Opening the Black-box: Deep Neural Networks as Weighted Conditional Knowledge Bases (Extended Abstract) ⋆

Laura Giordano and Daniele Theseider Dupré

DISIT - Università del Piemonte Orientale, Alessandria, Italy
laura.giordano@uniupo.it, dtd@uniupo.it

In this abstract we report the results of the paper *"Weighted defeasible knowledge bases and a multipreference semantics for a deep neural network model"* in *Proc. JELIA 2021* [15], which investigates the relationships between a multipreferential semantics for defeasible reasoning in knowledge representation and a deep neural network model. Weighted knowledge bases for description logics are considered under a "concept-wise" multipreference semantics. The semantics is further extended to fuzzy interpretations and exploited to provide a preferential interpretation of Multilayer Perceptrons.

Preferential approaches have been used to provide axiomatic foundations of non-monotonic and common sense reasoning [11, 31, 33, 26, 28, 32, 3, 22]. They have been extended to description logics (DLs), to deal with inheritance with exceptions in ontologies, by allowing for non-strict forms of inclusions, called *typicality or defeasible inclusions*, with different preferential semantics [19, 7] and closure constructions [9, 8, 20, 5, 34, 6, 16]. The paper exploits a concept-wise multipreference semantics as a semantics for weighted knowledge bases, i.e. knowledge bases in which 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") are given a positive or negative weight. For instance,

A multipreference semantics, taking into account preferences with respect to different concepts, was first introduced by the authors as a semantics for ranked DL knowledge bases [13]. For weighted knowledge bases, a different semantic closure construction is developed, still in the spirit of other semantic constructions in the literature, and is further extended to the fuzzy case.

A preference relation $<_{C_i}$ on the domain Δ of a DL interpretation can be associated to each concept C_i to represent the relative typicality of domain individuals with respect to C_i. Preference relations with respect to different concepts do not need to agree, as a domain element x may be more typical than y as a horse but less typical as a zebra. The plausibility/implausibility of properties for a concept is represented by their (positive or negative) weight. For instance, a weighted TBox (called $\mathcal{T}_{Employee}$) associated to concept $Employee$ might contain the following weighted defeasible inclusions:
(d_1) $\mathbf{T}(Employee) \sqsubseteq Young$, - 50 (d_3) $\mathbf{T}(Employee) \sqsubseteq \exists has\_classes.\top$, -70
(d_2) $\mathbf{T}(Employee) \sqsubseteq \exists has\_boss.Employee$, 100;
meaning is that, while an employee normally has a boss, he is not likely to be young or have classes. Furthermore, between the two defeasible inclusions (d_1) and (d_3), the second one is considered to be less plausible than the first one.

Multipreference interpretations are defined by adding to standard DL interpretations, which are pairs $\langle \Delta, \cdot^I \rangle$, where Δ is a domain, and $\cdot^I$ an interpretation function, the preference relations $<_{C_1}, \ldots, <_{C_n}$ associated with a set of distinguished concepts $C_1, \ldots, C_n$. The definition of a global preference relation $<$ from the $<_{C_i}$'s, leads to the definition of a notion of *concept-wise multipreference interpretation (cwm-interpretation)*, where concept $\mathbf{T}(C)$ is interpreted as the set of all $<$-minimal C elements. A simple notion of global preference $<$ exploits Pareto combination of the preference relations $<_{C_i}$, but a more sophisticated notion of preference combination has been considered in [14], by taking into account the specificity relation among concepts. It has been proven [14] that global preference in a cwm-interpretation determines a KLM-style preferential interpretation, and cwm-entailment satisfies the KLM postulates of a preferential consequence relation [26].

While in previous work [17, 18], the concept-wise multipreference semantics is used to provide a preferential interpretation of Self-Organising Maps [24], psychologically and biologically plausible neural network models, [15] investigates its relationships with Multilayer Perceptrons (MLPs), a deep neural network model. A deep network is considered after the training phase, when the synaptic weights have been learned, to show that it can be associated a preferential DL interpretation with multiple preferences, as well as a semantics based on fuzzy DL interpretations and another one combining fuzzy interpretations with multiple preferences. The three semantics allow the input-output behavior of the network to be captured by interpretations built over a set of input stimuli through a simple construction, which exploits the activity level of neurons for the stimuli. Logical properties can be verified over such models by model checking.

The idea underlying fuzzy-multipreference interpretations [15] is to extend a fuzzy DL interpretation with a set of induced preferences. In a fuzzy DL interpretation I, the interpretation of a concept C_h is a mapping $C_h^I : \Delta \rightarrow [0, 1]$, associating to each $x \in \Delta$ the degree of membership of x in C_h. In MLPs, each unit h can be associated to a concept C_h and, for a given domain Δ of input stimuli, the activation value of unit h for a stimulus x, can be interpreted as the degree of membership of x in concept C_h. The fuzzy interpretation of concepts induces an ordering $<_{C_h}$ on the domain Δ, which can be regarded as the preference relation associated to concept C_h. This allows a notion of typicality to be defined in a fuzzy interpretation. Logical properties of the neural network (both typicality inclusions and fuzzy axioms) can then be verified over such interpretations by model checking. It has been proven that, also in the fuzzy case, the concept-wise multipreference semantics has interesting properties and satisfies most of the KLM properties of a preferential consequence relation [12].

The definition of the concept-wise preferences starting from a weighted conditional KB exploits a closure construction in the same spirit of the one considered by Lehmann [29] to define the lexicographic closure, but more similar to Kern-Isberner's c-interpretations [22, 23], in which the world ranks are generated as a sum of impacts of falsified conditionals. Here, the (positive or negative) weights of the satisfied defaults are summed, but in a concept-wise manner, so to determine the plausibility of a domain elements with respect to certain concepts, by considering the modular structure of the KB. To guarantee that such preferences are coherent with the fuzzy interpretation of concepts, a notions of *coherent (fuzzy) multipreference interpretation)* has been introduced.

To prove that the fuzzy multipreference interpretation, built from a network $\mathcal{N}$ for a given set of input stimuli (a domain Δ), is a model of $\mathcal{N}$ in a logical sense, the multilayer network is mapped to a conditional knowledge base $K^{\mathcal{N}}$ containing, for each neuron k, a set of weighted defeasible inclusions. If C_k is the concept name associated to unit k and $C_{j_1}, \ldots, C_{j_m}$ are the concept names associated to units $j_1, \ldots, j_m$, whose output signals are the input signals for unit k, with synaptic weights $w_{k,j_1}, \ldots, w_{k,j_m}$, then unit k cab be associated a set $\mathcal{T}_{C_k}$ of weighted typicality inclusions: $\mathbf{T}(C_k) \sqsubseteq C_{j_1}$ with w_{k,j_1}, ..., $\mathbf{T}(C_k) \sqsubseteq C_{j_m}$ with w_{k,j_m}. The fuzzy multipreference interpretation built from a network $\mathcal{N}$ over a domain Δ can be proven to be a model of the knowledge base $K^{\mathcal{N}}$ under the some conditions on the activation functions.

The correspondence between neural network models and fuzzy systems has been first investigated by Kosko in his seminal work [25]. As a difference, have adopted the usual way of viewing concepts in fuzzy DLs [35, 30, 4], and we have used fuzzy concepts within a concept-wise multipreference semantics, based on a semantic closure construction. The first combination of fuzzy logic with the preferential semantics of conditional KBs has been studied by Casini and Straccia [10], who have developed a rational closure construction for propositional Gödel logic.

The possibility of exploiting the concept-wise multipreference semantics to provide a semantic interpretation of a neural network model has been first explored for Self-Organising Maps (SOMs), psychologically and biologically plausible neural network models [24]. A multi-preferential semantics can be used to provide a logical model of the SOM behavior after training [17, 18], based on the idea of associating different preference relations to categories. The model can be used to learn or validate conditional knowledge from the empirical data used for training and generalization, by model checking of logical properties. A similar approach has been adopted in [15] for the MLPs. Due to the diversity of the two neural models we expect that the approach might be extended to other neural network models and learning approaches.

A logical interpretation of a neural network can be useful from the point of view of explainability, in view of a trustworthy, reliable and explainable AI [1, 21, 2]. For MLPs, the strong relationship between a multilayer network and a weighted KB opens to the possibility of adopting a conditional DLs as a basis for neuro-symbolic integration. While a neural network, once trained, is able and fast in classifying the new stimuli (that is, it is able to do instance checking), all other reasoning services such as satisfiability, entailment and model-checking are missing. These capabilities would be needed for dealing with tasks combining empirical and symbolic knowledge, e.g., proving whether the network satisfies some (strict or conditional) properties; learning the weights of a conditional KB from empirical data, and combine the defeasible inclusions extracted from a neural network with other defeasible or strict inclusions for inference. To make these tasks possible, the development of proof methods for such logics is a preliminary step. An open problem is whether the notion of fuzzy-multipreference entailment is decidable, for which DLs fragments and under which choice of fuzzy logic combination functions. Another issue is whether the mapping of multilayer networks to weighted conditional knowledge bases can be extended to more complex neural network models, such as Graph neural networks [27], or whether different logical formalisms and semantics would be needed.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. 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., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)
3. Benferhat, S., Cayrol, C., Dubois, D., Lang, J., Prade, H.: Inconsistency management and prioritized syntax-based entailment. In: Proc. IJCAI'93, Chambéry, France, August 28 - September 3, 1993. pp. 640–647. Morgan Kaufmann
4. Bobillo, F., Straccia, U.: The fuzzy ontology reasoner fuzzydl. Knowl. Based Syst. **95**, 12–34 (2016)
5. Bonatti, P.A., Sauro, L.: On the logical properties of the nonmonotonic description logic DL^{N}. Artif. Intell. **248**, 85–111 (2017)
6. Britz, K., Casini, G., Meyer, T., Moodley, K., Sattler, U., Varzinczak, I.: Principles of KLM-style defeasible description logics. ACM Trans. Comput. Log. **22**(1), 1:1–1:46 (2021)
7. Britz, K., Heidema, J., Meyer, T.: Semantic preferential subsumption. In: Brewka, G., Lang, J. (eds.) KR 2008. pp. 476–484. AAAI Press, Sidney, Australia (September 2008)
8. Casini, G., Meyer, T., Varzinczak, I.J., , Moodley, K.: Nonmonotonic Reasoning in Description Logics: Rational Closure for the ABox. In: 26th International Workshop on Description Logics (DL 2013). CEUR Workshop Proceedings, vol. 1014, pp. 600–615 (2013)
9. Casini, G., Straccia, U.: Rational Closure for Defeasible Description Logics. In: Janhunen, T., Niemelä, I. (eds.) JELIA 2010. LNCS, vol. 6341, pp. 77–90. Springer, Helsinki (Sept 2010)
10. Casini, G., Straccia, U.: Towards rational closure for fuzzy logic: The case of propositional Gödel logic. In: Proceedings of the 19th International Conferences on Logic for Programming, Artificial Intelligence and Reasoning (LPAR-13). Lecture Notes in Computer Science, Advanced Research in Computing and Software Science, vol. 8312, pp. 213–227. Springer Verlag (2013)
11. Delgrande, J.: A first-order conditional logic for prototypical properties. Artificial Intelligence **33**(1), 105–130 (1987)
12. Giordano, L.: On the KLM properties of a fuzzy DL with Typicality. In: 16th European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU 2021. Springer (2021), to appear.
13. Giordano, L., Dupré, D.T.: An ASP approach for reasoning in a concept-aware multipreferential lightweight DL. Theory Pract. Log. Program. **20**(5), 751–766 (2020)
14. Giordano, L., Dupré, D.T.: An ASP approach for reasoning in a concept-aware multipreferential lightweight DL. Theory and Practice of Logic programming, TPLP **10(5)**, 751–766 (2020)
15. Giordano, L., Dupré, D.T.: Weighted defeasible knowledge bases and a multipreference semantics for a deep neural network model. In: Faber, W., Friedrich, G., Gebser, M., Morak, M. (eds.) Proc17th European Conf. on Logics in AI, JELIA 2021, May 17-20. LNCS, vol. 12678, pp. 225–242. Springer (2021), https://doi.org/10.1007/978-3-030-75775-5_16
16. Giordano, L., Gliozzi, V.: A reconstruction of multipreference closure. Artif. Intell. **290** (2021)
17. Giordano, L., Gliozzi, V., Dupré, D.T.: On a plausible concept-wise multipreference semantics and its relations with self-organising maps. In: Calimeri, F., Perri, S., Zumpano, E. (eds.) CILC 2020, Rende, Italy, October 13-15, 2020. CEUR, vol. 2710, pp. 127–140 (2020)
18. Giordano, L., Gliozzi, V., Dupré, D.T.: A conditional, a fuzzy and a probabilistic interpretation of self-organising maps. CoRR **abs/2103.06854** (2021), https://arxiv.org/abs/2103.06854

19. Giordano, L., Gliozzi, V., Olivetti, N., Pozzato, G.L.: Preferential Description Logics. In: LPAR 2007. LNAI, vol. 4790, pp. 257–272. Springer, Yerevan, Armenia (October 2007)
20. Giordano, L., Gliozzi, V., Olivetti, N., Pozzato, G.L.: Semantic characterization of rational closure: From propositional logic to description logics. Artif. Intell. **226**, 1–33 (2015)
21. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (2019)
22. Kern-Isberner, G.: Conditionals in Nonmonotonic Reasoning and Belief Revision - Considering Conditionals as Agents, LNCS, vol. 2087. Springer (2001)
23. Kern-Isberner, G., Eichhorn, C.: Structural inference from conditional knowledge bases. Stud Logica **102**(4), 751–769 (2014)
24. Kohonen, T., Schroeder, M., Huang, T. (eds.): Self-Organizing Maps, Third Edition. Springer Series in Information Sciences, Springer (2001)
25. Kosko, B.: Neural networks and fuzzy systems: a dynamical systems approach to machine intelligence. Prentice Hall (1992)
26. Kraus, S., Lehmann, D., Magidor, M.: Nonmonotonic reasoning, preferential models and cumulative logics. Artificial Intelligence **44**(1-2), 167–207 (1990)
27. Lamb, L.C., d'Avila Garcez, A.S., Gori, M., Prates, M.O.R., Avelar, P.H.C., Vardi, M.Y.: Graph neural networks meet neural-symbolic computing: A survey and perspective. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. pp. 4877–4884. ijcai.org (2020)
28. Lehmann, D., Magidor, M.: What does a conditional knowledge base entail? Artificial Intelligence **55**(1), 1–60 (1992). https://doi.org/http://dx.doi.org/10.1016/0004-3702(92)90041-U
29. Lehmann, D.J.: Another perspective on default reasoning. Ann. Math. Artif. Intell. **15**(1), 61–82 (1995)
30. Lukasiewicz, T., Straccia, U.: Managing uncertainty and vagueness in description logics for the semantic web. J. Web Semant. **6**(4), 291–308 (2008)
31. Makinson, D.: General theory of cumulative inference. In: Non-Monotonic Reasoning, 2nd International Workshop, Grassau, FRG, June 13-15, 1988, Proceedings. pp. 1–18 (1988)
32. Pearl, J.: System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. In: TARK'90, Pacific Grove, CA, USA, 1990. pp. 121–135. Morgan Kaufmann
33. Pearl, J.: Probabilistic Reasoning in Intelligent Systems Networks of Plausible Inference. Morgan Kaufmann (1988)
34. Pensel, M., Turhan, A.: Reasoning in the defeasible description logic $EL_\perp$ - computing standard inferences under rational and relevant semantics. Int. J. Approx. Reasoning **103**, 28–70 (2018)
35. Straccia, U.: Towards a fuzzy description logic for the semantic web (preliminary report). In: The Semantic Web: Research and Applications, Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29 - June 1, 2005, Proceedings. Lecture Notes in Computer Science, vol. 3532, pp. 167–181. Springer (2005)