Towards Knowledge-driven Distillation and Explanation of Black-box Models

Roberto Confalonieri[a], Pietro Galliani[a], Oliver Kutz[a], Daniele Porello[b], Guendalina Righetti[a] and Nicolas Troquard[a]

[a]Faculty of Computer Science, Free University of Bozen-Bolzano, Italy
[d]DAFIST, Università di Genova

Abstract

We introduce and discuss a knowledge-driven distillation approach to explaining black-box models by means of two kinds of interpretable models. The first is perceptron (or threshold) connectives, which enrich knowledge representation languages such as Description Logics with linear operators that serve as a bridge between statistical learning and logical reasoning. The second is Trepan Reloaded, an approach that builds post-hoc explanations of black-box classifiers in the form of decision trees enhanced by domain knowledge. Our aim is, firstly, to target a model-agnostic distillation approach exemplified with these two frameworks, secondly, to study how these two frameworks interact on a theoretical level, and, thirdly, to investigate use-cases in ML and AI in a comparative manner. Specifically, we envision that user-studies will help determine human understandability of explanations generated using these two frameworks.

Keywords

explainable AI, knowledge distillation, perceptron logic, Trepan Reloaded, decision trees

1. Introduction

Since the development of expert systems in the mid-1980s [1], explainable Artificial Intelligence (xAI) has been promoting decision models that are transparent, i.e., that are able to explain *why* and *how* decisions are being made. More recent successes in machine learning technology, together with episodes of unfair and discriminating decisions taken by black-box models, have brought explainability back into the focus [2]. This has led to a plethora of new approaches for explanations of black-box models [3], aiming to achieve explainability without sacrificing system performance, and approaches to knowledge discovery in databases [4], aiming to combine Semantic Web data with the data mining and knowledge discovery process.

Two of the most important problems that the two areas above have tried to address, and that are currently of great practical and theoretical interest are those of:

3rd International Workshop on Data meets Applied Ontologies in Explainable Artificial Intelligence. DAO-XAI @ BAKS 2021

roberto.confalonieri@unibz.it (R. Confalonieri); pietro.galliani@unibz.it (P. Galliani); oliver.kutz@unibz.it (O. Kutz); daniele.porello@unige.it (D. Porello); Guendalina.Righetti@unibz.it (G. Righetti); Nicolas.Troquard@unibz.it (N. Troquard)

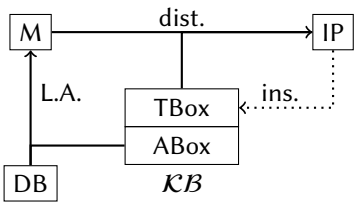

Figure 1: Our proposed workflow. A learning algorithm (L.A.) extracts from the ABox component of a knowledge base ($\mathcal{KB}$) and/or from data of a database (DB) a model (M). This model is then distilled to an intepretable proxy (IP) which may be added to the knowledge base's TBox.

- **Interpretable Machine Learning** [5, 6]: often, it is not sufficient for a model to perform accurate predictions. We also want our model to be *interpretable*, that is, to provide a simple, human-understandable explanation of *why* it makes its predictions.
- **Integration of Prior Domain Knowledge** [7, 8]: classification or regression tasks may not exist in a vacuum, but rather they may concern (and often, *do* concern) topics for which there exists a considerable amount of *domain knowledge* of the kind that is easily represented in terms of logical knowledge bases. This information may be of use both to direct the learning algorithm and to reformulate or generalize its conclusions.

Nonetheless, the interplay between these two problems has remained mostly unexplored. Only a few of these approaches have considered how to integrate and use domain knowledge to foster interpretable machine learning, and to drive the explanation process (e.g., [9]). Furthermore, very few approaches have addressed explanations from a user point of view [10], in particular, analysing what makes for a good explanation [11], how these are perceived and understood by humans, and how to use these findings to measure the understandability of explanations of black-box models.

We propose to address these problems with an approach that may be seen as an instance of *knowledge distillation* [12]. As shown in Figure 1, we propose to first train a (non-necessarily human-interpretable) model on data, and then attempt to approximate the resulting model by reducing it to an *interpretable proxy*. Depending on the nature of the model first trained and of the proxy, this may be trivial (for example if the model first trained is simple and interpretable enough to serve as its own proxy), rather less so (e.g., if the trained model is a multi-layer neural network and we wish for our interpretable proxy to be a linear model), or perhaps considerably difficult (e.g., if the trained model is a complex ensemble model). In general, then, the best way forward will be a modular, possibly model-agnostic, distillation approach through which an interpretable proxy is extracted by *evaluating* that model on arbitrary inputs and learning adaptively a simpler model that best imitates it. We envision that this 'knowledge distillation' procedure could be done by following at least two different approaches to approximate the original machine learning model. On one hand, we could use *threshold* (or "*Tooth*") *expressions*, which extend knowledge representation models by means of linear classifiers [13]. On the other hand, we could use Trepan Reloaded, a model-agnostic approach that provides symbolic explanations, under the form of decision trees, of a black-box model [14, 15].

Since distillation will make use of the knowledge base as well as of the machine learning

model, the resulting interpretable proxy will *integrate* the model and the logical information of the knowledge base to which it will be possibly added later, closing in this way a symbolic integration cycle. The cycle will foster knowledge reuse and sharing.

As pointed out above, another important aspect, which has been nevertheless almost overlooked, is the evaluation of human-understandability of explanations [10, 15]. Research in the social sciences has extensively studied what stands for human-understandable explanations, and how humans conceive and share explanations [10]. Other works studied and proposed how the understandability of explanations can be measured [16]. We will use these works as a basis to design experiments aiming to compare the explanations distilled using *threshold expressions* and Trepan Reloaded.

2. Explanations via Weighted Threshold Operators

Weighted Threshold Operators are n-ary logical operators which compute a weighted sum of their arguments and verify whether it reaches a certain threshold. These operators have been extensively studied in the context of circuit complexity theory, and they are also known in the neural network community under the alternative name of *perceptrons*. In [13], threshold operators were studied in the context of Knowledge Representation, focusing in particular on Description Logics (DLs). In brief, if $C_1 \dots C_n$ are concept expressions, $w_1 \dots w_n \in \mathbb{R}$ are weights, and $t \in \mathbb{R}$ is a threshold, we can introduce a new concept $\mathbb{W}^t(C_1 : w_1 \dots C_n : w_n)$ to designate those individuals d such that $\sum\{w_i : C_i \text{ applies to } d\} \geq t$.

In the context of DL and concept representation, such threshold expressions are natural and useful, as they provide a simple way to describe the class of the individuals that satisfy 'enough' of a certain set of desiderata. For example, let us consider the *Felony Score Sheet* used in the State of Florida[1], in which various aspects of a crime are assigned points, and a threshold must be reached to decide compulsory imprisonment. For example, possession of cocaine corresponds to 16 points if it is the primary offense and to 2.4 points otherwise, a victim injury describable as "moderate" corresponds to 18 points, and a failure to appear for a criminal proceeding results in 4 points. Imprisonment is compulsory if the total is greater than 44 points and not compulsory otherwise. A knowledge base describing the laws of Florida would need to represent this score sheet as part of its definition of its **CompulsoryImprisonment** concept, for instance as

$$\mathbb{W}^{44}(\textbf{CocainePrimary} : 16, \textbf{ModerateInjuries} : 18, \dots).$$

While it would be possible to also describe it (or any other Boolean function) in terms of more ordinary logical connectives (e.g., by a DNF expression), a definition in terms of threshold expressions is far simpler and more readable. As such, the definition is more transparent and more explainable.

We refer the interested reader to [13, 17] for a more in-depth analysis of the properties of this operator. Having threshold expressions in a language of knowledge representation has notable advantages. First, in psychology and cognitive science, the combination of two or more concepts has a more subtle semantics than set theoretic operations. As shown in [18], threshold

[1] http://www.dc.state.fl.us/pub/scoresheet/cpc_manual.pdf (accessed: 13 July 2021)

operators can represent complex concepts more faithfully regarding the way in which humans think of them. For this reason, explanations provided using threshold expressions are in principle more accessible to human agents. Second, as illustrated in [17], since a threshold expression is a linear classification model, it is possible to use standard linear classification algorithms (such as the Perceptron Algorithm, Logistic Regression, or Linear SVM) to learn its weights and its threshold given a set of assertions about individuals (that is, given an ABox).

Extensions of Description Logic involving threshold operators have also been discussed in [19, 20]. The approaches presented in these two papers are, however, very different from the one summarised above: the former paper, indeed, changes the semantics of DL by associating *graded membership functions* to models and requiring them for the interpretation of expressions, while the latter one extends the semantics of the DL $\mathcal{ALC}$ by means of weighted alternating parity tree automata. The approach described above is, in comparison, more direct: no changes are made to the definitions of the models of the DL(s) to which threshold operators are added, and the language is merely extended by means of the above-described operators. Provided that the language of the original DL contains the ordinary Boolean operators, adding the threshold operators to it does no increase the expressive power (as already noted in [13]), but does not increase the complexity of reasoning either [21].

3. Explanations via Decision Trees

In the ML literature, techniques for explaining black-box models are typically classified as local and global methods [3]. Whilst local methods take into account specific examples and provide local explanations, global methods aim to provide an overall approximation of the behavior of the black-box model. Global explanations are usually preferable over local explanations, because they provide a more general view about the decision making process of a black-box. An attempt to aggregate local explanations into global one was proposed in [22].

A seminal explanation method to explaining black-box classifiers is Trepan [23]. Trepan is a tree induction algorithm that recursively extracts decision trees from oracles, in particular from feed-forward neural networks. The algorithm is model-agnostic, and it can in principle be applied to explain any black-box classifier (e.g., Random Forest).

Trepan combines the learning of the decision tree with a trained machine learning classifier (the oracle). At each learning step, the oracle's predicted labels are used instead of known real labels. The use of this oracle serves two purposes: first, it helps to prevent the tree from overfitting to outliers in the training data. Second, and more importantly, it helps to build more accurate trees.

To produce enough examples to reliably generate test conditions on lower branches of the tree, Trepan draws extra artificial query instances that are submitted to the neural network as if they were real data. The features of these query instances are based on the distribution of the underlying data. Both the query instances and the original data are submitted to the neural network 'oracle', and its outputs are used to build the tree.

An extension of the Trepan algorithm, called Trepan Reloaded, was proposed to take into account explicit knowledge, modeled by means of ontologies, in [14]. Trepan Reloaded uses a modified information gain that, in the creation of split nodes, gives priority to features asso-

ciated with more general concepts defined in a domain ontology. This was achieved by means of an information content measures defined using refinement operators [24]. Linking explanations to structured knowledge in the form of ontologies, brings multiple advantages. It does not only enrich explanations (or the elements therein) with semantic information—thus facilitating effective knowledge transmission to users—but it also creates a potential for supporting the customisation of explanations to specific user profiles [25].

To measure the effects of the ontology on the understandability of explanations with human users an on-line user study was conducted. The study showed that decision trees generated by Trepan Reloaded, thus taking domain knowledge into account, were more understandable than those generated without the use of domain knowledge [14, 15].

4. Evaluating Human Understandability of Explanations

Decision trees and threshold expressions appear to have complementary pros and cons as explanatory tools for black-box classifiers. Decision trees have the advantage of having clear visual representations. A human user can easily follow them to understand what factors lead the classifier to reach which conclusion in which circumstances; but on the other hand, especially in the case of very large trees, it can be difficult for a user to follow the overall structure of the decision tree or use it to engage in counterfactual reasoning (e.g., "would the final decision of the classifier have been YES rather than NO if feature C1 had been different?"). Threshold expressions, on the other hand, are arguably of less immediate interpretability for a user; but have the advantage of specifying clearly which factors influence positively or negatively the decision of the classifier, and up to which (comparative) degree, thus making it easier for a user to evaluate the effect that changing certain specific input features would have on the outcome.

Previous work attempting to measure the understandability of symbolic decision models, and decision trees in particular [26, 27], proposed syntactic complexity measures based on the model's structure. The syntactic complexity of an explanation can be measured, for instance, in the case of decision trees, by counting the number of internal nodes or leaves, or in the case of logical formulas, by counting the number of symbols adopted. Having a measure like syntactic complexity, that can be easily computed, is useful from an application perspective. E.g., it may be used to prevent excessive complexity in building decision trees and threshold expressions when explaining a black-box. On the other hand, the syntactic complexity does not necessarily capture precisely the understandability of explanations by users. A direct measure of user understandability is how accurately a user can employ a given explanation to perform a decision. Another measure of cognitive difficulty is the reaction time (RT) or response latency [28]. RT is a standard measure used by cognitive psychologists and has become a staple measure of complexity in the domain of design and user interfaces [29]. Understandability depends on the cognitive load experienced by users, e.g., in using the decision model to classify instances and in understanding the features in the model itself. However, for practical processing human understandability needs to be approximated by an objective measure.

We will compare two characterisations of the understandability of explanations: (i) Understandability based on the syntactic complexity of an explanation (number of internal nodes, leaves, symbols used in a weighted formulas, etc.), and (ii) Understandability based on users'

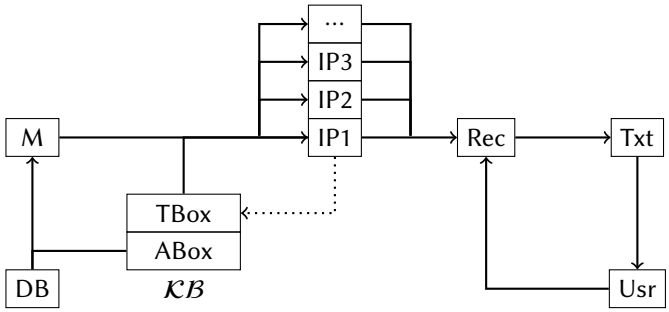

Figure 2: Our eventual aim. From the black-box model (M), learned from a database (DB) and/or from the ABox of a knowledge base ($\mathcal{KB}$) we distil, via the knowledge base $\mathcal{KB}$, a library of interpretable models IP1, IP2, ...with different advantages and disadvantages. A recommender system (Rec) then chooses among these models the one to present to the user (Usr) in form of a narrative generated by a textual translator (Txt). The user in turn interacts with the recommender system.

performances and subjective ratings, reflecting, for instance, the cognitive load by users in carrying out tasks using a given explanation format.

We aim at conducting a user study to measure and compare the understandability of explanations given in the form of decision trees and threshold expressions with human users. This can be done in domains where explanations are critical, e.g., justice, finance or medicine. Conducting and analysing such experiments can provide useful insights under which conditions and tasks one representation is deemed more understandable than the other one by users.

5. Outlook to Future Work

In this work we briefly introduced two novel and promising approaches to distilling black-box models into explainable models while making use of domain knowledge. As discussed in the previous section, the natural next step consists in investigating experimentally the respective advantages and disadvantages of these two approaches, with an eye towards a characterisation of the scenarios in which either provides models that are more understandable and/or closer to the black box model than the other.

Much besides that of course remains to be done. For instance, to further improve understandability, an ulterior processing step translating either kind of model into a textual description might be worth implementing, e.g., using narratives [30], as well as a way to use background knowledge to adjust the explanatory model to the needs of different stakeholders [25].

Ultimately, we aim to integrate these two approaches (or more) into a unified meta-approach that can use multiple modes of explanation for different aspects of a black-box model, automatically choosing among the available options the ones that are best suited to provide a faithful and understandable representation to that specific aspect of the model. This is an ambitious endeavour, which would culminate in the automated distillation of a single black-box model into a *library* of explainable models, different both in kind and in complexity (e.g., number of decision nodes, weights), whose availability to the user is mediated by a recommender system and a textual translator.

References

[1] M. R. Wick, W. B. Thompson, Reconstructive expert system explanation, Art. Intelligence 54 (1992) 33–70.

[2] R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, A historical perspective of explainable artificial intelligence, WIREs Data Mining and Knowledge Discovery 11 (2021). doi:https://doi.org/10.1002/widm.1391.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comp. Surv. 51 (2018) 1–42.

[4] P. Ristoski, H. Paulheim, Semantic Web in data mining and knowledge discovery: A comprehensive survey, Journal of Web Semantics 36 (2016) 1–22. URL: https://www.sciencedirect.com/science/article/pii/S1570826816000020. doi:10.1016/J.WEBSEM.2016.01.001.

[5] A. Vellido, J. D. Martín-Guerrero, P. J. Lisboa, Making machine learning models interpretable, in: 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2012, pp. 163–172.

[6] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[7] W. Liao, Q. Ji, Learning Bayesian network parameters under incomplete data with domain knowledge, Pattern Recognition 42 (2009) 3046–3056.

[8] S. Mei, J. Zhu, J. Zhu, Robust regbayes: Selectively incorporating first-order logic domain knowledge into bayesian models, in: International Conference on Machine Learning, 2014, pp. 253–261.

[9] X. Renard, N. Woloszko, J. Aigrain, M. Detyniecki, Concept tree: High-level representation of variables for more interpretable surrogate decision trees, CoRR abs/1906.01297 (2019). URL: http://arxiv.org/abs/1906.01297.

[10] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. URL: http://www.sciencedirect.com/science/article/pii/S0004370218305988. doi:https://doi.org/10.1016/j.artint.2018.07.007.

[11] R. Confalonieri, T. R. Besold, T. Weyde, K. Creel, T. Lombrozo, S. Mueller, P. Shafto, What makes a good explanation? Cognitive dimensions of explaining intelligent machines, in: Proc. of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019, 2019. URL: https://mindmodeling.org/cogsci2019/papers/0013/index.html.

[12] Y. Cheng, D. Wang, P. Zhou, T. Zhang, A survey of model compression and acceleration for deep neural networks, arXiv preprint arXiv:1710.09282 (2017).

[13] D. Porello, O. Kutz, G. Righetti, N. Troquard, P. Galliani, C. Masolo, A toothful of concepts: Towards a theory of weighted concept combination, in: Proc. of the 32nd International Workshop on Description Logics, volume 2373, CEUR-WS, 2019.

[14] R. Confalonieri, T. Weyde, T. R. Besold, F. M. del Prado Martín, Trepan Reloaded: A Knowledge-driven Approach to Explaining Black-box Models, in: Proc. of the 24th European Conference on Artificial Intelligence, volume 325 of *Frontiers in Artificial Intelligence and Applications*, IOS press, 2020, pp. 2457–2464. doi:10.3233/FAIA200378.

[15] R. Confalonieri, T. Weyde, T. R. Besold, F. M. del Prado Martín, Using ontologies to enhance human understandability of global post-hoc explanations of black-box models,

Artificial Intelligence 296 (2021). doi:https://doi.org/10.1016/j.artint.2021.103471.
[16] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: challenges and prospects, CoRR abs/1812.04608 (2018).
[17] P. Galliani, O. Kutz, D. Porello, G. Righetti, N. Troquard, On knowledge dependence in weighted description logic, in: Proc. of the 5th Global Conference on Artificial Intelligence (GCAI 2019), 2019, pp. 17–19.
[18] G. Righetti, D. Porello, O. Kutz, N. Troquard, C. Masolo, Pink panthers and toothless tigers: Three problems in classification, in: Proc. of the 5th Int. Workshop on Artificial Intelligence and Cognition, Manchester, September 10–11, 2019.
[19] F. Baader, G. Brewka, O. F. Gil, Adding threshold concepts to the description logic $\mathcal{EL}$, in: C. Lutz, S. Ranise (Eds.), Frontiers of Combining Systems, Springer International Publishing, Cham, 2015, pp. 33–48.
[20] F. Baader, A. Ecke, Reasoning with prototypes in the description logic $\mathcal{ALC}$ using weighted tree automata, in: Language and Automata Theory and Applications, Springer International Publishing, Cham, 2016, pp. 63–75.
[21] P. Galliani, G. Righetti, O. Kutz, D. Porello, N. Troquard, Perceptron connectives in knowledge representation, in: Proceedings of 22nd International Conference on Knowledge Engineering and Knowledge Management (EKAW 2020), 2020.
[22] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, GLocalX - From Local to Global Explanations of Black Box AI Models, Artificial Intelligence 294 (2021) 103457. URL: https://doi.org/10.1016/j.artint.2021.103457. doi:10.1016/j.artint.2021.103457.
[23] M. W. Craven, J. W. Shavlik, Extracting tree-structured representations of trained networks, in: NIPS 1995, MIT Press, 1995, pp. 24–30.
[24] N. Troquard, R. Confalonieri, P. Galliani, R. Peñaloza, D. Porello, O. Kutz, Repairing Ontologies via Axiom Weakening, in: AAAI 2018, 2018, pp. 1981–1988.
[25] M. Hind, Explaining Explainable AI, XRDS 25 (2019) 16–19. doi:10.1145/3313096.
[26] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, B. Baesens, An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models, Decision Support Systems 51 (2011) 141–154.
[27] R. Piltaver, M. Luštrek, M. Gams, S. Martinčić-Ipšić, What makes classification trees comprehensible?, Expert Syst. Appl. 62 (2016) 333–346.
[28] F. C. Donders, On the speed of mental processes., Acta Psychologica 30 (1969) 412–31.
[29] J. B. William Lidwell, Kritina Holden, Universal. Principles of Design., Rockport, 2003.
[30] P. Gervás, E. Concepción, C. León, G. Méndez, P. Delatorre, The long path to narrative generation, IBM J. Res. Dev. 63 (2019) 8:1–8:10.