Technische
Universität
Berlin

TU Berlin | FG MLKOM | Einsteinufer 17 - Sekr. EN-16 10587 Berlin

prof. dr hab. inż. Maciej Ogorzałek
Biuro Studiów Doktoranckich Szkoły Doktorskiej
Nauk Ścisłych i Przyrodniczych

Wydział Fizyki, Astronomii i Informatyki
Stosowanej UJ
ul. prof. S. Łojasiewicza 11, 30-348 Kraków

Fakultät IV
Elektrotechnik und Informatik
Institut für Softwaretechnik und
Theoretische Informatik

Fachgebiet Maschinelles Lernen und
Kommunikation

Leitung
Prof. Dr. Wojciech Samek

Sekretariat EN-16
Einsteinufer 17
10587 Berlin

Telefon +49 (0)30 31002 417
wojciech.samek@tu-berlin.de

Berlin, 18.12.2023

Report for the Dissertation “Interpretable Deep Learning with Prototypical Parts for Supervised and Weakly-Supervised Learning”

Considerable advances have been made in the field of Artificial Intelligence (AI), especially with Deep Neural Networks (DNNs) achieving impressive performances on a multitude of domains. While their success stems from ample data availability and widespread use, their complexity presents a critical challenge: they operate as opaque black-boxes in decision-making, lacking transparent reasoning. Consequently, their decisions can be biased and unpredictable, hampering trustworthiness, particularly in high-stakes scenarios and regulatory contexts. Addressing this issue, Explainable AI (XAI) has emerged as a field aiming to mitigate these limitations. The primary approaches in XAI involve either training a black-box model and subsequently explaining its predictions using dedicated techniques or designing models that are inherently self-explanatory, enabling them not only to make accurate predictions but also to provide explanations alongside these predictions from the outset.

The excellent dissertation of Dawid Damian Rymarczyk focuses on the latter approach. It contributes by (1) introducing two novel prototypical parts-based models overcoming the limitations of the famous ProtoPNet model, (2) generalizing prototypical parts and soft neural decision trees to the regression problem, and (3) generalizing the prototypical parts approach to Multiple Instance Learning (MIL) problems. These very important methodological contributions are of large practical value as they help to overcome important limitations (lack of scalability, no shared semantics no modeling of interactions), and thus may drastically extend the application range of models such as ProtoPNet or Attention-based MIL approaches.

Contents and Structure of the Dissertation

The thesis is well written and demonstrates that Dawid Damian Rymarczyk has acquired profound theoretical and practical knowledge in the topical research field of explainable AI, in particular on self-explainable models. The thesis adeptly distills complex concepts into a clear, concise narrative that remains comprehensive without oversimplification. The methods proposed are meticulously outlined and a set of insightful experiments is reported. Further details are described in the five scientific papers, published at top-tier conferences, at the end of the thesis.

Chapter 1 serves as an introduction to the thesis. It briefly summarizes the motivation and outline of the thesis. Furthermore, this chapter provides an overview of the scientific publications of the candidate, of which five are the basis for the thesis. The author has co-authored six additional papers, which are not explicitly included but certainly mark important progress towards and beyond the thesis. A formatting issue “prede[U+FB01]ned” can be found on page 9.

Chapter 2 describes the research scope of the dissertation, highlighting the main methodological contributions and briefly describing the conducted experiments.

Chapter 3 provides a detailed description of the contributions. This chapter is divided into two parts. Part 3.1 focuses on the contributions related to the prototypical parts-based models, while part 3.2 describes the contributions in the field of interpretable multiple instance learning. Each contribution is presented in a very clear manner, starting with a motivation statement describing the state-of-the-art and its limitations, a brief summary of the contribution, a detailed description including the necessary mathematical equations, a brief presentation of the main results, and finally a summary paragraph.

The first contribution presented in this chapter is the ProtoPShare method. This method tackles two limitations of the famous ProtoPNet approach, namely the poor scalability (i.e., large number of prototypes) and the lack of shared semantics (i.e., prototypes of different classes are pushed away although they may have similar semantics). The proposed merge-pruning process enables sharing of prototypical parts across multiple data classes. Experimental results show that ProtoPShare performs well even when the number of prototypes is reduced to 1/4. A user study was performed to investigate the effectiveness of data-dependent vs. data-independent measures of prototype similarity. In addition, an interesting theoretical analysis of ProtoPShare is provided in the published paper.

The second contribution further extends ProtoPShare and allows to learn prototypical parts sharing from scratch (i.e., without relying on ProtoPNet initialization). This approach very nicely formalizes the problem and allows to directly optimize for a self-explainable model with shared prototypes. Different “tricks” are introduced (e.g., novel focal similarity function, Gumbel-Softmax estimator, additional constraints) to avoid suboptimal solutions. Results show that this novel ProtoPool approach outperforms other models on different datasets and different base architectures. Also here a user study is performed, showing that ProtoPool generates prototypes with higher saliency scores than the baseline models.

The last contribution in the first part of chapter 3 focuses on the generalization of the prototypical parts methodology to the graph regression problem. The proposed Prototypical Graph Regression Soft Trees (ProGReST) model uses prototypical parts and combines them with Soft Neural Trees to construct the prototypical parts model. Also here the author introduces different regularization tricks in the training process to make the ProGReST model interpretable and accurate. Results show benefits of the proposed model with respect to interpretability, i.e., useful prototypical parts can be identified, performance and training efficiency.

The second part the chapter focuses on the interpretability of the MIL. In the first contribution a self-attention mechanism is combined with Attention-based MIL Pooling. A key advantage of this approach is the ability to capture dependencies between instances within a bag through the self-attention module. The interactions can be measured either by dot products or by using a kernel function. Results on two medical datasets indicate that the self-attention AbMILP method and its kernel variants outperform the baselines. However, the type of kernel used does not seem to have a large effect on the results.

The final contribution of the thesis introduces a ProtoMIL model. Due to the use of prototypical parts the models not only assures local interpretability, but also a global one. A specifically designed regularization enforces the derivation of prototypes from the underrepresented class (positive label). Results show the effectiveness of the approach on three histopathology datasets. Table 3.5 contains a typo “… thre large-scale …”.

Chapter 4 of the thesis lists the candidate’s achievements, including a list of publications that are not part of the thesis, a list of grants, an overview over research co-operations and a summary of his service to the scientific community.

Chapter 5 concludes the thesis with a short summary. The five key scientific papers are attached at the end of the dissertation.

Impact and Grading of the Dissertation

The thesis is very clear and well-written. It effectively presents the constraints of existing self-explanatory models and the tactics to surpass these limitations in a highly educational manner. The format and organization of thesis was effective in communicating and connecting the research contributions. A clear strength of the thesis is that it not only proposes novel self-explainable models, but also has a strong focus on overcoming practical limitation (scalability, sharing of semantics, modeling interactions) of current state-of-the-art, moreover, it improves the theoretical understanding by examining the changes in the network's predictions after the prototypes' merge in the case of ProtoPShape. The experimental analyses presented in the thesis are of very good quality. The embedding of the proposed methods into the wider XAI literature could have been clearer. Also the conclusion in chapter 5 is very short. An explicit discussion of the limitations of the proposed methods and an outlook on future work would have been beneficial, but are missing in the thesis.

Overall, the thesis clearly makes substantial novel contributions to the field of XAI and has the potential to make an impact in practice by overcoming the limitations of popular self-explainable models such as ProtoPNet. The work described in the thesis has been published at top-tier ML conferences. The achievements of the Ph.D. candidate are excellent, also beyond his great research work. He has shown that he is able to secure funding, establish new collaborations and contribute to the work of other researchers through co-authorship. These are great achievements and a clear indication for a successful scientific career.

Given these commendable achievements, I am confident in stating that the dissertation fulfills all the requisites for a PhD thesis. Consequently, I wholeheartedly recommend Dawid Rymarczyk, MSc Eng's doctoral dissertation for public defense and distinction.

Samek

Prof. Dr. Wojciech Samek