Rule-Based Explanations of Machine Learning Classifiers Using Knowledge Graphs

Orfeas Menis Mastromichalakis, Edmund Dervakos, Alexandros Chortaras, Giorgos Stamou

Artificial Intelligence and Learning Systems Laboratory
School of Electrical and Computer Engineering
National Technical University of Athens
menorf@ails.ece.ntua.gr, eddiedervakos@islab.ntua.gr, {achort,gstam}@cs.ntua.gr

Abstract

The use of symbolic knowledge representation and reasoning as a way to resolve the lack of transparency of machine learning classifiers is a research area that has lately gained a lot of traction. In this work, we use knowledge graphs as the underlying framework providing the terminology for representing explanations for the operation of a machine learning classifier escaping the constraints of using the features of raw data as a means to express the explanations, providing a promising solution to the problem of the understandability of explanations. In particular, given a description of the application domain of the classifier in the form of a knowledge graph, we introduce a novel theoretical framework for representing explanations of its operation, in the form of query-based rules expressed in the terminology of the knowledge graph. This allows for explaining opaque black-box classifiers, using terminology and information that is independent of the features of the classifier and its domain of application, leading to more understandable explanations but also allowing the creation of different levels of explanations according to the final end-user.

Introduction

Machine learning systems' explanations need to be represented in a human-understandable form, employing the standard domain terminology and this is why symbolic AI systems play a key role in the eXplainable AI (XAI) field of research (Murdoch et al. 2019; Guidotti et al. 2019; Arrieta et al. 2020). Of great importance in the area are the so-called *rule-based explanation* methods. Many of them rely on statistics to generate lists of if-then rules which mimic the behaviour of a classifier (Yang, Rudin, and Seltzer 2017; Ming, Qu, and Bertini 2019), or extract rules in the form of decision trees (Craven and Shavlik 1995; Confalonieri et al. 2019), while some methods make use of logics (Lehmann, Bader, and Hitzler 2010; Sarker et al. 2017) and extract rules in a form that it can be argued to be the desirable form of explanations (Pedreschi et al. 2019). Some recent methods utilize additional information about the data (such as objects depicted in an image) (Ciravegna et al. 2020), or external semantic information for the data (Panigutti, Perotti, and Pedreschi 2020). The requirement for expressing explanations in terms of domain knowledge with formal semantics has motivated the use of knowledge graphs (KG) (Hogan et al. 2020) in XAI (Tiddi and Schlobach 2022). Knowledge graphs allow for the development of mutually agreed-upon terminology to describe a domain in a human-understandable and computer-readable manner. In this respect, knowledge graphs have emerged as a promising complement or extension to machine learning approaches for explainability (Lecue 2019) like explainable recommender systems (Ai et al. 2018) that make use of knowledge representation, explainable Natural Language Processing pipelines (Silva, Freitas, and Handschuh 2019) which utilize knowledge graphs such as WordNet, and computer vision approaches, explainable by incorporating external knowledge (Alirezaie et al. 2018).

Following this line of work, we approach the problem of explaining the operation of opaque, black-box deep learning classifiers as follows: a) using domain knowledge, we construct a set of characteristic semantically-described items in the form of a knowledge graph (which we call an *explanation dataset*), b) we check the output of the unknown classifier against these items, and c) we describe the common characteristics of class instances in a human-understandable form (as if-then rules). For the latter, we propose a novel framework for representing global, post hoc explanations as first-order logic expressions produced through *semantic queries* over the knowledge graph, covering interesting common properties of the class instances. In this way, the problem of extracting logical rules is approached as a *semantic query reverse engineering problem*. Specifically, in order to extract rules of the form "if an image depicts X then it is classified as Y", we acquire the set of items classified as "Y" and then we reverse engineer a semantic query bound to have this set of items as certain answers. The use of semantic queries in our framework allows us to utilize the strong theoretical and practical results in the area of semantic query answering (Calvanese et al. 2007; Trivela et al. 2020). The query reverse engineering problem has been studied in the context of databases (Tran, Chan, and Parthasarathy 2014) and more recently of SPARQL queries (Arenas, Diaz, and Kostylev 2016). Recent studies also consider similar reverse engineering approaches in order to separate data or define queries to describe them (Jung et al. 2020; Cima, Croce, and Lenzerini 2021). Here, we use expressive knowledge graphs, where the answers to queries are considered

to be *certain answers*, whose computation involves reasoning. This approach through the semantic queries allows us to produce explanations for complex problems that other methods struggle with even simple cases, as we can also see through our experiments (see for example the experiments with CLEVR-Hans3 where our method respects the set nature of the problem, while other methods fail to represent properly the concept of sets).

Recent works have adopted a methodology akin to ours for achieving explainability, leveraging knowledge graphs, and framing the problem as a query reverse engineering task, facilitated by heuristic algorithms (Liartis et al. 2021, 2023). Our work is also relative to other global explanation methods, for instance, in the computer vision domain, global explanations of classifiers typically have the form of concept attribution or concept importance methods (Ghorbani et al. 2019; Wu et al. 2020), which extract important concepts (such as "black and white stripes") and present them as global explanations. Our proposed approach can be an alternative to methods such as concept attribution, offering more expressive explanations than concepts (e.g. including relations), independence of data features, and structured representation of terminology.

Our approach attempts to exceed two major drawbacks that most existing explanation methods have. Concerning the vocabulary they use, most approaches generate rules in terms of the feature space of the black-box classifier. However, it is argued that when the feature space of the classifier is sub-symbolic raw data, providing explanations in terms of features might lead to unintuitive, or even misleading results (Mittelstadt, Russell, and Wachter 2019). Through the proposed explanation dataset that contains the semantic description of the raw data, we can control the vocabulary used for the explanations, curating it so that the terminology is useful, understandable, and intuitive to the end user. Another major problem that is more rarely discussed in literature, is that there is not a "universal" explanation that fits all the end users. There might exist multiple explanations for an event because there might exist multiple causes, so we need to provide the right explanations to the right people. For this, we will need to gain insights from the social sciences as discussed in (Miller 2019), with the following example from (Hanson 1965) showing how the existence of multiple explanations, raises the question of *relevance*. In a fatal car accident *"consider how the cause of death might have been set out by the physician as 'multiple haemorrhage', by the barrister as 'negligence on the part of the driver', by the carriage-builder as 'a defect in the brakelock construction', by a civic planner as 'the presence of tall shrubbery at that turning'. None is more true than any of the others, but the particular context of the question makes some explanations more relevant than others.*" Choosing the relevant explanation for the end user can be vital for understandability and trustworthiness. In the above example, the different explanations are not parts of one explanation, we can imagine them as explanations at different *levels*. XAI methods need to be able to differentiate according to such different levels in order to be able to provide relevant explanations. For example, in the case of a medical assisting AI system, the doctor that uses the system would expect completely different information regarding the operation of the model, compared to the AI engineer that designed and implemented it. These two end-user groups require different levels of explanations and mixing information from different levels may prove to be misleading and non-understandable. The proposed explanation dataset allows for the selection not only of the terminology used in the explanations but also the appropriate information so that the produced explanations are relevant according to the end users.

Our approach allows us to produce *global* explanations for classifiers in any domain, as long as there are available semantic descriptions of the data. This is crucial for our method since the quality of the explanations heavily depends on this semantic description. The use of enriched data to explain deep learning models is an approach that is gaining more and more ground on the area, because there has been much controversy over black-box explanation methods that utilize only raw data to mimic the behaviour of a classifier (Rudin 2018), claiming that if the mime (explainer) is actually good, we should get rid of the opaque model and use the mime instead, so there is no point on creating mime explainers that utilize the same raw data as the model under investigation does, because they are either not good enough, or the model under investigation is redundant. We consider our method to be a form of explanation and we call it that way in order to be compatible with the existing literature, but as suggested in (Rudin 2018), a more appropriate and more descriptive title would be "summary of predictions" or "summary statistics", since our method is independent of the classifier features, and the explanations are produced by summarizing the predictions of the classifier or alternatively by a statistical analysis of the output of the model.

The rest of the paper is structured as follows: First, we introduce the background material and notation. Next, we describe the proposed theoretical framework for rule-based explanations in terms of knowledge. Then we discuss our approach to the problem as a query reverse engineering task. After that, we present the experiments including a comparative evaluation. We discuss the merits of providing explanations in terms of knowledge, we show that our approach performs similarly with state-of-the-art rule-based explainers in settings where the classifier is feature-based while showing a clear improvement in more realistic settings of deep learning classifiers (with raw data like images as input), especially in the presence of expressive knowledge. Finally, we summarize our main contributions and propose directions for future research.

Background

Description Logics The framework is described using the formalism of Description Logics (Baader et al. 2003), allowing for explanations that utilize expressive knowledge, and reasoning. Specifically, let $\mathcal{V} = \langle \mathsf{CN}, \mathsf{RN}, \mathsf{IN} \rangle$ be a *vocabulary*, where CN, RN, IN are mutually disjoint finite sets of *concept*, *role* and *individual* names, respectively. For example, a concept name might be Dog, an individual name could be the name of a specific dog, for example $\mathsf{snoopy\_42}$, and a role name could represent a specific relation, such

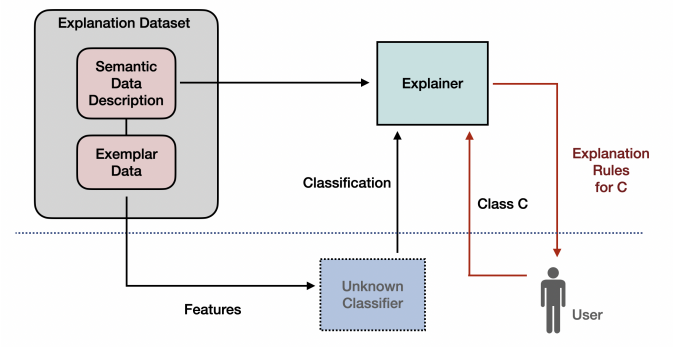

Figure 1: Explainer in operation

as hasParent. Let also $\mathcal{T}$ and $\mathcal{A}$ be a terminology (TBox) and an assertional database (ABox), respectively, over $\mathcal{V}$. The ABox contains assertions of the form $C(a)$, or $r(a, b)$, $C \in \mathsf{CN}, a, b \in \mathsf{IN}, r \in \mathsf{RN}$, for example it might contain the assertion Dog(snoopy_42), indicating that snoopy_42 is a dog. The TBox contains terminological axioms that use constructors of the specific description logics dialect, and elements of $\mathcal{V}$. For example, a TBox might contain the axiom $\mathsf{Dog} \sqsubseteq \mathsf{Mammal}$, indicating that all dogs are mammals, and $\mathsf{Dog} \sqsubseteq \exists \mathsf{hasParent}.\mathsf{Dog}$, indicating that all dogs have a parent that is itself a dog. The pair $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ is a *(DL) knowledge base* (KB). The semantics of KBs are defined in the standard model theoretical way using interpretations. Given a non-empty domain Δ, an interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ assigns a set $C^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ to each $C \in \mathsf{CN}$, a set $r^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ to each $r \in \mathsf{RN}$, and an $a^{\mathcal{I}} \in \Delta$ to each $a \in \mathsf{IN}$. $\mathcal{I}$ is a *model* of a KB $\mathcal{K}$ iff it satisfies all assertions in $\mathcal{A}$ and all axioms in $\mathcal{T}$.

Conjunctive Queries In this paper, we focus on single answer variable conjunctive queries. Given a vocabulary $\mathcal{V}$, such a query is an expression $q(x) = \{x | \exists y_1 \dots \exists y_k (c_1 \wedge \dots \wedge c_n)\}$, where x is the answer variable, y_i are variables, and each c_i is an atom $C(u)$ or $r(u, v)$, where $C \in \mathsf{CN}$, $r \in \mathsf{RN}$, and u, v are variables. A query q_2 *subsumes* a query q_1 (we write $q_1 \leq_S q_2$) iff there is a substitution θ s.t. $q_2\theta \subseteq q_1$. If q_1, q_2 are mutually subsumed, they are *syntactically equivalent*. Given a KB $\mathcal{K}$, an individual a is a *certain answer* for a query q over $\mathcal{K}$, if in every model $\mathcal{I}$ of $\mathcal{K}$, there is a match π for q such that $\pi(x) = a^{\mathcal{I}}$. We denote the set of certain answers (*answer set*) to q by $\mathsf{cert}(q, \mathcal{K})$. For example, the query $q(x) = \mathsf{Mammal}(x)$ will have as certain answers all individuals that are Mammals (according to $\mathcal{K}$) so in our case, $\mathsf{cert}(q, \mathcal{K}) = \mathsf{snoopy\_42}$.

Rules A (definite Horn) *rule* is a First Order Logic (FOL) expression of the form $\forall x_1 \dots \forall x_n \ (c_1, \dots, c_n \Rightarrow c_0)$, usually written as $c_1, \dots, c_n \rightarrow c_0$, where the c_is are atoms and x_i all appearing variables. In a rule over a vocabulary $\mathcal{V}$, each c_i is either $C(u)$ or $r(u, v)$, where $C \in \mathsf{CN}$, $r \in \mathsf{RN}$.

Classifiers A classifier is viewed as a function $F : \mathcal{D} \rightarrow \mathcal{C}$, where $\mathcal{D}$ is a domain of item feature data (e.g. images, audio, text), and $\mathcal{C}$ a set of classes (e.g. Dog, Cat).

Rule-based Global Explanations

Our approach to the extraction of rule-based global explanations is shown in Fig. 1. The *explainer* takes as input the output of an *unknown classifier* to specific items (the *exemplar data*) and a *class* C from the user and computes *explanation rules for* C, in the form of definite Horn rules. The explanation rules are expressed using a standard vocabulary (e.g. terms from domain ontologies), which should be understandable and useful to the end-user. To compute the explanation rules, the explainer has access also to *semantic data descriptions* associated with the exemplar data items, expressed in the same vocabulary. The exemplar data, which are the items fed to the unknown classifier, and their associated semantic data descriptions comprise an *explanation dataset*.

In this paper, we consider semantic data descriptions that are expressed as DL knowledge bases, and in order to compute the explanation rules, we use semantic query answering technologies, taking advantage of the semantic interrelation of rules and conjunctive queries over DL knowledge bases (Motik and Rosati 2010). Intuitively, given a class C and using semantic query answering, the explainer computes and expresses as rules the conjunctive queries that have as answers individuals representing the exemplar data items that the unknown classifier classifies as C. Because the exemplar data are consumed by the classifier, we consider that each exemplar data item consists of all the information that the classifier needs to classify it (the necessary *features*). The association of a semantic data description to each such item is modeled by the explanation dataset.

Definition 1 (Explanation Dataset)**.** *Let $\mathcal{D}$ be a domain of item feature data, $\mathcal{C}$ a set of classes, and $\mathcal{V} = \langle \mathsf{IN}, \mathsf{CN}, \mathsf{RN} \rangle$ a vocabulary such that $\mathcal{C} \cup \{\mathsf{Exemplar}\} \subseteq \mathsf{CN}$. Let also $\mathsf{EN} \subseteq \mathsf{IN}$ be a set of* exemplars*. An* explanation dataset *$\mathcal{E}$ in terms of $\mathcal{D}$, $\mathcal{C}$, $\mathcal{V}$ is a tuple $\mathcal{E} = \langle \mathcal{M}, \mathcal{S} \rangle$, where $\mathcal{M} : \mathsf{EN} \rightarrow \mathcal{D}$ is a mapping from the exemplars to the item feature data, and $\mathcal{S} = \langle \mathcal{T}, \mathcal{A} \rangle$ is a DL KB over $\mathcal{V}$ such that $\mathsf{Exemplar}(a) \in \mathcal{A}$ iff $a \in \mathsf{EN}$, the elements of $\mathcal{C}$ do not appear in $\mathcal{S}$, and* Exemplar *and the elements of* EN *do not appear in $\mathcal{T}$.*

Intuitively, $\mathcal{D}$ contains items that can be fed to a classifier. Each such item is represented in the associated semantic data description by an individual (exemplar) $a \in \mathsf{EN}$, which is mapped to the respective feature data by $\mathcal{M}$. The knowledge base $\mathcal{S}$ contains the semantic data descriptions about all individuals in EN. The concept Exemplar is used solely to identify the exemplars within $\mathcal{A}$ (since other individual may exist) and should not appear elsewhere. The classes $\mathcal{C}$ should not appear in $\mathcal{S}$ so as not to take part in any reasoning process. The explanation dataset thus provides items with which we can probe the black-box classifier to explain it, by making use of the semantic descriptions of the items, in the context of the underlying knowledge.

Given an explanation dataset, an unknown classifier, and a class C, the aim of the explainer is to detect the semantic properties and relations of the exemplar data items that are classified by the unknown classifier to class C, and represent them in a human-understandable form, as rules utilizing the terminology of the knowledge.

Definition 2 (Explanation Rule)**.** *Let $F : \mathcal{D} \to \mathcal{C}$ be a classifier , $\mathcal{E} = \langle \mathcal{M}, \mathcal{S} \rangle$ an explanation dataset in terms of $\mathcal{D}$, $\mathcal{C}$ and an appropriate vocabulary $\mathcal{V} = \langle \mathsf{CN}, \mathsf{RN}, \mathsf{IN} \rangle$. Given a concept $C \in \mathcal{C}$, the rule*

$$\mathsf{Exemplar}(x), c_1, c_2, \ldots, c_n \to C(x)$$

where c_i is an atom $D(u)$ or $r(u, v)$, where $D \in \mathsf{CN}$, $r \in \mathsf{RN}$, and u, v are variables, is an explanation rule *of F for class C over $\mathcal{E}$. We denote the rule by $\rho(F, \mathcal{E}, C)$, or simply by ρ whenever the context is clear. We may also omit $\mathsf{Exemplar}(x)$ from the body, since it is a conjunct of any explanation rule.*

Explanation rules describe *sufficient* conditions for an item to be classified in class C by a classifier. E.g., if the classifier classified images depicting wild animals in a zoo class, an explanation rule could be $\mathsf{Exemplar}(x), \mathsf{Image}(x), \mathsf{depicts}(x, y), \mathsf{WildAnimal}(y) \to \mathsf{ZooClass}(x)$, assuming that $\mathsf{Image}, \mathsf{WildAnimal} \in \mathsf{CN}$, $\mathsf{depicts} \in \mathsf{RN}$, and $\mathsf{ZooClass} \in \mathcal{C}$. It is important that explanation rules refer only to individuals $a \in \mathsf{EN}$ that correspond to items $\mathcal{M}(a) \in \mathcal{D}$; this is guaranteed by the conjunct $\mathsf{Exemplar}(x)$ in the explanation rule body.

Given a classifier $F : \mathcal{D} \to \mathcal{C}$ and a set of individuals $\mathcal{I} \subseteq \mathsf{EN}$, the positive set (pos-set) of F on $\mathcal{I}$ for class $C \in \mathcal{C}$ is $\mathsf{pos}(F, \mathcal{I}, C) = \{a \in \mathcal{I} : F(\mathcal{M}(a)) = C\}$.

Definition 3 (Explanation Rule Correctness)**.** *Let $F : \mathcal{D} \to \mathcal{C}$ be a classifier, $\mathcal{E} = \langle \mathcal{M}, \mathcal{S} \rangle$ an explanation dataset in terms of $\mathcal{D}$, $\mathcal{C}$ and an appropriate vocabulary $\mathcal{V}$, and $\rho(F, \mathcal{E}, C)$ an explanation rule. The rule ρ is* correct *over F and $\mathcal{E}$ if and only if*

$$\mathsf{fol}(\mathcal{S} \cup \{\mathsf{Exemplar} \sqsubseteq \{a \mid a \in \mathsf{EN}\}\} \cup \{C(a) \mid a \in \mathsf{pos}(F, \mathsf{EN}, C)\}) \models \rho$$

where $\mathsf{fol}(\mathcal{K})$ is the first-order logic translation of DL KB $\mathcal{K}$.

For the rest of the paper we will only consider correct rules, and when F and $\mathcal{E}$ are not ambiguous we will refer to correct rules over F and $\mathcal{E}$ simply as *correct rules*. The intended meaning of a correct explanation rule is that for every $a \in \mathsf{EN}$, if the body of the rule holds, then the classifier classifies $\mathcal{M}(a)$ to the class indicated in the head of the rule. Intuitively, an explanation rule is correct if it is a logical consequence of the underlying knowledge extended by the axiom $\mathsf{Exemplar} \sqsubseteq \{a \mid a \in \mathsf{EN}\}$ (which forces $\mathsf{Exemplar}(x)$ to be true in an interpretation $\mathcal{I}$ only for $x = a^{\mathcal{I}}$ with $a \in \mathsf{EN}$). For instance, the rule of the previous example $\mathsf{Exemplar}(x), \mathsf{Image}(x), \mathsf{depicts}(x, y), \mathsf{WildAnimal}(y) \to \mathsf{ZooClass}(x)$ would be correct for the KB $\mathcal{S}_1 = \langle \mathcal{T}_1, \mathcal{A}_1 \rangle$, where $\mathcal{A}_1 = \{\mathsf{Image(a)}, \mathsf{depicts(a, b)}, \mathsf{Wolf(b)}\}$ and $\mathcal{T}_1 = \{\mathsf{Wolf} \sqsubseteq \mathsf{WildAnimal}\}$ if $\mathsf{a} \in \mathsf{pos}(F, \mathsf{EN}, \mathsf{ZooClass})$, while it would not be correct for the KB $\mathcal{S}_2 = \langle \emptyset, \mathcal{A}_1 \rangle$, nor would it be correct for $\mathcal{S}_1$ if $\mathsf{a} \notin \mathsf{pos}(F, \mathsf{EN}, \mathsf{ZooClass})$. Checking whether a rule is correct is a reasoning problem which can be solved by using standard DL reasoners. On the other hand, finding rules which are correct is an inverse problem that is much harder to solve.

Explanation Rules From Queries

There is a clear resemblance between a query as described in the Background section, and the body of an explanation rule as defined in Def. 2. Thus, by representing the bodies of explanation rules as queries, the computation of explanations can be treated as a query reverse engineering problem.

Definition 4 (Explanation Rule Query)**.** *Let $F : \mathcal{D} \to \mathcal{C}$ be a classifier, $\mathcal{E} = \langle \mathcal{M}, \mathcal{S} \rangle$ an explanation dataset in terms of $\mathcal{D}$, $\mathcal{C}$ and an appropriate vocabulary $\mathcal{V}$, and $\rho(F, \mathcal{E}, C)$: $\mathsf{Exemplar}(x), c_1, c_2, \ldots, c_n \to C(x)$ an explanation rule. The query*

$$q_\rho \doteq \{\mathsf{Exemplar}(x), c_1, c_2, \ldots, c_n\}_x$$

is the explanation rule query *of explanation rule ρ.*

This definition establishes a 1-1 relation (up to variable renaming) between ρ and q_ρ. To compute queries corresponding to explanation rules that are guaranteed to be correct, we prove Theorem 1 (see the supplementary material for proof).

Theorem 1. *Let $F : \mathcal{D} \to \mathcal{C}$ be a classifier, $\mathcal{E} = \langle \mathcal{M}, \mathcal{S} \rangle$ an explanation dataset in terms of $\mathcal{D}$, $\mathcal{C}$ and an appropriate vocabulary $\mathcal{V}$, $\rho(F, \mathcal{E}, C)$: $\mathsf{Exemplar}(x), c_1, c_2, \ldots, c_n \to C(x)$ an explanation rule, and q_ρ the explanation rule query of ρ. The explanation rule ρ is correct if and only if*

$$\mathsf{cert}(q_\rho, \mathcal{S}) \subseteq \mathsf{pos}(F, \mathsf{EN}, C)$$

Theorem 1 allows us to compute guaranteed correct rules, by finding a query q for which $\mathsf{cert}(q, \mathcal{S}) \subseteq \mathsf{pos}(F, \mathsf{EN}, C)$. Intuitively, an explanation rule query is correct for class C, if all of its certain answers are mapped by $\mathcal{M}$ to feature data which is classified in class C. It follows that a query with one certain answer which is an element of the pos-set is a correct rule query, as is a query q for which $\mathsf{cert}(q, \mathcal{S}) = \mathsf{pos}(F, \mathsf{EN}, C)$. Thus, it is useful to define a *recall* metric for explanation rule queries by comparing the set of certain answers with the pos-set of a class C:

$$\mathsf{recall}(q, \mathcal{E}, C) = \frac{|\mathsf{cert}(q, \mathcal{S}) \cap \mathsf{pos}(F, \mathsf{EN}, C)|}{|\mathsf{pos}(F, \mathsf{EN}, C)|}.$$

Given the above, one approach to the problem of finding explanation rules for an explanation dataset is to reduce it to forming candidate queries, computing their answers, and assessing the correctness of the corresponding rules. The computation of arbitrary candidate explanation rule queries for the KB $\mathcal{S}$ of an explanation dataset is in general hard since it involves exploring the query space $\mathcal{Q}$ of all queries that can be constructed using the underlying vocabulary $\mathcal{V}$ and getting their certain answers for $\mathcal{S}$. Difficulties arise even in simple cases since the query space is in general infinite. However, the set of all possible distinct answer sets is finite, and in most cases, it is expected to be much smaller than its upper limit, the powerset 2^{IN}. There are works that approach this problem in a similar way to ours, but employ heuristic algorithms in order to alleviate the computational complexity of the problem (Liartis et al. 2021). In this work, we use a simple exhaustive algorithm based on (Chortaras, Giazitzoglou, and Stamou 2019) in order to maintain the guarantees of the framework and showcase its potential, without being concerned about computational optimizations which are beyond the scope of this paper. We briefly describe and discuss this algorithm below for transparency and reproducibility reasons.

Algorithm 1: QuerySpaceDAG

Data: Vocabulary $\mathcal{V}$, KB $\mathcal{K}$, a maximum query depth $k \geq 0$
Result: Query space DAG $\mathcal{G}$
Compute the set $\mathcal{B}$ of all non-syntactically equivalent queries $\{C_1(x), \ldots, C_n(x)\}_x$, where $C_i \in \mathsf{CN} \setminus \{\mathsf{Exemplar}\}$, $n \geq 1$.
Compute the set $\mathcal{F}$ of all non-syntactically equivalent queries $\{r_1(u_1, v_1), \ldots, r_n(u_n, v_n)\}_{x,y}$, where $r_i \in \mathsf{RN}$, $n \geq 1$, each u_i, v_i is either x or y and $u_i \neq v_i$.
Initialize an empty set of queries $\mathcal{Q}$.
for $i = 0 \ldots k$ **do**
  Compute the set $\mathcal{T}_i$ of all trees of depth i.
  foreach $t \in \mathcal{T}_i$ **do**
    Assign to each node v of t a distinct variable $\mathsf{var}(v)$. Assign x to the root of t.
    Construct all non-syntactically equivalent queries q obtained from t by adding to the body of q: i) for each node v of t, the body of an element of $\mathcal{B} \cup \{\emptyset\}$ after renaming x to $\mathsf{var}(v)$, ii) for each edge (v_1, v_2) of t, the body of an element of $\mathcal{F}$ after renaming x to $\mathsf{var}(v_1)$ and y to $\mathsf{var}(v_2)$, and iii) $\mathsf{Exemplar}(x)$.
    Condense all qs and add them to $\mathcal{Q}$.
  end
end
while there are $q_1, q_2 \in \mathcal{Q}$ s.t. $\mathsf{cert}(q_1, \mathcal{K}) = \mathsf{cert}(q_2, \mathcal{K})$ **do**
  remove q_1, q_2 from $\mathcal{Q}$ and add $q_1 \sqcap q_2$ to $\mathcal{Q}$.
end
Arrange the elements of $\mathcal{Q}$ in a DAG $\mathcal{G}$, making q_1 a child of q_2 iff $\mathsf{cert}(q_1, \mathcal{K}) \subset \mathsf{cert}(q_2, \mathcal{K})$.
return the transitive reduction of $\mathcal{G}$

Alg. 1 explores a useful finite subset of $\mathcal{Q}$, namely the tree-shaped queries of a maximum depth k (Glimm et al. 2007). It constructs all possible such queries (that include $\mathsf{Exemplar}(x)$ in the body), obtains their answers, and arranges them in a directed acyclic graph (the *query space DAG*) using the subset relation on the answer sets. The queries are constructed in the for loop, and then the while loop replaces queries having the same answer set by their intersection. The *intersection* $q_1 \sqcap q_2$ of two instance queries q_1, q_2 with answer variable x is the query $\mathsf{cond}(q_1 \cup q_2\theta)$, where θ renames each variable appearing in q_2 apart from x to a variable not appearing in q_1. Thus, from all possible queries with the same answers, the algorithm keeps only the *most specific* query q of all such queries. Intuitively, this is the most detailed query. Finally, the queries are arranged in a DAG. By construction, each node of the DAG is a query representing a distinct answer set.

Theorem 2. *Let $F : \mathcal{D} \to \mathcal{C}$ be a classifier, $\mathcal{E} = \langle \mathcal{M}, \mathcal{S} \rangle$ an explanation dataset in terms of $\mathcal{D}$, $\mathcal{C}$ and an appropriate vocabulary $\mathcal{V}$, and $\rho(F, \mathcal{E}, C)$ a correct tree-shaped explanation rule of maximum depth k. The DAG constructed by Alg. 1 contains a query $q_{\rho'}$ corresponding to a correct explanation rule $\rho'(F, \mathcal{E}, C)$ with the same recall as ρ, s.t. $q_{\rho'} \leq_S q_\rho$.*

Given Theorem 2 (see the supplementary material for proof) the nodes corresponding to correct rules for some $\mathsf{pos}(F, \mathsf{EN}, C)$ can be reached by traversing the graph starting from the root and finding the first nodes whose answer sets are subsets of $\mathsf{pos}(F, \mathsf{EN}, C)$. These nodes correspond to the explanation rules with the highest recall whose underlying queries do not subsume each other, and the descendants of that nodes provide all subsumed queries corresponding to correct explanation rules with smaller recall. The DAG has a unique root because answer sets are subsets of $\mathsf{cert}(\{\mathsf{Exemplar}(x)\}_x, \mathcal{S})$.

Experiments and Evaluation

In this section, we evaluate the proposed approach, which we call KGrules. We conduct experiments on tabular and image data, investigating how explanation datasets of different sizes and expressivities affect the explanations, we compare our work with other rule-based explanation methods, and discuss the quality and usability of the results.

Tabular Classifier

The first set of experiments is conducted on the Mushroom[1] dataset which contains data in tabular form with categorical features. Our proposed approach is overkill for such a dataset, since its representation as a DL KB does not contain roles nor a TBox, however on this dataset we can compare the proposed method with the state-of-the-art. To represent the dataset as a knowledge base in order to run Alg.1, we create a concept for each combination of categorical feature name and value, ending up with $|\mathsf{CN}| = 123$ and an individual for each row of the dataset. Then we construct an ABox where the type of each individual is asserted based on the values of its features and the aforementioned concepts. To measure the quality of the generated explanations, the dataset is split in three parts: (I) A classification-training set on which we train a simple two-layer Multi-Layer Perceptron (MLP) classifier (that achieved 100% test accuracy), (II) an explanation-training set which we use to generate explanations for the predictions of the classifier with the methods under evaluation, and (III) an explanation-testing set on which we measure the fidelity (ratio of input instances on which the predictions of the model and the rules agree, over total instances) of the rules. We also measure the number of rules and the average rule length for each case. We compare our method with Skope-Rules[2] and rule-matrix (Ming, Qu, and Bertini 2019) which implements scalable bayesian rule lists (Yang, Rudin, and Seltzer 2017), on different sizes of explanation-training sets. The results are shown in Table 1. All methods perform similarly with respect to fidelity, with no clear superiority of any method since they all achieve near perfect performance, probably because of the simplicity of the dataset. Despite the fact that our method is not tailored to such problems like the other two methods, we can see that it has similar performance to the state of the art for rule-based explanations of tabular classifiers. However, although our method is much more expressive than others (and we can see it in the following subsections where our method is able to explain complex set-related classifiers much more

[1] https://archive.ics.uci.edu/ml/datasets/mushroom
[2] https://github.com/scikit-learn-contrib/Skope-Rules

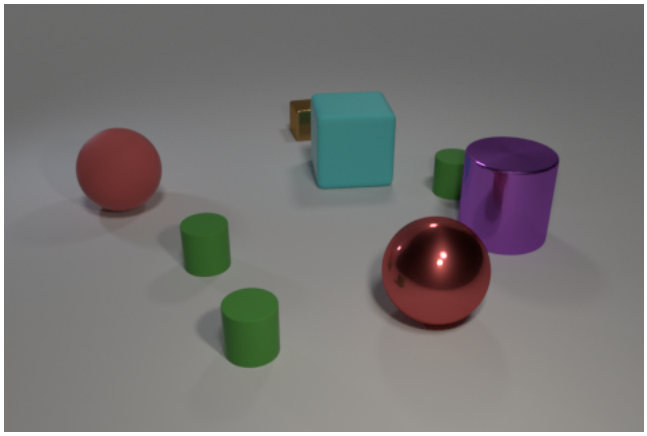

Figure 2: An image from the CLEVR-Hans3 dataset.

efficiently) certain limitations (e.g. the lack of negation) results in this case to more and slightly longer rules compared to the other two methods.

Image Classifier: CLEVR-Hans3

Although the explanation of tabular classifiers has been addressed in many cases in literature, of great interest is the (more challenging) explanation of deep learning models that take as input raw data, like images or text. Hence, for the second set of experiments, we employ CLEVR-Hans3 (Stammer, Schramowski, and Kersting 2020) which is a dataset of images with intentionally added biases in the train and validation set which are absent in the test set. For example the characteristic of the first class is that all images include a large cube and a large cylinder, but the large cube is always gray in the training and validation sets, while it has a random color in the test set. This makes it ideal for the evaluation of XAI frameworks since it creates classifiers with known biases. On this dataset, we conduct two experiments. Firstly, we explore the effect of the size of the explanation dataset by testing whether our method can predict the known description of the three classes. Secondly, we evaluate our method on a real image classifier and compare the results with other rule-based methods.

For representing available annotations as a DL KB, we define an individual name for each image and for each object depicted therein, and a concept name for each color, size, shape, and material of the objects. We also include a role name contains, to connect images to objects they depict. Then, in the ABox, we assert the characteristics of each object and link them to the appropriate images by using the role. For example, in Fig. 2 we can see a sample image from the CLEVR-Hans3 dataset with id i, which is described in the ABox of our explanation datasets with the assertions: {Exemplar(i), contains(i,o_1), Red(o_1), Sphere(o_1), Large(o_1), Rubber(o_1), contains(i,o_2), Green(o_2), Cylinder(o_2), Small(o_2), Rubber(o_2), ... }.

For comparing with other methods we also create a tabular version of the dataset, in which each object's characteristics are one-hot encoded, and an image is represented as a concatenation of the encodings of the objects it depicts.

Size	Method	Fidelity	Nr. of Rules	Avg. Length
100	KGrules	**97.56%**	11	5
	RuleMatrix	94.53%	3	2
	Skope-Rules	97.01%	3	2
200	KGrules	98.37%	11	5
	RuleMatrix	97.78%	4	2
	Skope-Rules	**98.49%**	4	2
600	KGrules	99.41%	13	4
	RuleMatrix	**99.43%**	6	1
	Skope-Rules	98.52%	4	2

Table 1: Performance on the Mushroom dataset.

Using the true labels of the data allows us to use the description of each class as ground truth explanations. Table 2 shows a condensed version of the explanation rules produced by Alg. 1 for an ideal classifier (accuracy=100%) with explanation datasets of various sizes, along with ground truth and the explanation rule with highest recall per class for a real classifier. The full explanations are obtained by adding to that condensed versions the conjuncts Exemplar(x), and contains(x,t), for all other appearing variables $t \neq x$, as well as the tail of the rule ($\rightarrow ClassX$) for the respective class X. All explanations on the ideal classifier achieved recall=100%. We can see that with explanation datasets with 600 or more exemplars we are able to predict the ground truth for all 3 classes. Even with 20 exemplars, we are able to produce the ground truth explanation for one of the classes and with 40 or more exemplars we produce ground truth explanations for 2 out of the 3 classes and almost for the third class too (only one characteristic of one object missing).

In order to produce accurate explanations it seems useful to have individuals close to the "semantic border" of the classes, i.e. individuals of different classes with similar descriptions. Intuitively, such individuals guide the algorithm to produce a more accurate explanation in a similar manner that near-border examples guide a machine learning algorithm to approximate better the separating function. Following this intuition, we experiment with two of the small explanation datasets that almost found the perfect explanations (size of 40 and 80). By strategically choosing individuals , we are able to obtain two small explanation datasets, one of size 43 and one of size 82, that when used by Alg. 1 produce the ground truth explanations for all 3 classes. This indicates the importance of the curation of the explanation dataset, which is not an easy task, and the selection of "good" individuals for the explanation dataset is not trivial.

Finally we use our framework to explain a real classifier trained on CLEVR-Hans3, and compare our explanations with Skope-Rules and RuleMatrix. The classifier we use is a ResNet34-based model, that achieved overall 99,4% validation accuracy and 71,2% test accuracy (probably due to the confounded train and validation sets). More details about the classifier's performance can be found in the supplementary material.

We curate an explanation set with 100 images so that it also accurately explains the ground truth, and we used it to

Nr. of images	Class 1	Class 2	Class 3
20	✓	Small(y), Metal(y), Cube(y)	Yellow(y), Small(y), Blue(z), Large(z), Sphere(z)
40 / 60	✓	✓	Yellow(y), Small(y), Blue(z), Large(z), Sphere(z)
80 / 100 / 200 / 400	✓	✓	Yellow(y), Small(y), Sphere(y), Blue(z), Sphere(z)
600 / 800 / 1000	✓	✓	✓
Ground Truth	(*Gray(y)*) , Large(y), Cube(y), Large(z), Cylinder(z)	Small(y), (*Metal(y)*), Sphere(y), Small(z), Metal(z), Cube(z)	Yellow(y), Small(y), Sphere(y), Blue(z), Large(z), Sphere(z)
Real Classifier	Large(y), Cube(y), Gray(z), Large(z), Large(w), Cylinder(w)	Small(y), Metal(y), Cube(y)	✓

Table 2: Explanations on CLEVR-Hans3. The concepts in parentheses are the confounding factors in the ground truth row. Check mark (✓) indicates that the explanation is the same as the ground truth (without the confounding factors).

explain the real classifier. Table 3 shows that our method significantly outperforms the other rule-based methods in terms of fidelity, with a notable smaller number of rules which is used as an indication of understandability of a rule-set. The set nature of the input data (each image contains a set of objects with specific characteristics) shows the limitations of other rule-based methods in such realistic problems. We are able to reproduce the rules created by our method using the tabular format that is fed to the other classifiers, showing that the data format is not a limitation in terms of fidelity, but it requires a large number of rules, which indicates the usefulness of rule-based methods like ours that do not only work on tabular data. Investigating the explanations produced by our method, we are also able to detect potential biases of the classifier due to the confounding factors of the dataset. For example, regarding the first class (all images contain a large cube and a large cylinder), the rule with the highest recall produced for the real classifier is: $\mathsf{contains}(x, y)$, $\mathsf{Gray}(y)$, $\mathsf{Large}(y)$, $\mathsf{contains}(x, z)$, $\mathsf{Cylinder}(z)$, $\mathsf{Large(z)} \rightarrow \mathsf{Class}_1(x)$ showing the existence of a large cylinder, and detecting the potential color bias of

Method	Fidelity	Nr. of Rules	Avg. Length
KGrules	**85.07%**	4	5
RuleMatrix	58.09%	42	2
Skope-Rules	77.18%	20	3

Table 3: Performance on the CLEVR-Hans3 dataset.

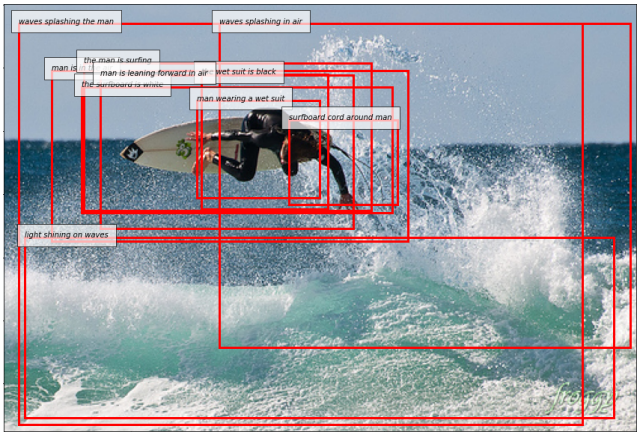

Figure 3: An image from the Visual Genome dataset.

another large object created by the intentional bias of the train and validation set (the large cube is always gray in the train and validation sets).

Image Classifier: Visual Genome

As a third experiment, we evaluate our framework on an explanation dataset of real-world images, described by an expressive knowledge, that includes roles and a TBox. Specifically, we utilize the Visual Genome dataset (VGD) (Krishna et al. 2017) which contains richly annotated images, including descriptions of regions, attributes of depicted objects and relations between them. On this dataset, we attempt to explain image classifiers trained on ImageNet. We define three super-classes of ImageNet classes which contain a) Domestic, b) Wild and c) Aquatic Animals , because they are more intuitive to perform a qualitative evaluation, when compared to the fine-grained ImageNet classes. We represent the available VGD annotations as a DL KB, where the ABox consists of the scene graphs for each image, in which each node and edge is labeled with a WordNet (WN) synset and the TBox consists of the WN hypernym-hyponym hierarchy. In the ABox we also include assertions about which objects are depicted by an image in order to connect the exemplar data with the scene graphs. For example, the image in Fig. 3 with id i is described in the ABox by the assertions: {Exemplar(i), contains(i,person$_1$), contains(i,sea$_1$), surfer.n.01(person$_1$), ocean.n.01(sea$_1$), blue.s.01(sea$_1$), travel.v.01(person$_1$,sea$_1$)} , and the TBox contains the axioms: {ocean.n.01 $\sqsubseteq$ body_of_water.n.01, surfer.n.01 $\sqsubseteq$ swimmer.n.02, ...}.

Since in the original VGD annotations are linked to WN automatically, there are errors, thus we chose to manually curate a subset of 100 images. This is closer to the intended

Network	Rules
VGG-16	$\mathsf{artifact}(y), \mathsf{dog}(z), \mathsf{brown}(w) \rightarrow \mathsf{Domestic}(x)$ $\mathsf{green}(y), \mathsf{plant}(z), \mathsf{organ}(w) \rightarrow \mathsf{Wild}(x)$ $\mathsf{whole}(y), \mathsf{ocean}(z) \rightarrow \mathsf{Aquatic}(x)$
WRN	$\mathsf{animal}(y), \mathsf{wear}(y, z), \mathsf{artifact}(z) \rightarrow \mathsf{Domestic}(x)$ $\mathsf{green}(y), \mathsf{plant}(z), \mathsf{nose}(w) \rightarrow \mathsf{Wild}(x)$ $\mathsf{surfboard}(y) \rightarrow \mathsf{Aquatic}(x)$
ResNext	$\mathsf{artifact}(y), \mathsf{dog}(z), \mathsf{brown}(w) \rightarrow \mathsf{Domestic}(x)$ $\mathsf{ear}(y), \mathsf{plant}(z), \mathsf{nose}(w) \rightarrow \mathsf{Wild}(x)$ $\mathsf{fish}(y), \mathsf{structure}(z) \rightarrow \mathsf{Aquatic}(x)$

Table 4: Explanation rules utilizing the animal explanation dataset. Rules are shown in condensed form: the full rules are obtained by adding the conjuncts $\mathsf{contains}(x, t)$ for all appearing variables $x \neq t$.

use-case of our proposed method, in which experts would curate explanation datasets for specific domains. We explain three different neural architectures[3]: VGG-16 (Simonyan and Zisserman 2014), Wide-ResNet (WRN) (Zagoruyko and Komodakis 2016) and ResNeXt (Xie et al. 2016), trained for classification on the ImageNet dataset. The context of VGD is too complex to be transformed into tabular form in a useful and valid way for the other rule-based methods. Table 4 shows the correct rules of maximum recall for each class and each classifier. We discuss three key explanations:

1. Wide ResNet: $\mathsf{surfboard}(y) \rightarrow$ *Aquatic(x)*. It seems that the classifier has a bias accepting surfer/surfboard images as aquatic animals probably due to the sea environment of the images; further investigation finds this claim to be consistent, showing the potential of this framework in detecting biases.

2. Wide ResNet: $\mathsf{animal}(y)$, $\mathsf{wear}(y, z)$, $\mathsf{artifact}(z) \rightarrow$ *Domestic(x)*. It is interesting to compare this explanation with another correct rule for the same classifier with lower recall: $\mathsf{animal}(y)$, $\mathsf{collar}(z) \rightarrow$ *Domestic(x)*. By considering roles between objects we get a more accurate (higher recall) and informative explanation, denoting the tendency of the classifier to classify as *Domestic* any animal that wears something man-made. This example shows how more complex queries enhance the insight (wearing an artifact) while less expressive ones might only see a part of it (collar). Here we can also see one of the effects of the TBox hierarchy on the explanations, since this rule covers many sub-cases (like dog wears collar, and cat wears bowtie) that would require multiple rules if it wasn't for the grouping that stems from the TBox.

3. ResNeXt: $\mathsf{nose}(y)$, $\mathsf{plant}(z)$, $\mathsf{ear}(w) \rightarrow$ *Wild*. Although this explanation provides information that is related to the natural environment of the images classified as *Wild* (plant), we see also some rather odd concepts (nose, ear). While this could be a strange bias of the classifier, it is probably a flaw of the explanation dataset. We discovered that images are not consistently annotated with body parts, like noses and ears. Thus, through the explanations, we can also detect weaknesses of the explanation set. The rules are limited by the available knowledge, so we should constantly evaluate the quality and expressivity of the knowledge that is used in order to produce accurate and useful explanations.

We also investigated explanation datasets with different levels of information like details about the image brightness, contrast, sharpness, saturation, etc. Our experimentation gives a good intuition regarding the usefulness of the different levels of explanation. Consider an application similar to the classifier used in this subsection, that classifies animals. The end-user of the application when asking *"why this animal was classified as 'elephant'"* would expect an explanation in terms of characteristics of animals, like *"because the animal has a trunk, it is grey, etc."* but not an explanation about the characteristics of the image like *"because the image has high brightness and low saturation"*, let alone an explanation that mixes all these things like *"because the image has high brightness and the animal is grey"*. The latter explanations would not answer the user's question, would probably be misleading, and would reduce the user's trust in the application, even if they did not indicate a bias. On the other hand, the AI engineer that developed this application would be very interested in explanations concerning the images' characteristics because this could indicate potential biases and guide them for further investigation.

Conclusions

In this work, we introduced a framework for representing explanations for ML classifiers in the form of rules, generated explanations for various classifiers and datasets, and compared our work with other methods. Our proposed query-based rules showed superiority in terms of expressivity compared to existing methods, producing consistent explanations in cases where other methods struggled. We believe that the transparency of the proposed explanation dataset, combined with the guarantees of framework and algorithm improve *user awareness* when compared with other rule-based explanation methods. Additionally, the control over the terminology as well as the selection of information that the explanation dataset offers allow for the creation of more understandable and relevant explanations. Regarding understandability, however, this should be evaluated in a human study, which we plan to conduct in the future. In addition, we are in the process of creating explanation datasets in collaboration with area experts for the domains of medicine and music. We are investigating what constitutes a "good" explanation dataset with regard to its size, distribution, and represented information. Finally, we are exploring improvements and optimizations for the algorithmic part, like heuristic solutions, approaches from other areas like Inductive Logic Programming (ILP), adaptations of our existing algorithm to more DL dialects, relaxations for getting approximate solutions faster, and modifications in order to generate different types of explanations such as local, counterfactual, and prototype explanations.

[3]https://pytorch.org/vision/stable/models.html

References

Ai, Q.; Azizi, V.; Chen, X.; and Zhang, Y. 2018. Learning Heterogeneous Knowledge Base Embeddings for Explainable Recommendation. *Algorithms*,11(9): 137.

Alirezaie, M.; Längkvist, M.; Sioutis, M.; and Loutfi, A. 2018. A Symbolic Approach for Explaining Errors in Image Classification Tasks. In *IJCAI 2018*.

Arenas, M.; Diaz, G. I.; and Kostylev, E. V. 2016. Reverse Engineering SPARQL Queries. In *WWW*, 239–249. ACM.

Arrieta, A. B.; Rodríguez, N. D.; Ser, J. D.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58: 82–115.

Baader, F.; Calvanese, D.; McGuinness, D.; Patel-Schneider, P.; and Nardi, D. 2003. *The description logic handbook: Theory, implementation and applications*. Cambridge university press.

Calvanese, D.; Giacomo, G. D.; Lembo, D.; Lenzerini, M.; and Rosati, R. 2007. Tractable Reasoning and Efficient Query Answering in Description Logics: The *DL-Lite* Family. *J. Aut. Reason.*, 39(3): 385–429.

Chortaras, A.; Giazitzoglou, M.; and Stamou, G. 2019. Inside the Query Space of DL Knowledge Bases. In *Description Logics*, volume 2373 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Cima, G.; Croce, F.; and Lenzerini, M. 2021. *Query Definability and Its Approximations in Ontology-Based Data Management*, 271–280. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384469.

Ciravegna, G.; Giannini, F.; Gori, M.; Maggini, M.; and Melacci, S. 2020. Human-Driven FOL Explanations of Deep Learning. In *IJCAI*, 2234–2240. ijcai.org.

Confalonieri, R.; del Prado, F. M.; Agramunt, S.; Malagarriga, D.; Faggion, D.; Weyde, T.; and Besold, T. R. 2019. An Ontology-based Approach to Explaining Artificial Neural Networks. *CoRR*, abs/1906.08362.

Craven, M. W.; and Shavlik, J. W. 1995. Extracting Tree-Structured Representations of Trained Networks. In *NIPS*, 24–30. MIT Press.

Ghorbani, A.; Wexler, J.; Zou, J. Y.; and Kim, B. 2019. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32.

Glimm, B.; Horrocks, I.; Lutz, C.; and Sattler, U. 2007. Conjunctive Query Answering for the Description Logic SHIQ. In *IJCAI*, 399–404.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5): 93:1–93:42.

Hanson, N. R. 1965. *Patterns of discovery: An inquiry into the conceptual foundations of science*. CUP Archive.

Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; de Melo, G.; Gutiérrez, C.; Gayo, J. E. L.; Kirrane, S.; Neumaier, S.; Polleres, A.; Navigli, R.; Ngomo, A. N.; Rashid, S. M.; Rula, A.; Schmelzeisen, L.; Sequeda, J. F.; Staab, S.; and Zimmermann, A. 2020. Knowledge Graphs. *CoRR*, abs/2003.02320.

Jung, J. C.; Lutz, C.; Pulcini, H.; and Wolter, F. 2020. Logical Separability of Incomplete Data under Ontologies.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.*, 123(1): 32–73.

Lecue, F. 2019. On the role of knowledge graphs in explainable AI. *Semantic Web*, 11: 1–11.

Lehmann, J.; Bader, S.; and Hitzler, P. 2010. Extracting reduced logic programs from artificial neural networks. *Appl. Intell.*, 32(3): 249–266.

Liartis, J.; Dervakos, E.; Menis-Mastromichalakis, O.; Chortaras, A.; and Stamou, G. 2021. Semantic Queries Explaining Opaque Machine Learning Classifiers. In *DAO-XAI*.

Liartis, J.; Dervakos, E.; Menis-Mastromichalakis, O.; Chortaras, A.; and Stamou, G. 2023. Searching for explanations of black-box classifiers in the space of semantic queries. *Semantic Web*, (Preprint): 1–42.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38.

Ming, Y.; Qu, H.; and Bertini, E. 2019. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Trans. Vis. Comput. Graph.*, 25(1): 342–352.

Mittelstadt, B. D.; Russell, C.; and Wachter, S. 2019. Explaining Explanations in AI. In *FAT*, 279–288. ACM.

Motik, B.; and Rosati, R. 2010. Reconciling description logics and rules. *J. ACM*, 57(5): 30:1–30:62.

Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; and Yu, B. 2019. Interpretable machine learning: definitions, methods, and applications. *CoRR*, abs/1901.04592.

Panigutti, C.; Perotti, A.; and Pedreschi, D. 2020. Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In *FAT**, 629–639. ACM.

Pedreschi, D.; Giannotti, F.; Guidotti, R.; Monreale, A.; Ruggieri, S.; and Turini, F. 2019. Meaningful Explanations of Black Box AI Decision Systems. In *AAAI*, 9780–9784. AAAI Press.

Rudin, C. 2018. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.

Sarker, M. K.; Xie, N.; Doran, D.; Raymer, M.; and Hitzler, P. 2017. Explaining Trained Neural Networks with Semantic Web Technologies: First Steps. In *NeSy*, volume 2003 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Silva, V. D. S.; Freitas, A.; and Handschuh, S. 2019. Exploring Knowledge Graphs in an Interpretable Composite Approach for Text Entailment. In *AAAI*, 7023–7030. AAAI Press.

Simonyan, K.; and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition.

Stammer, W.; Schramowski, P.; and Kersting, K. 2020. Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting with their Explanations. *arXiv preprint arXiv:2011.12854*.

Tiddi, I.; and Schlobach, S. 2022. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.*, 302: 103627.

Tran, Q. T.; Chan, C. Y.; and Parthasarathy, S. 2014. Query reverse engineering. *VLDB J.*, 23(5): 721–746.

Trivela, D.; Stoilos, G.; Chortaras, A.; and Stamou, G. 2020. Resolution-based rewriting for Horn-*SHIQ* ontologies. *Knowl. Inf. Syst.*, 62(1): 107–143.

Wu, W.; Su, Y.; Chen, X.; Zhao, S.; King, I.; Lyu, M. R.; and Tai, Y.-W. 2020. Towards global explanations of convolutional neural networks with concept attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8652–8661.

Xie, S.; Girshick, R. B.; Dollár, P.; Tu, Z.; and He, K. 2016. Aggregated Residual Transformations for Deep Neural Networks. *CoRR*, abs/1611.05431.

Yang, H.; Rudin, C.; and Seltzer, M. I. 2017. Scalable Bayesian Rule Lists. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, 3921–3930. PMLR.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. *CoRR*, abs/1605.07146.