An ASP approach for reasoning on neural networks under a finitely many-valued semantics for weighted conditional knowledge bases (Extended Abstract)

Laura Giordano , Daniele Theseider Dupré
DISIT - Università del Piemonte Orientale, Italy
laura.giordano@uniupo.it, dtd@uniupo.it

This extended abstract reports about the work in (Giordano and Theseider Dupré 2022), concerning an ASP approach for reasoning in a "concept-wise" multi-preferential semantics for weighted conditional knowledge bases, in the multi-valued case, and its use in the verification of some multilayer networks. New results and ASP encodings, which take advantage of weak constraints, have been investigated in (Alviano, Giordano, and Theseider Dupré 2023).

The work stems from the area of conditional and preferential reasoning. Preferential approaches to common sense reasoning (Delgrande 1987; Makinson 1988; Kraus, Lehmann, and Magidor 1990; Pearl 1990; Lehmann and Magidor 1992; Benferhat et al. 1993; Booth and Paris 1998; Kern-Isberner 2001) have been recently extended to Description Logics (DLs), to deal with inheritance with exceptions in ontologies, by allowing non-strict inclusions, called *defeasible* or *typicality* inclusions, of the form $\mathbf{T}(C) \sqsubseteq D$ (meaning "the typical C's are D's" or "normally C's are D's"). Different preferential semantics (Britz, Heidema, and Meyer 2008; Giordano et al. 2009; Giordano et al. 2015; Britz et al. 2021) and closure constructions have been proposed starting from Casini and Straccia's work (2010).

In recent work, a concept-wise multi-preferential semantics has been proposed as a semantics of ranked knowledge bases (KBs) in a lightweight description logic (Giordano and Theseider Dupré 2020), in which defeasible or typicality inclusions are given a rank, a natural number representing their strength. The idea underlying the multi-preferential semantics is that different preferences should be associated to different concepts and, for instance, for two individuals x and y, and two concepts, *Swimmer* and *Student*, x might be more typical than y as a swimmer ($x <_{Swimmer} y$) but less typical than y as a student ($y <_{Student} x$).

This semantics has been shown to have some desirable properties from the knowledge representation point of view and has also been extended to the fuzzy case (Giordano and Theseider Dupré 2021). In both the two-valued and fuzzy case, it has been exploited to provide a preferential interpretation to Multilayer Perceptrons (MLPs) (Haykin 1999), an approach previously considered (Giordano, Gliozzi, and Theseider Dupré 2022) for self-organising maps (SOMs) (Kohonen, Schroeder, and Huang 2001). Considering as domain a set of input stimuli presented to the network, one can build a semantic interpretation describing the input-output behavior of the network as a multi-preferential interpretation, where preferences are associated to concepts, which has suggested a model checking approach for post-hoc verification of both SOMs and MLPs. In particular, the model checking approach has been exploited in the verification of typicality properties of a multilayer networks, trained to recognize emotions from input features, based on a Datalog encoding of the model checking problem in the finite-valued case (Bartoli et al. 2022).

For MLPs, based on a fuzzy multi-preferential semantics for weighted KBs, a deep neural network can actually be regarded as a weighted conditional knowledge base (Giordano and Theseider Dupré 2021). This rises the issue of defining proof methods for reasoning with weighted conditional knowledge bases.

Weighted conditional $\mathcal{ALC}$ knowledge bases with typicality have been considered through some different semantic constructions. For reasoning with the concept-wise multi-preferential entailment under the so called φ-*coherent semantics*, the finite many-valued case has been considered, a case well studied for DLs (García-Cerdaña, Armengol, and Esteva 2010; Bobillo and Straccia 2011; Borgwardt and Peñaloza 2013).

An Answer Set Programming (ASP) approach has been proposed for the boolean fragment of $\mathcal{ALC}$, which neither contains roles, nor universal and existential restrictions. The problem of deciding φ-coherent entailment from a weighted knowledge base (in a finitely-valued description logic, with Gödel or with Łukasiewicz combination functions) is reformulated as the problem of computing preferred answer sets of an ASP program. The problem of verifying φ-coherent entailment of a typicality inclusion $\mathbf{T}(C) \sqsubseteq D \geq \alpha$ from a weighted knowledge base K (a subsumption problem), can be reformulated as a problem of generating answer sets representing φ-coherent models of the KB, and then selecting preferred answer sets, where a distinguished domain element aux_C is intended to represent a typical C-element. For the selection of preferred answer sets, those maximizing the degree of membership of aux_C in concept C, *asprin* (Brewka et al. 2015) is used. It is then verified that inclusion $C \sqsubseteq D \geq \alpha$ holds in all the preferred answer sets. Our proof method is sound and complete for the computation of φ-coherent entailment in the fragment considered, and provides a Π_2^p complexity upper-bound for φ-coherent entailment over a

finite set of values.

In the paper, as a proof of concept, we have experimented our approach over some weighted KBs corresponding to some of the trained multilayer feedforward networks considered by Thrun et al. (1991), exploiting ASP to verify properties of the network expressed as typicality properties in the finite many-valued case. This is a step towards explainability of the black-box, in view of a trustworthy, reliable and explainable AI (Adadi and Berrada 2018; Guidotti et al. 2019), and of an integrated use of symbolic reasoning and neural network models.

The paper does not provide the exact complexity of the problem and only describes a proof-of-concept implementation. An $P^{NP[log]}$-completeness result for φ-coherent entailment in the many-valued case and new ASP encodings have been investigated by Alviano, Giordano and Theseider Dupré (2023), taking advantage of weak constraints, possibly without the need for weights. Such encodings allow to deal with weighted knowledge bases with larger search spaces.

References

Adadi, A., and Berrada, M. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6:52138–52160.

Alviano, M.; Giordano, L.; and Theseider Dupré, D. 2023. Complexity and scalability of defeasible reasoning in many-valued weighted knowledge bases. *CoRR* abs/2303.04534. Accepted as Tech. Commun. at ICLP 2023.

Bartoli, F.; Botta, M.; Esposito, R.; Giordano, L.; and Theseider Dupré, D. 2022. An ASP approach for reasoning about the conditional properties of neural networks: an experiment in the recognition of basic emotions. In *Datalog 2.0 2022*, volume 3203 of *CEUR Workshop Proc.*, 54–67. CEUR-WS.org.

Benferhat, S.; Cayrol, C.; Dubois, D.; Lang, J.; and Prade, H. 1993. Inconsistency management and prioritized syntax-based entailment. In *Proc. IJCAI'93, Chambéry,*, 640–647.

Bobillo, F., and Straccia, U. 2011. Reasoning with the finitely many-valued Łukasiewicz fuzzy Description Logic SROIQ. *Inf. Sci.* 181(4):758–778.

Booth, R., and Paris, J. B. 1998. A note on the rational closure of knowledge bases with both positive and negative knowledge. *J. Log., Lang. Inf.* 7(2):165–190.

Borgwardt, S., and Peñaloza, R. 2013. The complexity of lattice-based fuzzy description logics. *J. Data Semant.* 2(1):1–19.

Brewka, G.; Delgrande, J. P.; Romero, J.; and Schaub, T. 2015. asprin: Customizing answer set preferences without a headache. In *Proc. AAAI 2015*, 1467–1474.

Britz, K.; Casini, G.; Meyer, T.; Moodley, K.; Sattler, U.; and Varzinczak, I. 2021. Principles of KLM-style defeasible description logics. *ACM Trans. Comput. Log.* 22(1):1:1–1:46.

Britz, K.; Heidema, J.; and Meyer, T. 2008. Semantic preferential subsumption. In Brewka, G., and Lang, J., eds., *KR 2008*, 476–484. Sidney, Australia: AAAI Press.

Casini, G., and Straccia, U. 2010. Rational Closure for Defeasible Description Logics. In Janhunen, T., and Niemelä, I., eds., *JELIA 2010*, volume 6341 of *LNCS*, 77–90. Helsinki: Springer.

Delgrande, J. 1987. A first-order conditional logic for prototypical properties. *Artif. Intell.* 33(1):105–130.

García-Cerdaña, A.; Armengol, E.; and Esteva, F. 2010. Fuzzy description logics and t-norm based fuzzy logics. *Int. J. Approx. Reason.* 51(6):632–655.

Giordano, L., and Theseider Dupré, D. 2020. An ASP approach for reasoning in a concept-aware multipreferential lightweight DL. *Theory Pract. Log. Program.* 10(5):751–766.

Giordano, L., and Theseider Dupré, D. 2021. Weighted defeasible knowledge bases and a multipreference semantics for a deep neural network model. In *Proc. JELIA 2021, May 17-20*, volume 12678 of *LNCS*, 225–242. Springer.

Giordano, L., and Theseider Dupré, D. 2022. An ASP approach for reasoning on neural networks under a finitely many-valued semantics for weighted conditional knowledge bases. *Theory Pract. Log. Program.* 22(4):589–605.

Giordano, L.; Gliozzi, V.; Olivetti, N.; and Pozzato, G. L. 2009. ALC+T: a preferential extension of Description Logics. *Fundamenta Informaticae* 96:1–32.

Giordano, L.; Gliozzi, V.; Olivetti, N.; and Pozzato, G. L. 2015. Semantic characterization of rational closure: From propositional logic to description logics. *Art. Int.* 226:1–33.

Giordano, L.; Gliozzi, V.; and Theseider Dupré, D. 2022. A conditional, a fuzzy and a probabilistic interpretation of self-organizing maps. *J. Log. Comput.* 32(2):178–205.

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.

Haykin, S. 1999. *Neural Networks - A Comprehensive Foundation*. Pearson.

Kern-Isberner, G. 2001. *Conditionals in Nonmonotonic Reasoning and Belief Revision - Considering Conditionals as Agents*, volume 2087 of *LNCS*. Springer.

Kohonen, T.; Schroeder, M.; and Huang, T., eds. 2001. *Self-Organizing Maps, Third Edition*. Springer Series in Information Sciences. Springer.

Kraus, S.; Lehmann, D.; and Magidor, M. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artif. Intell.* 44(1-2):167–207.

Lehmann, D., and Magidor, M. 1992. What does a conditional knowledge base entail? *Artif. Intell.* 55(1):1–60.

Makinson, D. 1988. General theory of cumulative inference. In *Non-Monotonic Reasoning, 2nd International Workshop, Grassau, FRG, June 13-15, 1988, Proceedings*, 1–18.

Pearl, J. 1990. System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. In *TARK'90, Pacific Grove, CA, USA*, 121–135.

Thrun, S. et al. 1991. A Performance Comparison of Different Learning Algorithms. Technical Report CMU-CS-91-197, Carnegie Mellon University.