From Common Sense Reasoning to Neural Network Models: a Multi-preferential Approach to Explainability

Laura Giordano [1], **Valentina Gliozzi** [2], **Daniele Theseider Dupré** [1]

[1] Dipartimento di Scienze e Innovazione Tecnologica e Centro di Ricerca Interdipartimentale sull'Intelligenza Artificiale (AI@UPO) - Università del Piemonte Orientale

[2] Dipartimento di Informatica e Centro di Logica, Linguaggio e Cognizione (LLC) - Università di Torino

laura.giordano@uniupo.it, valentina.gliozzi@unito.it, dtd@uniupo.it

Abstract

This short paper reports about a line of research at Università del Piemonte Orientale and Università di Torino, concerning a semantic interpretation of neural networks, based on a "concept-wise" multi-preferential semantics for weighted conditionals and its use to provide a logical interpretation to some neural network models. The approach has been applied to Self-Organising Maps (SOMs) and Multilayer Perceptrons (MLPs). For MLPs, a deep network can be regarded as a conditional knowledge base, in which the synaptic connections correspond to weighted conditionals.

1 Introduction

Preferential approaches to common sense reasoning (e.g., [Kraus *et al.*, 1990]) have their roots in conditional logics [Lewis, 1973; Nute, 1980], and have been recently extended to Description Logics (DLs), to deal with inheritance with exceptions in ontologies, by allowing non-strict form of inclusions, called *defeasible* or *typicality* inclusions.

Different preferential semantics and closure constructions have been proposed for such defeasible DLs. In this short paper, we report about a concept-wise multipreference semantics [Giordano and Theseider Dupré, 2020], introduced as a semantics of ranked knowledge bases in a lightweight DL to account for preferences with respect to different concepts, and proposed as a semantics for some neural network models.

We deal with both an unsupervised model, Self-organising maps (SOMs) [Kohonen *et al.*, 2001], which is considered as a psychologically and biologically plausible neural network model, and a supervised one, Multilayer Perceptrons (MLPs) [Haykin, 1999]. Learning algorithms in the two cases are quite different but our aim was to capture in a semantic interpretation the behavior of the network after training and not to deal with the learning process.

In both cases, considering the domain of all input stimuli presented to the network during training (or generalization), a semantic interpretation describing the *input-output behavior* of the network can be provided as a multipreference interpretation, where preferences are associated to concepts. For SOMs, the learned categories $C_1, \ldots, C_n$ are regarded as concepts so that a preference relation over the domain of input stimuli is associated with each category [Giordano *et al.*, 2020a; Giordano *et al.*, 2021]. For MLPs, each unit of intererst in the deep network (including hidden neurons) can be associated with a concept and with a preference relation on the domain [Giordano and Theseider Dupré, 2021b].

The idea is that, given two input stimuli x and y, and two categories/concepts, e.g., $Horse$ and $Zebra$, the neural model can, for example, assign to x a degree of membership in $Horse$ which is higher than the degree of membership of y, so that x can be regarded as more typical than y as a horse ($x <_{Horse} y$), while x could be less typical than y as a zebra ($y <_{Zebra} x$). A preferential interpretation can be built over the domain of input stimuli, and used for checking properties such as: are the instances of category C_1 also instances of C_2? Are typical instances of C_1 also instances of C_2? This verification can be done by *model-checking* given the multipreference interpretation.

For MLPs, the relationship between the logic of commonsense reasoning and deep neural networks is even stronger, as a deep neural network can itself be regarded as a conditional knowledge base, i.e., as a set weighted conditionals. This has been achieved by developing a concept-wise fuzzy multipreference semantics for a DLs with weighted defeasible inclusions.

The strong relationship between neural networks and conditional logics of commonsense reasoning raises several issues from the standpoint of knowledge representation, from the standpoint of neuro-symbolic integration, as well as from the standpoint of explainable AI [Adadi and Berrada, 2018; Guidotti *et al.*, 2019; Arrieta *et al.*, 2020]. We will hint at some of these issues in the following.

2 The concept-wise multipreference semantics

The concept-wise multipreference semantics (cw^m-semantics) has been introduced as a semantics for ranked $\mathcal{EL}$ knowledge bases [Giordano and Theseider Dupré, 2020], and later extended to weighted knowledge bases [Giordano and Theseider Dupré, 2021b]. In both cases the knowledge base contains *strict* (i.e., standard) inclusions and defeasible or typicality inclusions $\mathbf{T}(C) \sqsubseteq D$ (meaning "the typical Cs are Ds" or "normally Cs are Ds") with a rank (resp. a weight). They correspond to KLM conditionals $C \mathrel{|\!\sim} D$ [Kraus *et al.*, 1990]. Ranks (weights) of defeasible inclusions represent their strength (plausibility/implausibility). The

preferential semantics of ranked and weighted knowledge bases are defined in terms of concept-wise multipreference interpretations, based on different constructions.

Concept-wise multipreference interpretations (cw^m-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 $C_1, \ldots, C_n$, representing the relative typicality of domain individuals with respect to these concepts. Each preference relation $<_{C_i}$ is a modular and well-founded strict partial order on Δ. Preferences with respect to different concepts do not need to agree, as we have seen. A global preference relation $<$ can be defined from the $<_{C_i}$'s, and 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 $<_{C_i}$ (a more sophisticated global preference, taking into account specificity, has also been considered). It has been proven [Giordano and Theseider Dupré, 2020] that global preference in a cw^m-interpretation determines a KLM-style preferential interpretation, and cw^m-entailment satisfies the KLM postulates of a preferential consequence relation [Kraus *et al.*, 1990].

3 A preferential interpretation of Self-Organising Maps

Once a SOM has learned to categorize, the result of the categorization can be seen as a concept-wise multipreference interpretation over a domain of input stimuli, in which a preference relation is associated with each concept (learned category). The combination of preferences into a global one (following the approach described above) defines a KLM-style preferential model of the SOM. More precisely, once the SOM has learned to categorize, to assess category generalization, Gliozzi and Plunkett [2019] define the map's disposition to consider a new stimulus y as a member of a known category C as a function of the *distance* of y from the *map's representation* of C. The relative distance $rd(x, C_i)$ of a stimulus x from a category C_i can be used to build a binary preference relation $<_{C_i}$ among the stimuli in Δ with respect to category C_i [Giordano *et al.*, 2020a; Giordano *et al.*, 2020b], by letting $x <_{C_i} y$ if and only if $rd(x, C_i) > rd(y, C_i)$ (x is more typical than y with respect to category C_i if its relative distance from category C_i is lower than the relative distance of y).

This preferential model can be exploited to learn or validate conditional knowledge from empirical data, by verifying conditional formulas over the preferential interpretation constructed from the SOM. Both a two-valued and a fuzzy semantics have been considered [Giordano *et al.*, 2021]. In both cases, model checking can be used for the verification of inclusions (either defeasible inclusions or fuzzy inclusion axioms) over the respective models of the SOM (for instance, do the most typical penguins belong to the category Bird with at least a degree of membership 0.8?). A probabilistic account can also be given based on Zadeh's probability of fuzzy events [Zadeh, 1968].

4 A preferential interpretation of Multilayer Perceptrons

The input-output behaviour of MLPs can be captured in a similar way as for SOMs by constructing a preferential interpretation over the domain Δ of the input stimuli considered during training (or generalization) [Giordano and Theseider Dupré, 2021b]. Each neuron k of interest can be associated to a concept C_k and, for each distinguished concept C_j, a preference relation $<_{C_j}$ is defined over the domain Δ based on the activity values, $y_j(v)$, of neuron j for each input $v \in \Delta$. In a similar way, a fuzzy interpretation of the network can be constructed over the domain Δ, as well as a fuzzy-multipreference semantics.

All the three semantics allow the input-output behavior of the network to be captured by interpretations built over a set of input stimuli through simple constructions, which exploits the activity level of neurons for the stimuli. For the fuzzy-multipreference interpretations, the idea is to enrich a fuzzy DL interpretation I with a set of induced preferences: the interpretation of a concept C_h is a mapping $C_h^I : \Delta \to [0, 1]$, associating to each $x \in \Delta$ the degree of membership of x in C_h. The activation value of unit h for a stimulus x in the network (assumed to be in the interval $[0, 1]$) is taken as the degree of membership of x in concept C_h. The fuzzy interpretation also induces an ordering $<_{C_h}$ on Δ. This allows a notion of typicality to be defined in a fuzzy interpretation. Let us call $\mathcal{M}_{\mathcal{N}}^{f,\Delta}$ the fuzzy multipreference interpretation built from network $\mathcal{N}$ over a domain Δ.

As for SOMs, logical properties of the network (typicality properties and fuzzy axioms) can then be verified by model checking over such an interpretation. Evaluating properties involving hidden units might be of interest (see, for instance, the discussion in [Giordano and Theseider Dupré, 2021b] of the well known Hinton's family example [1986]). If the properties of interest involve some specific units, only the concepts associated to those units may be considered in the language to build the interpretation.

All the three kinds of interpretations considered above for MLPs describe the input-output behavior of the network. However, the fuzzy multipreference interpretation $\mathcal{M}_{\mathcal{N}}^{f,\Delta}$ described above can be also proven to be a model of the neural network $\mathcal{N}$ in a logical sense, by mapping the multilayer network into a weighted conditional knowledge base.

4.1 Weighted $\mathcal{ALC}$ knowledge bases

We introduce the definition of weighted conditional knowledge bases through an example, and give some hints about the two-valued and fuzzy multipreference semantics.

A weighted $\mathcal{ALC}$ knowledge base, other than sets of standard inclusion axioms (Tbox $\mathcal{T}$) and assertions (Abox $\mathcal{A}$), contains, for a set $\mathcal{C} = \{C_1, \ldots, C_k\}$ of distinguished $\mathcal{ALC}$ concepts, a set of weighted typicality inclusions of the form $\mathbf{T}(C_i) \sqsubseteq D$, with a positive or negative weight (a real number). In the fuzzy case, $\mathcal{T}$ and $\mathcal{A}$ contain fuzzy axioms.

As an example, a knowledge base with $\mathcal{T}$ containing the inclusions $Penguin \sqsubseteq Bird$ and $Black \sqcap Grey \sqsubseteq \bot$ may also include the following weighted defeasible inclusions:

(d_1) $\mathbf{T}(Bird) \sqsubseteq Fly$, +20

(d_2) $\mathbf{T}(Bird) \sqsubseteq \exists has\_Wings.\top$, +50
(d_3) $\mathbf{T}(Bird) \sqsubseteq \exists has\_Feathers.\top$, +50;
(d_4) $\mathbf{T}(Penguin) \sqsubseteq Fly$, - 70
(d_5) $\mathbf{T}(Penguin) \sqsubseteq Black$, +50;
(d_6) $\mathbf{T}(Penguin) \sqsubseteq Grey$, +10;

The meaning is that a bird normally has wings, has feathers and flies, but having wings and feathers is more plausible than flying, although flying is regarded as being plausible. For a penguin, flying is not plausible (d_4 has a negative weight), and being black is more plausible than being grey.

A two-valued semantics for weighted $\mathcal{ALC}$ knowledge bases can be defined with a semantic closure construction in the spirit of Lehmann's lexicographic closure [Lehmann, 1995], but more similar to Kern-Isberner's semantics of c-representations [Kern-Isberner, 2001; Kern-Isberner and Eichhorn, 2014], 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. For a domain element x in Δ, and a distinguished concept C_i, the weight $W_i(x)$ of x wrt C_i is defined as the sum of the weights w_h^i of the typicality inclusions $\mathbf{T}(C_i) \sqsubseteq D_{i,h}$ verified by x (and is $-\infty$ when x is not an instance of C_i). From the weights $W_i(x)$ the *preference relation* $\leq_{C_i}$ can be defined by letting $x \leq_{C_i} y$ iff $W_i(x) \geq W_i(y)$. The higher the weight of x wrt C_i the higher its typicality relative to C_i. This closure construction defines preferences $<_{C_i}$ (strict modular partial orders) and allows for the definition of *concept-wise multipreference interpretations* as in Section 2.

In the fuzzy case, the fuzzy logic combination functions are used for complex concepts to compute the $W_i(x)$'s and to determine the associated preference relations. To guarantee that the preferences determined from the knowledge base are coherent with the fuzzy interpretation of concepts, a notion of *coherent (fuzzy) multipreference interpretation* is also introduced.

4.2 MLPs as conditional knowledge bases

A multilayer network $\mathcal{N}$ can be mapped to a weighted conditional knowledge base $K^{\mathcal{N}}$ as follows. If the output signal of unit j is input to unit k, with synaptic weight w, the typicality inclusion $\mathbf{T}(C_k) \sqsubseteq C_j$, with weight w, is included in the knowledge base, where concept names C_k and C_j are associated with the respective units.

Let us consider the fuzzy multipreference interpretation $\mathcal{M}_{\mathcal{N}}^{f,\Delta}$ built from $\mathcal{N}$ over a domain Δ of input stimuli, as described earlier, and assume that, all units are considered, i.e., a concept C_k is introduced in the language for each unit k. It has been proven [Giordano and Theseider Dupré, 2021b] that the interpretation $\mathcal{M}_{\mathcal{N}}^{f,\Delta}$ is a coherent fuzzy multipreference model of the knowledge base $K^{\mathcal{N}}$, under some condition on the activation functions in $\mathcal{N}$. The properties that are entailed from $K^{\mathcal{N}}$ are satisfied by $\mathcal{M}_{\mathcal{N}}^{f,\Delta}$, for any choice of the input stimuli in the domain Δ.

5 Discussion and Conclusions

The multipreference semantics has been shown to satisfy the KLM properties in the two-valued case [Giordano and Theseider Dupré, 2020], and most of the KLM properties in the fuzzy case, depending on their reformulation and on the fuzzy combination functions considered [Giordano, 2021].

While a neural network, once trained, is able and fast in classifying the new stimuli (that is, it is able to do instance checking), other reasoning services such as satisfiability, entailment and model-checking are missing. Such reasoning tasks are useful for validating knowledge that has been learned, including proving whether the network satisfies some (strict or conditional) properties.

Much work has been devoted to the combination of neural networks and symbolic reasoning (e.g., the work by d'Avila Garcez et al. [2001; 2009; 2019] and Setzu et al. [2021]), as well as to the definition of new computational models [Lamb *et al.*, 2020; Serafini and d'Avila Garcez, 2016; Hohenecker and Lukasiewicz, 2020; Le-Phuoc *et al.*, 2021]. The work summarized in this paper also opens to the possibility of adopting conditional logics as a basis for neuro-symbolic integration, e.g., learning the weights of a conditional knowledge base from empirical data, and combining the defeasible inclusions extracted from a neural network with other defeasible or strict inclusions for inference.

To make these tasks possible, proof methods for such logics are needed. In the two-valued case multipreference entailment is decidable for weighted $\mathcal{EL}^{\bot}$ knowledge bases, and proof methods exploit Answer Set Programming (ASP) encodings of the concept-wise multipreference semantics [Giordano and Theseider Dupré, 2021a], using *asprin* [Brewka *et al.*, 2015] to achieve defeasible reasoning. In the fuzzy case, an open problem is whether the notion of fuzzy-multipreference entailment is decidable, and under which choice of fuzzy logic combination functions. Undecidability results for fuzzy description logics motivate the investigation of decidable approximations of fuzzy-multipreference entailment.

An important issue for future work concerns the experimental evaluation of the approach and its effectiveness to achieve explainability. Due to the diversity of the neural network models considered, we expect that the construction of a preferential interpretation can be applied to other network models and learning approaches.

References

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

[Arrieta *et al.*, 2020] A. Barredo Arrieta, N. Díaz Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115, 2020.

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

[d'Avila Garcez *et al.*, 2001] A. S. d'Avila Garcez, K. Broda, and D. M. Gabbay. Symbolic knowledge extraction from trained neural networks: A sound approach. *Artif. Intell.*, 125(1-2):155–207, 2001.

[d'Avila Garcez *et al.*, 2009] A. S. d'Avila Garcez, L. C. Lamb, and D. M. Gabbay. *Neural-Symbolic Cognitive Reasoning*. Cognitive Technologies. Springer, 2009.

[d'Avila Garcez *et al.*, 2019] A. S. d'Avila Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *FLAP*, 6(4):611–632, 2019.

[Giordano and Theseider Dupré, 2020] L. Giordano and D. Theseider Dupré. An ASP approach for reasoning in a concept-aware multipreferential lightweight DL. *Theory and Practice of Logic programming, TPLP*, 10(5):751–766, 2020.

[Giordano and Theseider Dupré, 2021a] L. Giordano and D. Theseider Dupré. Weighted conditional $\mathcal{EL}$ knowledge bases with integer weights: an ASP approach. In *ICLP 2021 Tech. Commun.*, volume 345 of *EPTCS*, pages 70–76, 2021.

[Giordano and Theseider Dupré, 2021b] L. Giordano and D. Theseider Dupré. Weighted defeasible knowledge bases and a multipreference semantics for a deep neural network model. In *JELIA 2021*, volume 12678 of *LNCS*, pages 225–242. Springer, 2021.

[Giordano *et al.*, 2020a] L. Giordano, V. Gliozzi, and D. Theseider Dupré. On a plausible concept-wise multipreference semantics and its relations with self-organising maps. In *CILC 2020*, volume 2710 of *CEUR*, pages 127–140, 2020.

[Giordano *et al.*, 2020b] L. Giordano, V. Gliozzi, and D. Theseider Dupré. Towards a conditional interpretation of self organising maps. In *Italian Workshop on Explainable Artificial Intelligence, XAI.it 2020*, volume 2742 of *CEUR*, pages 127–134, 2020.

[Giordano *et al.*, 2021] L. Giordano, V. Gliozzi, and D. Theseider Dupré. A conditional, a fuzzy and a probabilistic interpretation of self-organising maps. *CoRR*, abs/2103.06854, 2021. To appear in Journal of Logic and Computation.

[Giordano, 2021] L. Giordano. On the KLM properties of a fuzzy DL with Typicality. In *ECSQARU 2021*. Springer, 2021.

[Gliozzi and Plunkett, 2019] V. Gliozzi and K. Plunkett. Grounding bayesian accounts of numerosity and variability effects in a similarity-based framework: the case of self-organising maps. *Journal of Cognitive Psychology*, 31(5–6), 2019.

[Guidotti *et al.*, 2019] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2019.

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

[Hinton, 1986] G. Hinton. Learning distributed representation of concepts. In *Proceedings 8th Annual Conference of the Cognitive Science Society. Erlbaum, Hillsdale, NJ*, 1986.

[Hohenecker and Lukasiewicz, 2020] P. Hohenecker and T. Lukasiewicz. Ontology reasoning with deep neural networks. *J. Artif. Intell. Res.*, 68:503–540, 2020.

[Kern-Isberner and Eichhorn, 2014] G. Kern-Isberner and C. Eichhorn. Structural inference from conditional knowledge bases. *Stud Logica*, 102(4):751–769, 2014.

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

[Kohonen *et al.*, 2001] T. Kohonen, M.R. Schroeder, and T.S. Huang, editors. *Self-Organizing Maps, Third Edition*. Springer Series in Information Sciences. Springer, 2001.

[Kraus *et al.*, 1990] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44(1-2):167–207, 1990.

[Lamb *et al.*, 2020] L. C. Lamb, A. S. d'Avila Garcez, M. Gori, M. O. R. Prates, P. H. C. Avelar, and M. Y. Vardi. Graph neural networks meet neural-symbolic computing: A survey and perspective. In *IJCAI 2020*, pages 4877–4884, 2020.

[Le-Phuoc *et al.*, 2021] D. Le-Phuoc, T. Eiter, and A. Le-Tuan. A scalable reasoning and learning approach for neural-symbolic stream fusion. In *AAAI 2021, February 2-9*, pages 4996–5005. AAAI Press, 2021.

[Lehmann, 1995] D. J. Lehmann. Another perspective on default reasoning. *Ann. Math. Artif. Intell.*, 15(1):61–82, 1995.

[Lewis, 1973] D. Lewis. *Counterfactuals*. Basil Blackwell Ltd, 1973.

[Nute, 1980] D. Nute. Topics in conditional logic. *Reidel, Dordrecht*, 1980.

[Serafini and d'Avila Garcez, 2016] L. Serafini and A. S. d'Avila Garcez. Learning and reasoning with logic tensor networks. In *AI*IA 2016*, volume 10037 of *LNCS*, pages 334–348. Springer, 2016.

[Setzu *et al.*, 2021] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti. GlocalX - from local to global explanations of black box AI models. *Artif. Intell.*, 294:103457, 2021.

[Zadeh, 1968] L. Zadeh. Probability measures of fuzzy events. *J.Math.Anal.Appl*, 23:421–427, 1968.