A conditional, a fuzzy and a probabilistic interpretation of self-organising maps (Extended Abstract)

Laura Giordano[1] , **Valentina Gliozzi**[2] , **Daniele Theseider Dupré**[1]
[1] DISIT - Università del Piemonte Orientale, Italy
[2] Center for Logic, Language and Cognition & Dipartimento di Informatica, Università di Torino, Italy
{laura.giordano, dtd}@uniupo.it, valentina.gliozzi@unito.it

This extended abstract reports about the work in (Giordano, Gliozzi, and Theseider Dupré 2022), concerning a logical interpretation of Self-Organising Maps (SOMs) (Kohonen, Schroeder, and Huang 2001), based on a multi-preferential semantics for weighted conditionals, as well as on a fuzzy semantics. The work stems from the area of conditional and preferential reasoning. In fact, preferential approaches to common sense reasoning (e.g., by Pearl (1990), by Kraus Lehmann and Magidor (1990), by Lehmann (1992), by Benferhat et al. (1993)) 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 (Giordano et al. 2007; Britz, Heidema, and Meyer 2008) and closure constructions (starting from Casini and Straccia's work (2010)) have been proposed for defeasible DLs.

Fuzzy description logics have also been widely studied in the literature for representing vagueness in DLs (see (Lukasiewicz and Straccia 2009) for a survey), based on the idea that concepts and roles can be interpreted as fuzzy sets and fuzzy binary relations.

The paper aims at developing a logical interpretation of SOMs after training. SOMs have been proposed as possible candidates to explain the psychological mechanisms underlying category generalisation. They are psychologically and biologically plausible neural network models that can also learn after limited exposure to positive category examples, without any need of contrastive information. We consider a "concept-wise" multi-preferential semantics, which has been first introduced as a semantics of ranked knowledge bases in a lightweight DL (Giordano and Theseider Dupré 2020), and takes into account preferences with respect to different concepts. It is shown that both the multi-preferential semantics and a fuzzy semantics can be used to provide a logical interpretation of SOMs, and to allow for the verification of properties of a trained SOM by model checking.

Both interpretations are based on the idea of associating each learned category to a concept in the language of the simple description logic $\mathcal{LC}$, which does not allow for roles and role restrictions, but allows for the boolean combination of concepts. We show that the learning process in self-organising maps produces, as a result, either a *fuzzy model*, in which each concept (or learned category) is interpreted as a fuzzy set over the domain of input stimuli, or a *multipreference model* by associating a preference relation to each concept (each learned category). Both models can be exploited to extract or validate knowledge from the empirical data used in the learning process and the validation can be done by model checking. The verification of logical properties of a neural network can be useful for post-hoc explanation, in view of a trustworthy, reliable and explainable AI (Adadi and Berrada 2018; Guidotti et al. 2019).

Concerning the preferential semantics, based on the assumption that the abstraction process in the SOM is able to identify the most typical members of a given category, in the semantic representation, we identify some specific stimuli as the *typical exemplars* of the category, and define a preference relation among exemplars. To this purpose, we use the notion of distance of an input stimulus from a category representation. 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 being more typical than y as a horse ($x <_{Horse} y$), but less typical than y as a zebra ($y <_{Zebra} x$). A preferential interpretation can be built over the domain of input stimuli (plus the *best matching units*), and used for checking properties such as: "are typical instances of C_1 also instances of C_2?", by exploiting the fact that the map is organized topologically.

To develop a fuzzy interpretation of SOMs as *fuzzy DL interpretations*, the paper exploits the notion of relative distance introduced by Gliozzi and Plunkett (2019) in their similarity-based account of category generalization based on SOMs. This is done by interpreting each category (concept) as a fuzzy set mapping each input stimulus to a value in $[0, 1]$, based on the map's generalization degree of category membership to the stimulus as in (Gliozzi and Plunkett 2019). A fuzzy model of the SOM is defined as a fuzzy $\mathcal{LC}$ interpretation. As for the multipreference semantics, model checking can be used for the verification of inclusions (strict, defeasible or fuzzy inclusions) over the fuzzy model of the SOM (e.g., "are the instances of category C_1 also instances of C_2 with a degree ≥ 0.8?"). Starting from the fuzzy interpretation of the SOM the paper also provides a probabilistic interpretation of this neural network model based on Zadeh's probability of fuzzy events (Zadeh 1968).

The strong relations between the logics of commonsense reasoning and SOMs also extend to other neural network models, in particular, to Multilayer Perceptrons (MLPs) (Haykin 1999). For MLPs, under a fuzzy multi-preferential semantics, a deep neural network can itself be regarded as a conditional knowledge base (Giordano and Theseider Dupré 2021), where conditional implications are associated to synaptic connections with their weights. Conditional implications with a weight can as well be extracted from a SOM.

Conditional logic belong to a family of logics which are normally used for hypothetical and counterfactual reasoning, for common sense reasoning, and for reasoning with exceptions. That one such logic can be used for capturing reasoning in a deep neural network model can be rather surprising. It suggests that slow thinking and fast thinking (Kahneman 2011) might be more related than expected.

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 or fuzzy) properties.

The work summarized in this abstract 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. Undecidability results for fuzzy description logics motivate the investigation of many-valued semantics for weighted conditional knowledge bases. In the finitely many-valued case multipreference entailment is decidable for weighted $\mathcal{LC}$ knowledge bases and can be computed based on ASP encodings (Giordano and Theseider Dupré 2022). Whether a mapping of multilayer networks to weighted conditional KBs can be extended to other neural network models is an issue for future investigation.

References

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

Benferhat, S.; Cayrol, C.; Dubois, D.; Lang, J.; and Prade, H. 1993. Inconsistency management and prioritized syntax-based entailment. In *Proc. IJCAI'93, Chambéry, France, August 28 - September 3, 1993*, 640–647. Morgan Kaufmann.

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.

Giordano, L., and Theseider Dupré, D. 2020. An ASP approach for reasoning in a concept-aware multipreferential lightweight DL. *Theory and Practice of Logic programming, TPLP* 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 *JELIA 2021*, 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. *ICLP 2022*, to appear in TPLP, preliminary version in https://arxiv.org/abs/2202.01123.

Giordano, L.; Gliozzi, V.; Olivetti, N.; and Pozzato, G. L. 2007. Preferential Description Logics. In *LPAR 2007*, volume 4790 of *LNAI*, 257–272. Yerevan, Armenia: Springer.

Giordano, L.; Gliozzi, V.; and Theseider Dupré, D. 2022. A conditional, a fuzzy and a probabilistic interpretation of self-organising maps. *Journal of Logic and Computation* 32(2):178–205.

Gliozzi, V., and Plunkett, K. 2019. 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).

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.

Kahneman, D. 2011. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

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. *Artificial Intelligence* 44(1-2):167–207.

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

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

Lukasiewicz, T., and Straccia, U. 2009. Description logic programs under probabilistic uncertainty and fuzzy vagueness. *Int. J. Approx. Reason.* 50(6):837–853.

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

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

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