11.
- [7] E. M. Oliveira, R. F. Reale, J. S. Martins, A methodological approach to model cbr-based systems, arXiv preprint arXiv:2009.04346 (2020).
- [8] I. Bichindaritz, C. Marling, S. Montani, Recent themes in case-based reasoning and knowledge discovery, in: The Thirtieth International Flairs Conference, 2017.
- [9] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, B. Séroussi, Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach, Artificial Intelligence in Medicine 94 (2019) 42–53.
- [10] M. M. Ershadi, A. Seifi, An efficient bayesian network for differential diagnosis using experts' knowledge, International Journal of Intelligent Computing and Cybernetics 13 (2020) 103–126.
- [11] E. M. Kenny, E. Ruelle, A. Geoghegan, L. Shalloo, M. O'Leary, M. O'Donovan, M. T. Keane, Predicting grass growth for sustainable dairy farming: A cbr system using bayesian case-exclusion and post-hoc, personalized explanation-by-example (xai), in: International Conference on Case-Based Reasoning, Springer, 2019, pp. 172–187.
- [12] P. Cunningham, D. Doyle, J. Loughrey, An evaluation of the usefulness of case-based explanation, in: International conference on case-based reasoning, Springer, 2003, pp. 122–130.

Developing a Decision Support System leveraging Distributed and Heterogeneous Sources: Case-Based Reasoning for Manufacturing Incident Handling

M. van der Pas^{1,2,*}

¹*Eindhoven University of Technology, Department of Industrial Engineering & Innovation Sciences, Eindhoven 5600MB, The Netherlands*

²*Semaku B.V., Eindhoven 5617BC, The Netherlands*

Abstract

Case-Based Reasoning is a proven method to provide decision support in a manufacturing context. However, data and knowledge relevant for the case representation is often spread over distributed sources, leading to challenges in the case representation and retrieval. Those challenges require different techniques that this PhD project aims to develop. Techniques for data collection and integration during the case representation, as well as similarity measurement during case retrieval. This paper describes the motivating problem, the research methods, and the current state and future plans.

Keywords

Incident Handling, Traceability, Case-Based Reasoning, Semantic Web, Knowledge Graph, Event Graph

1. Introduction

One of the challenges identified for Case-Based Reasoning (CBR) research is the acquisition of cases from heterogeneous and distributed data sources [1]. This challenge certainly also applies to complex manufacturing environments, where CBR can be applied to assist engineers with the handling of quality incidents. In case customers have an issue with a device, they might initiate a (quality) complaint at the company that produced the device. The company should then analyse the complaint and take suitable measures, like containment and corrective action. This complaint handling process is taking an increasing amount of effort, caused by the increasing product and production process complexity. Especially in the semiconductor industry [2], which is the main motivating use case for this project. During the handling process there are already some commonality checks done to find historic complaints related to a new complaint. However, due to the challenges described below only to limited extend.

In manufacturing companies there are many data available about the products and their production process, which can be used to describe a (complaint) case. These data are often spread over many different systems. Not only because of the different types of data that are of interest, but also because of the complexity of the semiconductor production process, consisting

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France

*Corresponding author.

 m.c.a.v.d.pas@tue.nl (M. v. d. Pas)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

of many different production steps spread over multiple facilities. At the same time it is costly to index all data in a central case base. Therefore, a system that supports engineers with identifying similar cases (historic complaints), will have to deal with distributed and heterogeneous sources. These characteristics introduce specific challenges, and this research will try to solve some of them, focusing on case representation and retrieval phase of the CBR cycle [3].

2. Research Plan, Objectives, and Approach

The goal of this project is to develop a (decision support) system that assists engineers in the handling of manufacturing incidents by providing cases similar to the new case they have at hand. The system should leverage data and knowledge from distributed and heterogeneous sources. As such, answer the following research question: **How to identify related quality incidents in a manufacturing environment with distributed and heterogeneous data and knowledge sources?**

2.1. Sub-projects

The project is divided into four sub-projects. The topic of the first sub-project is the development of a general framework for CBR-based decision support leveraging distributed and heterogeneous sources. The other sub-projects focus on specific components in this framework, for case representation and retrieval. More details about the evaluation of the components and system can be found in subsection 2.2.

I. Framework

How to design a system to find related quality incidents in a manufacturing environment with distributed and heterogeneous data and knowledge sources?

The first sub-project focuses on a framework to support the CBR-cycle, more specifically the case representation based on distributed sources. In distributed decision support systems [4] and CBR systems for knowledge management [5, 6, 7] it is common to use an agent-based approach. Chaudhury et al. [8] proposed a solution for CBR with distributed storage of cases. Similarly, Camarillo [9] proposed a knowledge management framework using CBR in an industrial context. However, both focus on combining cases from distributed case bases, while this research focuses on gathering data from distributed systems for case representation. Therefore, the system should be able to collect and integrate data from heterogeneous sources to describe a new case, refine and enrich the case representation, and provide similar cases back to the user. An overview of the main steps can be found in Figure 1. The system will to a large extend rely on Semantic Web Technologies, which are proven to be suitable for combining data and knowledge driven approaches [10]. The main components of this architecture are investigated as part of the other sub-projects. Once the components are developed, the framework will also be implemented and evaluated with the (source) systems in a case study.

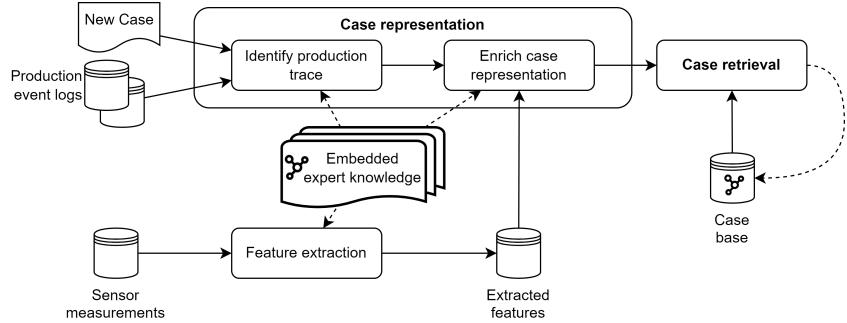


Figure 1: Main steps of the CBR-based decision support for quality incidents.

II. Case Representation: Trace identification

How to identify the entities and events that were involved in the production process of a case?

During the production process often multiple case identifiers are used and production batches are split and merged. For example, when multiple semi-finished goods are assembled into one device. This results in multi-dimensional event data, and introduces fuzziness and uncertainty in the trace. Therefore, it is a challenge to collect relevant data and information to build a case representation. The production trace [11] can serve as the foundation for the case representation. The production trace describes the production process of a device and consists of production events and related entities. It can be represented as an event graph, which is well suited to represent multi-dimensional event data [12], combining the time and relation dimension. In comparison to Esser and Fahland [12], this research aims to enrich the event graphs with knowledge encoded in ontologies. For example, Lee and Park [13] improved the traceability using information about the bill of materials. The developed technique should be able to generate an event graph consisting of events and entities on different levels of aggregation, which describe the production trace for the case at hand.

III. Case Representation: Data integration

How to integrate data from distributed and heterogeneous sources?

The first step of the representation phase introduced by Finnie and Sun [14] is to construct a case description. This research will focus on event and sensor data for describing the case, which are common in the manufacturing domain [15]. As it is costly to integrate and store all sensor data centrally, a method is required to reduce the data volume and dimensionality to be able to integrate it into one case representation. Wang et al. [16] propose to aggregate sensor data to events and integrate those events with graph structured context data. Similarly, the system developed by Gundersen et al. [17] abstracts sensor data to events using pattern matching. Those events are subsequently used in CBR to find similar situations from the past. In a similar way, this research aims to use machine learning techniques to extract features from a series of data points [18]. Those data points are generated by sensors on the production equipment and describe the production process of one device. The extracted features should correlate to quality incidents.

IV. Case Retrieval: Similarity Measurement

How to find similar quality incidents based on heterogeneous incident descriptions leveraging domain knowledge?

The goal of this sub-project is to develop a technique that can be used to retrieve cases similar to a novel case. Camarillo et al. [6] use a predefined set of attributes to describe the case and its context. To deal with the distributed and heterogeneous sources, a more flexible case representation format is required. Therefore, this research aims to use RDF (Resource Description Framework) knowledge graphs and corresponding graph-based similarity measurement techniques. Zhang et al. [19] also used knowledge graphs, but conclude that more work needs to be done on the similarity and knowledge reasoning. Furthermore, domain knowledge is required to conduct proper similarity measurement. This knowledge can be represented by taxonomies or ontologies [6, 20, 21, 5, 22]. There are various standards for describing ontologies/taxonomies using RDF, for example OWL¹ and SKOS², which as such can be integrated in the case representation.

2.2. Evaluation of the system

The sub-projects will result in different components of the system, which require different methods and data sets to validate and evaluate their functioning.

Sub-project II The data collection and integration solution can be validated using simulated or actual (from a semi-conductor use case) production events. However, there exists no data set with 'known-good production traces'. Therefore, the aim is to identify and reconstruct a number of traces for a validation data set, with the help of engineers.

Sub-project III The feature extraction technique can be evaluated using sensor data collected from equipment that is used in the production process. After most production steps a quality check is done. The results of this check can be used to validate if the derived features are indeed correlated to quality incidents.

Sub-project IV The aim is to evaluate the graph-based case comparison technique, using a data set from the semi-conductor use case described in the introduction (handling of customer complaints). The data set consists of historic complaints, traceability data (production events from different Manufacturing Execution Systems (MES)), and data from Product Life cycle Management (PLM) systems. In practice there are already some commonality checks done by the engineers, which can be used as a benchmark.

The integrated system The preferred method of evaluating the integrated system, which integrates the techniques developed in the sub-projects, is to combine the data sets used to evaluate those sub-projects. However, the challenge is that only a very small portion of produced devices result in a complaint, which in turn are only detected months to years after production. Therefore, it will be difficult to construct a data set with relevant sensor and complaint data. A

¹<https://www.w3.org/TR/owl2-primer/>

²<https://www.w3.org/TR/skos-primer/>

possible solution is to use simulated data, based on the data collected before. An alternative is to find a different use case, in which quality incidents occur with higher frequency, such that it is possible to construct a data set that contains related sensor and incident data.

3. Progress Summary

At the time of submission, most work is done on defining the framework for data collection and integration (sub-project I) in the context of the MAS4AI project. In future, the developed framework will be evaluated in a case study, using data and information sources from a manufacturing company. Next to the work on sub-project I, two case studies are in progress which look into modelling event graphs based on manufacturing events, and feature extraction from sensor data, respectively contributing to sub-project II and III. Both focus on a specific step in the semi-conductor production process.

Acknowledgements

This project is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 957204, the project MAS4AI (Multi-Agent Systems for Pervasive Artificial Intelligence for assisting Humans in Modular Production).

References

- [1] A. K. Goel, B. Diaz-Agudo, What's Hot in Case-Based Reasoning, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 5067–5069.
- [2] J. Aelker, T. Bauernhansl, H. Ehm, Managing complexity in supply chains: A discussion of current approaches on the example of the semiconductor industry, Procedia CIRP 7 (2013) 79–84. doi:10.1016/J.PROCIR.2013.05.014.
- [3] A. Aamodt, E. Plaza, Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, AI Communications 7 (1994) 39–59. doi:10.3233/AIC-1994-7104.
- [4] K. Sycara, D. D. Zeng, Multi-Agent Integration of Information Gathering and Decision Support, in: 12th European Conference on Artificial Intelligence, John Wiley & Sons, Ltd., 1996, pp. 549–553.
- [5] P. Reuss, K.-D. Althoff, A. Hundt, W. Henkel, M. Pfeiffer, Multi-Agent Case-Based Diagnosis in the Aircraft Domain, in: Proceedings of the ICCBR 2015, Frankfurt, 2015, pp. 43–52.
- [6] A. Camarillo, J. Ríos, K.-D. Althoff, Knowledge-based multi-agent system for manufacturing problem solving process in production plants, Journal of Manufacturing Systems 47 (2018) 115–127. doi:10.1016/j.jmsy.2018.04.002.
- [7] W. Luís, J. Carlos, E. Ferreira, F. Gonzalez, C. Gomes, R. M. Lorenzo, W. Lui's, L. Mikos, C. Espíndola, E. Ferreira, A combined multi-agent and case-based reasoning approach to support collaborative nonconformance problem solving in the thermoplastic injection moulding process, <https://doi.org/10.1080/09511920903440321> 23 (2010) 177–194. doi:10.1080/09511920903440321.

- [8] S. Chaudhury, T. Singh, P. S. Goswami, Distributed fuzzy case based reasoning, *Applied Soft Computing* 4 (2004) 323–343. doi:[10.1016/J.ASOC.2003.10.003](https://doi.org/10.1016/J.ASOC.2003.10.003).
- [9] A. Camarillo González, Knowledge Management framework based on an Agents Network to support Continuous Improvement in Manufacturing integrating Case-Based Reasoning and Product Lifecycle Management, Ph.D. thesis, Universidad Politécnica de Madrid, 2018. doi:[10.20868/UPM.thesis.53989](https://doi.org/10.20868/UPM.thesis.53989).
- [10] B. Steenwinckel, D. De Paepe, S. Vanden Hautte, P. Heyvaert, M. Bentefrit, P. Moens, A. Dimou, B. Van Den Bossche, F. De Turck, S. Van Hoecke, F. Ongenae, FLAGs: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge with machine learning, *Future Generation Computer Systems* 116 (2021) 30–48. doi:[10.1016/j.future.2020.10.015](https://doi.org/10.1016/j.future.2020.10.015).
- [11] R. Schuitemaker, X. Xu, Product traceability in manufacturing: A technical review, *Procedia CIRP* 93 (2020) 700–705. doi:[10.1016/J.PROCIR.2020.04.078](https://doi.org/10.1016/J.PROCIR.2020.04.078).
- [12] S. Esser, D. Fahland, Multi-Dimensional Event Data in Graph Databases, *Journal on Data Semantics* 10 (2021) 109–141. doi:[10.1007/S13740-021-00122-1/TABLES/6](https://doi.org/10.1007/S13740-021-00122-1/TABLES/6).
- [13] D. Lee, J. Park, RFID-based traceability in the supply chain, *Industrial Management and Data Systems* 108 (2008) 713–725. doi:[10.1108/02635570810883978/FULL/PDF](https://doi.org/10.1108/02635570810883978/FULL/PDF).
- [14] G. Finnie, Z. Sun, R5 model for case-based reasoning, *Knowledge-Based Systems* 16 (2003) 59–65. doi:[10.1016/S0950-7051\(02\)00053-9](https://doi.org/10.1016/S0950-7051(02)00053-9).
- [15] H.-N. Dai, H. Wang, G. Xu, J. Wan, M. Imran, Enterprise Information Systems Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies, *Enterprise Information Systems* 14 (2020) 1279–1303. doi:[10.1080/17517575.2019.1633689](https://doi.org/10.1080/17517575.2019.1633689).
- [16] H. Wang, P. Nguyen, J. Li, S. Kopru, G. Zhang, S. Katariya, S. Ben-Romdhane, GRANO: Interactive graph-based root cause analysis for cloud-native distributed data platform, *Proceedings of the VLDB Endowment* 12 (2019) 1942–1945.
- [17] O. Erik Gundersen, A. Aamodt, P. Skalle, A Real-Time Decision Support System for High Cost Oil-Well Drilling Operations PhD project View project Reproducible AI View project, *Ai Magazine* 34 (2012) 21. doi:[10.1609/aimag.v34i1.2434](https://doi.org/10.1609/aimag.v34i1.2434).
- [18] F. J. Baldán, J. M. Benítez, Multivariate times series classification through an interpretable representation, *Information Sciences* 569 (2021) 596–614. doi:[10.1016/J.INS.2021.05.024](https://doi.org/10.1016/J.INS.2021.05.024).
- [19] Y. Zhang, X. Liu, J. Jia, X. Luo, Knowledge representation framework combining case-based reasoning with knowledge graphs for product design, *Computer-Aided Design and Applications* 17 (2020) 763–782. doi:[10.14733/CADAPS.2020.763-782](https://doi.org/10.14733/CADAPS.2020.763-782).
- [20] G. Căndea, S. Kifor, C. Constantinescu, Usage of Case-based Reasoning in FMEA-driven Software, *Procedia CIRP* 25 (2014) 93–99. doi:[10.1016/J.PROCIR.2014.10.016](https://doi.org/10.1016/J.PROCIR.2014.10.016).
- [21] B. Chebel-Morello, M. K. Haouchine, N. Zerhouni, Reutilization of diagnostic cases by adaptation of knowledge models, *Engineering Applications of Artificial Intelligence* 26 (2013) 2559–2573. doi:[10.1016/J.ENGAPPAI.2013.05.001](https://doi.org/10.1016/J.ENGAPPAI.2013.05.001).
- [22] S. H. El-Sappagh, M. Elmogy, Case Based Reasoning: Case Representation Methodologies, *IJACSA) International Journal of Advanced Computer Science and Applications* 6 (2015).

DL-CBR Hybridization for Feature Generation and Similarity Assessment

Zachary Wilkerson

Indiana University, Bloomington, IN, 47408, United States

Abstract

Effective retrieval is essential to strong case-based reasoning performance, and retrieval quality is critically dependent on case indexing. Such indices are not always feasible to generate manually, and so a thorough exploration of how features and weights may be generated automatically (especially using deep learning) is necessary. To that end, this summary outlines a research plan for investigating structural influences on feature quality, how learned features may be used in concert with knowledge-engineered features, and how weights may be generated in feature-dense spaces created by feature learning. It also proposes a methodology for modular exploration of various models, training set sizes, numbers of features generated, etc., to provide a comprehensive foundation of index generation using deep learning. Finally, it points to already-published research that works towards some of these goals and illustrates how results from these existing projects inform future research plans.

Keywords

Case-based reasoning, Deep learning, Hybrid systems, Feature learning, Weight learning

1. Introduction

Case-based reasoning (CBR) performance relies significantly on retrieving useful cases from the case base. In turn, retrieval quality depends on indices used to characterize/discriminate between cases. High-quality indices can be derived through manual knowledge engineering (e.g., [1, 2]), but this approach can be costly and is not feasible in some domains. For example, indexing vocabularies may be unsatisfactory for poorly-understood domains or for complex tasks such as computer vision. An analogous challenge exists for weights as well—effective feature weights can augment feature information for indexing, but even provided a comprehensive feature set, it can be difficult and/or expensive to identify useful weighting information for those features.

These problems may be addressed using feature and weight learning. Initially, this was achieved using symbolic methods (e.g., [3]); however, with recent advances in deep learning (DL, esp. in domains such as computer vision), it is natural to consider how increased performance of DL architectures may be translated to CBR retrieval. Specifically, CBR systems can be described as inherently interpretable via case presentation, but black-box DL systems are traditionally viewed as more accurate for most domains; however, if a CBR system can be made more accurate by leveraging DL methods/structures, then resulting DL-CBR hybrid systems may be applicable to a wide variety of domains, representing a “best of both worlds” with regards

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France

✉ zachwilk@indiana.edu (Z. Wilkerson)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

to accuracy and explainability (e.g., “Twin systems” by Kenny and Keane in [4]). In this vein, recent research leverages neural networks to generate and refine feature information inferred from training examples for classification and/or explanation [5, 6, 7], but there still exist many areas of potential DL-CBR integration for which there has been little or no research exploration.

This research summary outlines strategies for exploring feature and weight learning using DL in greater depth, presenting both a blueprint detailing potential research objectives and methodologies as well as an overview of steps taken so far and the resulting publications.

2. Research Objectives and Methodology

This research broadly seeks to deeply explore methods for leveraging DL models to generate indexing information to supplement or replace information gathered through knowledge engineering for computer vision-related tasks. This overarching goal can be subdivided into the four primary investigation regions described below, which are followed by a proposed research plan outline for exploring them.

2.1. Exploring Methods for Generating High-Quality Features

Under the umbrella of network-based feature generation, there exist multiple potential variables that can influence feature quality. For one, different model structures provide unique pathways for feature generation, exemplified by different computer vision approaches (e.g., comparing AlexNet, Inception, DenseNet, transformers, and MLP-based models); additionally, architectures may be leveraged in different ways (e.g., using ensembles of networks for localized feature generation in the multi-net approach explored in [8]). For another, the way in which features are extracted from DL models affects feature quality (e.g., extracting features from different locations in a model). Analysis of the impact of such variables on feature quality is an essential foundation for optimizing DL-CBR hybrid system performance.

2.2. Using Knowledge-Engineered and Learned Features in Concert

While feature learning can be useful in domains for which generating features through knowledge-engineering is not feasible, more research is required to evaluate feature learning augmenting incomplete knowledge-engineered indices. This includes exploring methods for effectively using both feature sets in concert, for which it may be necessary to mitigate harmful effects of a “curse of dimensionality” as a result of extracted feature spaces being generally denser than knowledge-engineered ones. Integrating knowledge-engineered and network-generated features also requires investigation into potential discretization of continuous network-generated features as well as into the independence of generated features and/or their correlation with knowledge-engineered features.

2.3. Refining Weight Learning Methods for Feature-Dense Spaces

As mentioned above, using learned features can result in similarity assessment being performed in feature-dense spaces, for which conventional methods of weight learning (e.g., [4]) may

be less effective. To this end, it is important to consider ways in which such techniques may be refined to accommodate larger numbers of features, and/or to explore methods to extract weights directly from a network architecture, potentially in concert with feature extraction. This objective also encompasses how combinations of feature and weight learning methods influence retrieval quality, especially with regards to extracting feature weights from different network architectures.

2.4. Evaluating DL-CBR Hybrid Model Explainability

Investigation of hybrid systems leveraging interpretable CBR structures alongside more opaque DL systems demands contextualization relative to explainability. Innately, DL-CBR hybrid systems imply an overall architecture that is more interpretable than an out-of-the-box DL model but less so than a CBR system using only knowledge-engineered information. Thus, in addition to optimizing index quality to maximize retrieval accuracy, it is important to assess where on an explainability spectrum that this work sits and to take measures where possible to maximize interpretability.

2.5. Proposed Research Plan

The four research objectives described above encompass a diverse range of research avenues, and they present specific sub-goals that align with an overarching three-step process that guides the proposed research methodology. Specifically, this process begins with deep exploration of network-based index generation methods, including different ways in which generated indices may be integrated into a CBR system. The second step involves post-processing of these indices, particularly for optimizing CBR system performance/accuracy, but also potentially including discretization for better combination with knowledge-engineered index information where applicable. Finally, the resulting DL-CBR hybrid model is analyzed/contextualized with respect to explainability, especially in comparison to CBR systems using only knowledge-engineered indexing information.

To this end, initial experiments imitating established index generation methods (e.g., [5, 7], see next section for details) have established both a proof of concept that network-generated and knowledge-engineered features used in concert can enable greater CBR accuracy than either feature set used individually and that the network architecture/structure can have a substantial impact on feature quality. Next steps will focus on other network structure influences (esp. how different DL models affect feature quality), enabling a comprehensive analysis of network-generated feature sets augmenting knowledge-engineered feature sets.

Beyond this point, future experiments could continue in any of several directions. For one, feature weighting strategies may be more deeply investigated and/or revised in the context of potentially denser feature spaces created using network-generated features. For another, potential relationships between knowledge-engineered and network-generated features may be explored for explainability purposes. Such investigations would include dependency correlations between knowledge-engineered and network-generated features and/or methods by which continuous network-generated features may be discretized, along with the resulting impact on feature quality.

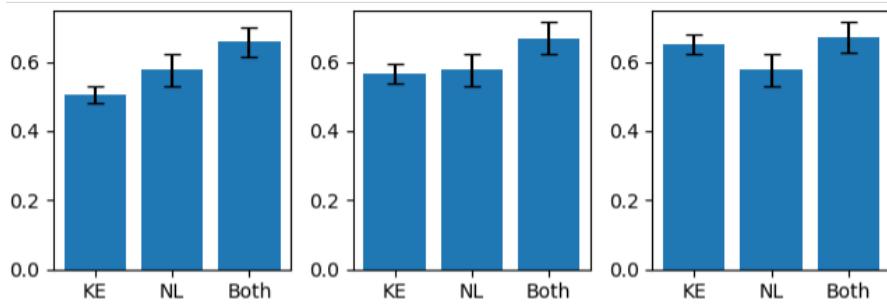


Figure 1: Comparison of retrieval accuracy using knowledge-engineered (KE) features, network-generated/learned (NL) features, or both sets together (First published in Case-Based Reasoning Research and Development, ICCBR 2021 by Springer [9]). From right to left, knowledge-engineered features are perturbed by greater magnitudes, resulting in less per-feature reliability.

3. Progress to Date

As of the writing of this summary, most progress to date has focused on providing a conceptual foundation and initial empirical tests for DL-CBR hybridization for retrieval, in line with the objectives presented above. Specifically, such explorations include using network-generated features in concert with knowledge-engineered features for greater classification accuracy, as well as investigating structural influences (e.g., network architecture/structure and feature extraction location) on feature quality, both using retrieval accuracy as proxy. The following subsections summarize the associated publications.

3.1. Augmenting Similarity Feature Engineering with Deep Learning

This research [9] assumes the availability of knowledge-engineered feature information for a given domain, but that such feature information may be incomplete and/or inaccurate. In these instances, existing feature information may be supplemented by additional learned features extracted from raw data using DL. In these instances, the inclusion of learned features improves retrieval accuracy by capturing indexing information to which humans might not be sensitive.

Results supporting this hypothesis are obtained using a zero-shot learning dataset for computer vision. Each image is associated with a unique case, and per-class feature information from the dataset is perturbed based on a random coefficient and combined with values extracted from the image using a convolutional neural network (CNN) to form the case's feature set. The two combined sets of values represent knowledge-engineered and network-generated features, respectively. Retrieval accuracy values for the aggregated feature set are compared against corresponding accuracy values using either component set exclusively (Figure 1).

Based on the outcomes from these initial tests, combining feature sets does improve retrieval accuracy. However, additional variables such as the reliability of knowledge-engineered features and the number of features extracted from the CNN may significantly influence the magnitude of accuracy improvement. In addition, preliminary tests regarding weight extraction in parallel with feature extraction suggest that more feature-dense spaces created by extracting features

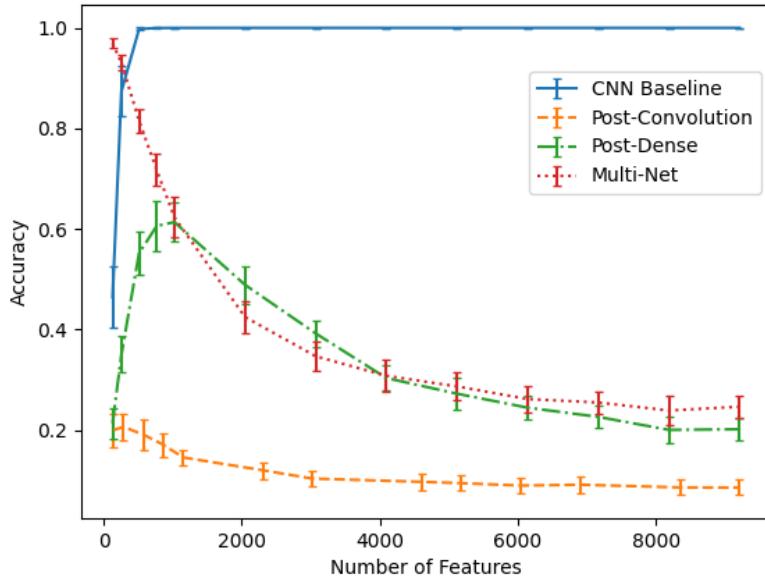


Figure 2: Comparison of different feature extraction methods with an end-to-end CNN classifier for different numbers of features for a basic CNN architecture (First published in Case-Based Reasoning Research and Development, ICCBR 2022 by Springer Nature [8]). Post-Convolution extracts features immediately after convolution and pooling steps, Post-Dense extracts features immediately after the dense layers (before the output layer), and Multi-Net extracts features as in Post-Dense from an ensemble of networks based on a candidate retrieved case’s class.

from CNNs seem to accommodate established weighting strategies poorly.

3.2. Exploring Structural Influences on Generated Feature Effectiveness

In contrast to the previous description, this work [8] specifically investigates how the way in which features are extracted affects feature quality. To this end, two feature extraction locations and a novel model ensemble structure that generates localized features are explored.

The experiments investigate the hypothesis that extracting features from later in the network results in higher-quality features from the perspective of the CBR system. Retrieval accuracy values are used as proxy for feature quality among the three proposed methods and a CNN baseline (Figure 2). Additionally, keeping in mind the consequences of feature-dense spaces discovered in the previous work, varying numbers of features are extracted from each model.

The results support the hypothesis and suggest that localized feature sets may be especially accurate for feature-sparse scenarios preferred by CBR systems (if at the cost of increased training time). The number of features does significantly impact retrieval performance as well, both as a “curse of dimensionality” for large numbers of features and as a minimum requirement for DL model convergence for smaller numbers of features.

4. Future Work

Future work will build on the current publications' findings while moving forward to address other objectives. In the short term, research will focus on exploring feature generation using both different DL models across multiple datasets and different experimental parameters (e.g., number of training examples). Later experiments will investigate potential weight generation methods, as well as how weight generation methods are affected by the number of features.

References

- [1] D. Leake, An indexing vocabulary for case-based explanation, in: Proceedings of the Ninth National Conference on Artificial Intelligence, AAAI Press, Menlo Park, CA, 1991, pp. 10–15.
- [2] R. Schank, M. Brand, R. Burke, E. Domeshuk, D. Edelson, W. Ferguson, M. Freed, M. Jona, B. Krulwich, E. Ohmaya, R. Osgood, L. Pryor, Towards a general content theory of indices, in: Proceedings of the 1990 AAAI spring symposium on Case-Based Reasoning, AAAI Press, Menlo Park, CA, 1990.
- [3] S. Bhatta, A. Goel, Model-based learning of structural indices to design cases, in: Proceedings of the IJCAI-93 Workshop on Reuse of Design, IJCAI, Chambery, France, 1993, pp. A1–A13.
- [4] E. M. Kenny, M. T. Keane, Twin-systems to explain artificial neural networks using case-based reasoning: Comparative tests of feature-weighting methods in ANN-CBR twins for XAI, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019.
- [5] S. Sani, N. Wiratunga, S. Massie, Learning deep features for kNN-based human activity recognition, in: Proceedings of ICCBR 2017 Workshops (CAW, CBRDL, PO-CBR), Doctoral Consortium, and Competitions co-located with the 25th International Conference on Case-Based Reasoning (ICCBR 2017), Trondheim, Norway, June 26–28, 2017, volume 2028 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017, pp. 95–103.
- [6] K. Martin, N. Wiratunga, S. Sani, S. Massie, J. Clos, A convolutional siamese network for developing similarity knowledge in the selfBACK dataset., in: A. A. Sanchez-Ruiz, A. Kofod-Petersen (Eds.), Proceedings of the ICCBR 2017 Workshop on Case-Based Reasoning and Deep Learning, CEUR Workshop Proceedings, 2017, pp. 85–94.
- [7] J. T. Turner, M. W. Floyd, K. M. Gupta, T. Oates, NOD-CC: A hybrid CBR-CNN architecture for novel object discovery, in: Case-Based Reasoning Research and Development, ICCBR 2019, 2019, pp. 373–387.
- [8] D. Leake, Z. Wilkerson, D. Crandall, Extracting case indices from convolutional neural networks: A comparative study, in: Case-Based Reasoning Research and Development, ICCBR 2022 (in print), 2022.
- [9] Z. Wilkerson, D. Leake, D. Crandall, On combining knowledge-engineered and network-extracted features for retrieval, in: Case-Based Reasoning Research and Development, ICCBR 2021, 2021, pp. 248–262.