Explainable Data Processing through Knowledge Compilation

PhD topic proposal (Inria Lille)
team LINKS

Co-advisor: Mikaël Monet[1,3], Co-advisor: Florent Capelli[2,3], and Supervisor: Sławek Staworko[2,3]

[1]Inria Lille
[2]University of Lille
[3]UMR 9189 CRIStAL

Abstract: The increasing usage of computing in every aspect of modern society has naturally raised a need for transparency in algorithms. Recently, methods from *knowledge compilation* – a field of Artificial Intelligence (AI) – have been proposed to help explain the decisions of certain algorithms used in machine learning. First, a model that is given in a formalism not easily interpretable (e.g., a neural network) is transformed into a new representation in a knowledge compilation formalism. Then, this new representation is used to answer various *interpretability queries* (such as for instance "How many inputs should be changed so that the entry becomes accepted by the model?"). If the target formalism has been correctly chosen, these queries can be answered efficiently on the new representation.

On the other hand, knowledge compilation has also been applied in the context of relational databases query evaluation, where it has been shown to be useful for solving complex tasks such as probabilistic query evaluation or enumerating query answers. So far however, knowledge compilation has not been considered for explainability in query evaluation.

The goal of this thesis is then twofold. First, the candidate will explore new ways into which knowledge compilation can help explain the decisions of AI algorithms, for instance by designing new relevant interpretability queries or developing new knowledge compilation algorithms. Second, they will investigate how knowledge compilation could be used for explainability in the context of relational databases.

Keywords: Explainable Artificial Intelligence, Knowledge Compilation, Relational Databases.

1 Topic presentation

Knowledge compilation [DM02] is a field of Artificial Intelligence (AI) that studies the relationships between diverse formalisms used to represent Boolean functions, as well as the tasks that can be solved efficiently over these formalisms (satisfiability, model counting, enumeration of models, etc.). The difficulty of a given task depends on the formalism used to represent the function; for instance, the *satisfiability problem* (SAT) can be solved in linear time if the function is represented as a Boolean formula in disjunctive normal form (DNF) but is NP-complete when it is represented as a formula in conjunctive normal form (CNF) or as a Boolean circuit. Similarly, the problem of *counting* the number of satisfying assignments (#SAT) can be solved in polynomial time if the function is given as a *free binary decision diagram* (FBDD), but is intractable if it is given as a DNF. The general idea of knowledge compilation is then the following. Imagine that we have some data and knowledge $\mathcal{D}$ that we want to analyze, through some queries $q_1, q_2, \ldots$. First, we choose a knowledge compilation formalism such that the queries $q_1, q_2, \ldots$ can be answered efficiently on that formalism. We then *compile* the data into that formalism, obtaining a new representation R of our data. Then, we use R to analyse the data by answering the queries. The compilation phase could be expensive, but the advantage of this approach is that it needs to be done only once, after which we can efficiently analyze our data, using its new representation. We next explain the two axes of research that would be pursued in this thesis, and that would be investigated in parallel. Results of both theoretical and practical nature are to be expected.

Knowledge compilation to explain AI algorithms. The knowledge compilation approach has recently been used to *explain* the decisions of algorithm from machine learning, such as Bayesian Classifiers [SCD18b], Binary Neural Networks [SSDC20], and Random Forests [CSGD20]. The model is transformed into a formalism of knowledge compilation such that some interesting *interpretability queries* can be answered efficiently. Examples of interpretability queries include finding *sufficient reasons* [SCD18b] (a sufficient reason being a set of feature values that suffice to classify a given entry[1]), determining the model's *robustness* [SCD18a], *counting completions* (having fixed a subset of features, what is the proportion of the entries that are accepted by the model) [BMPS20], finding *minimum cardinality changes* ("How many inputs should be changed so that the entry becomes accepted by the model?") [BMPS20] or again computing *Shapley values* [ABM20]. For this first axis of the thesis, we propose the following directions of research.

- Carrying a systematic study of which interpretabiliy query can be solved efficiently on which formalisms from knowledge compilation, in the spirit of [AKM20] and [BMPS20]. This would allow us to have a clearer picture of what can be done for each type of model.
- Designing new interpretability queries that have not yet been considered. Indeed, since this line of research is rather recent, new interpretability queries of interest are regularly proposed by the community (for instance, because they would be specific to certain applications).
- Investigating which type of model from machine learning can be compiled efficiently into which formalism of knowledge compilation, and determining if the *structure* of the data can help in the compilation phase (for instance, bounded-treewidth data).
- Proposing new knowledge compilation formalisms that would allow new queries to be answered, for instance formalisms that could mix logical and statistical approaches.
- Implementing the algorithms that we will have developed to experiment on real datasets.

[1] Also sometimes called a *prime implicant*.

Knowledge compilation to interpret (relational) query answers. Knowledge compilation has also been used in the context of relational databases query evaluation. It has been shown to be helpful for instance for *probabilistic query evaluation* [SORK11, JS13], or for *enumerating* query results with small delay [ABJM17]. There again, a database D and a query q are first compiled into a representation R from knowledge compilation, and R is then used to solve the desired task. The idea for this second axis of the proposal is to investigate how R could also be used to explain the query results. For instance, it has recently been shown that certain variants of *Shapley-values* can be computed efficiently on formalisms of knowledge compilation [ABM20]; and, independently, Shapley values have been proposed as a framework to quantify the importance of database facts to a query result [LBKS19, RKL20]. An interesting research direction would then be to show that the tractable cases identified by [LBKS19, RKL20] can be recaptured and even extended by using the knowledge compilation approach, and by developing implementation to show the practicality of this approach. Similarly, we would like to understand if the results for the various interpretability queries that have been considered for AI models (such as sufficient reasons or minimum cardinality changes) through knowledge compilation can be transferred into the relational databases world. For instance, a framework has recently been introduced to explain query answers in the context of *ontology-mediated query answering* [CLMV20], that actually uses the same notion of sufficient explanation that is considered in AI, so it would be interesting to investigate if a knowledge compilation approach for this could work.

2 Context and PhD advisors

The PhD will be carried out in LINKS[2], which is a joint research team between Inria Lille[3], the University of Lille[4], and the CRIStAL laboratory[5]. It would be supervised by Mikaël Monet[6] and co-supervised by Florent Capelli[7] and Sławek Staworko[8]. Mikaël Monet is a recently recruited Inria full-time researcher who works on theoretical aspects of uncertain data management, knowledge compilation, and more recently on applying symbolic and logical approaches to explainable AI. Florent Capelli is a *maître de conférences* at University of Lille and works on knowledge compilation and on solving complex aggregation problems for databases. Sławek Staworko is also maître de conférences at University of Lille, holds an *habilitation à diriger des recherches* since 2016, and works on graph databases, data quality and consistency, as well as on inference problems of database schemas.

References

[ABJM17] Antoine Amarilli, Pierre Bourhis, Louis Jachiet, and Stefan Mengel. A circuit-based approach to efficient enumeration. In *44th International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 111:1–111:15, 2017.

[ABM20] Marcelo Arenas, Pablo Barceló Leopoldo Bertossi, and Mikaël Monet. The tractability of SHAP-scores over deterministic and decomposable Boolean circuits. *arXiv preprint arXiv:2007.14045*, 2020.

[2] https://team.inria.fr/links/
[3] https://www.inria.fr/en/centre-inria-lille-nord-europe
[4] https://www.univ-lille.fr/home/
[5] https://www.cristal.univ-lille.fr/en/
[6] http://mikael-monet.net/
[7] http://florent.capelli.me/
[8] http://researchers.lille.inria.fr/ staworko/

[AKM20] Gilles Audemard, Frédéric Koriche, and Pierre Marquis. On tractable XAI queries based on compiled representations. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 17, pages 838–849, 2020.

[BMPS20] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. Model interpretability through the lens of computational complexity. *Advances in Neural Information Processing Systems*, 33, 2020.

[CLMV20] Ismail Ilkan Ceylan, Thomas Lukasiewicz, Enrico Malizia, and Andrius Vaicenavicius. Explanations for ontology-mediated query answering in description logics. In *Proceedings of the 24th European conference on artificial intelligence (ECAI 2020)*, 2020.

[CSGD20] Arthur Choi, Andy Shih, Anchal Goyanka, and Adnan Darwiche. On symbolically encoding the behavior of random forests. *arXiv preprint arXiv:2007.01493*, 2020.

[DM02] Adnan Darwiche and Pierre Marquis. A knowledge compilation map. *Journal of Artificial Intelligence Research*, 17:229–264, 2002.

[JS13] Abhay Jha and Dan Suciu. Knowledge compilation meets database theory: compiling queries to decision diagrams. *Theory of Computing Systems*, 52(3):403–440, 2013.

[LBKS19] Ester Livshits, Leopoldo Bertossi, Benny Kimelfeld, and Moshe Sebag. The Shapley value of tuples in query answering. *arXiv preprint arXiv:1904.08679*, 2019.

[RKL20] Alon Reshef, Benny Kimelfeld, and Ester Livshits. The Impact of Negation on the Complexity of the Shapley Value in Conjunctive Queries. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 285–297, 2020.

[SCD18a] Andy Shih, Arthur Choi, and Adnan Darwiche. Formal verification of Bayesian network classifiers. In *International Conference on Probabilistic Graphical Models*, pages 427–438, 2018.

[SCD18b] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining Bayesian network classifiers. In *Proceedings IJCAI*, pages 5103–5111, 2018.

[SORK11] Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. *Probabilistic Databases*. Morgan & Claypool, 2011.

[SSDC20] Weijia Shi, Andy Shih, Adnan Darwiche, and Arthur Choi. On tractable representations of binary neural networks. *arXiv preprint arXiv:2004.02082*, 2020.