

# From Common Sense Reasoning to Neural Network Models: a Conditional and Multi-preferential Approach for Explainability and Neuro-Symbolic Integration (an Overview)

Francesco Bartoli<sup>1</sup>, Marco Botta<sup>1</sup>, Roberto Esposito<sup>1</sup>, Laura Giordano<sup>2</sup>,  
Valentina Gliozzi<sup>1</sup> and Daniele Theseider Dupré<sup>2</sup>

<sup>1</sup>*Dipartimento di Informatica, Università di Torino, Italy*

<sup>2</sup>*DISIT - Università del Piemonte Orientale, Alessandria, Italy*

## Abstract

This short paper reports about a line of research exploiting a conditional logic of commonsense reasoning to provide a semantic interpretation to neural network models. A “concept-wise” multi-preferential semantics for conditionals is exploited to build a preferential interpretation of a trained neural network starting from its input-output behavior. The approach is general (model agnostic): it is based on a notion of metric distance to define preferences and has been first proposed for Self-Organising Maps (SOMs). For MultiLayer Perceptrons (MLPs), a deep network can as well be regarded as a (fuzzy) conditional knowledge base (KB), in which the synaptic connections correspond to weighted conditionals. This opens to the possibility of adopting conditional description logics as a basis for neuro-symbolic integration. Proof methods for many-valued weighted conditional KBs have been developed, based on Answer Set Programming and Datalog encodings to deal with the entailment and model-checking problems.

## Keywords

Preferential Description Logics, Typicality, Neural Networks, Explainability

## 1. Introduction

Preferential approaches to common sense reasoning (e.g., [1]) have their roots in conditional logics [2, 3], 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 [4, 5, 6, 7] and closure constructions (e.g., [8, 9, 10]) have been proposed for such defeasible DLs. In this paper, we report about a concept-wise multi-

---

*8th Workshop on Formal and Cognitive Reasoning, September 19, 2022, Trier, Germany*


✉ francesco.bartoli@edu.unito.it (F. Bartoli); marco.botta@unito.it (M. Botta); roberto.esposito@unito.it (R. Esposito); laura.giordano@uniupo.it (L. Giordano); gliozzi@di.unito.it (V. Gliozzi); dtd@uniupo.it (D. Theseider Dupré)

🌐 <http://informatica.unito.it/persono/marco.botta/> (M. Botta); <http://informatica.unito.it/persono/roberto.esposito> (R. Esposito); <http://people.unipmn.it/laura.giordano/> (L. Giordano); <http://www.di.unito.it/~gliozzi/> (V. Gliozzi); <http://people.unipmn.it/dtd/> (D. Theseider Dupré)

🆔 0000-0001-5366-292X (R. Esposito); 0000-0001-9445-7770 (L. Giordano); 0000-0001-6798-4380 (D. Theseider Dupré)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

preferential semantics [11], first introduced as a semantics of ranked knowledge bases in a description logic (DL) to account for preferences with respect to different concepts, and later extended to weighted conditional knowledge bases and proposed as a semantics for some neural network models [12, 13, 14].

We deal with both an unsupervised model, Self-organising maps (SOMs) [15], which is considered as a psychologically and biologically plausible neural network model, and a supervised one, MultiLayer Perceptrons (MLPs) [16]. Learning algorithms in the two cases are quite different but our project is to capture in a semantic interpretation the behavior of the network after training and not to provide a logical characterization of the learning process.

In both cases, considering a domain of input stimuli presented to the network e.g., during training or generalization), a semantic interpretation describing the *input-output behavior* of the network can be provided as a multi-preferential interpretation, where preferences are associated to concepts. For SOMs, the learned categories  $C_1, \dots, C_n$  are regarded as concepts so that a preference relation over the domain of input stimuli is associated with each category [12, 14]. For MLPs, each unit of interest in the deep network (including hidden units) can be associated with a concept and with a preference relation on the domain [13].

The idea is that, for 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* 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* on the preferential 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 multi-preferential semantics for DLs with weighted defeasible inclusions. Some different preferential closure constructions have been considered for weighted knowledge bases (the *coherent* [13], *faithful* [17] and  $\varphi$ -*coherent* [18] multi-preferential semantics), and their relationships with MLPs have been investigated (see [13, 18]).

Undecidability results for fuzzy DLs with general inclusion axioms [19, 20] have motivated the investigation of many-valued approximations of fuzzy multi-preferential entailment. The semantics above have been reconsidered in the finitely many-valued case. In [21] an ASP-based approach has been exploited for reasoning with weighted conditional KBs under  $\varphi$ -coherent entailment. Datalog with weakly stratified negation has been used for developing a model-checking approach for MLPs, still in the many-valued case [22, 23]. Both the entailment and the model-checking approaches have been experimented in the verification of properties of some trained multilayer feedforward networks and, specifically, in the verification of properties of neural networks for the recognition of basic emotions.

The strong relationships between neural networks and conditional logics of commonsense reasoning suggest that conditional logics can be used for the verification of properties of neural networks to explain their behavior. The possibility of combining symbolic knowledge with elicited knowledge in the same formalism is a step towards neuro-symbolic integration, in the

direction of a trustworthy and explainable AI [24, 25, 26].

## 2. The concept-wise multi-preferential semantics

The concept-wise multi-preferential semantics ( $\text{cw}^m$ -semantics) has been introduced as a semantics for ranked  $\mathcal{EL}$  knowledge bases [11], and later extended to weighted knowledge bases [13]. 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  $C$ s are  $D$ s” or “normally  $C$ s are  $D$ s”) with a rank (resp. a weight). They correspond to KLM conditionals  $C \sim D$  [1]. 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 multi-preferential interpretations, based on different constructions.

*Concept-wise multi-preferential interpretations* ( $\text{cw}^m$ -interpretations) are defined by adding to standard DL interpretations (which are pairs  $I = \langle \Delta, \cdot^I \rangle$ , where  $\Delta$  is a domain, and  $\cdot^I$  an interpretation function) the preference relations  $<_{C_1}, \dots, <_{C_n}$  associated with a set of distinguished concepts  $C_1, \dots, C_n$ , 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  $\Delta$ . Preferences with respect to different concepts do not need to agree, as we have seen. In the two-valued case, 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  $<_{C_i}$ , but a more sophisticated global preference, taking into account specificity, has also been considered [11]. It has been proven therein that global preference in a  $\text{cw}^m$ -interpretation determines a KLM-style preferential interpretation, and  $\text{cw}^m$ -entailment satisfies the KLM postulates of a preferential consequence relation [1].

In the Sections 3 and 4 we will see that, both for SOMs and for MLPs, a multi-preferential interpretation can be constructed from the input-output behavior of the network over a set of input stimuli, and can be used for model checking.

## 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 multi-preferential 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 [27] 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 distance  $d(x, C_i)$  of a stimulus  $x$  from a category  $C_i$  can be used to build a binary preference relation  $<_{C_i}$  among the stimuli in  $\Delta$  with respect to category  $C_i$  [14, 12], by letting  $x <_{C_i} y$  if and only if  $d(x, C_i) > d(y, C_i)$  ( $x$  is more typical than  $y$  with respect to category  $C_i$  if its distance from category  $C_i$  is lower than the relative distance of  $y$ ). Based on the assumption that the abstraction process in the SOM is able to identify the

most typical exemplars for a given category, in the semantic representation of a category, some specific stimuli (corresponding to the *best matching units*) are identified as the *typical exemplars* of the category.

A notion of *relative distance*, introduced by Gliozzi and Plunkett in their similarity-based account of category generalization based on self-organising maps [27], is also used for developing another semantic interpretation of SOMs based on *fuzzy DL interpretations*. This is done by interpreting each category (concept) as a function mapping each input stimulus to a value in  $[0, 1]$ , based on the *map's generalization degree* of category membership to the stimulus [27].

In both the two-valued and fuzzy case, the 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. 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?). Starting from the fuzzy interpretation of the SOM, a probabilistic interpretation of this neural network model is also provided [14], based on Zadeh's probability of fuzzy events [28], and on Montes et al. [29] recent characterization of the continuous t-norms compatible with Zadeh's probability of fuzzy events.

#### 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 a domain  $\Delta$  of input stimuli, e.g., those stimuli considered during training or generalization [13]. Each neuron  $k$  of interest for property verification can be associated to a distinguished concept  $C_k$ . For each concept  $C_k$ , a preference relation  $<_{C_k}$  is defined over the domain  $\Delta$  based on the activity values,  $y_k(v)$ , of neuron  $k$  for each input  $v \in \Delta$ . In a similar way, a fuzzy interpretation of the network can be constructed over the domain  $\Delta$ , as well as a fuzzy multi-preferential 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 exploit the activity level of neurons for the stimuli. In the fuzzy semantics, the interpretation of a concept  $C_k$  is a mapping  $C_k^I : \Delta \rightarrow [0, 1]$ , associating to each  $x \in \Delta$  the degree of membership of  $x$  in  $C_k$ . The activation value  $y_k(x)$  of neuron  $k$  for a stimulus  $x$  in the network (assumed to be in the interval  $[0, 1]$ ) is taken to be the degree of membership of  $x$  in concept  $C_h$ . The fuzzy interpretation also induces a preference  $<_{C_h}$  on  $\Delta$ .

The interpretation of boolean concepts is defined by fuzzy combination functions, as usual in fuzzy DLs [30, 31]. This also allows a preference relation  $<_C$  to be associated to any concept  $C$ , and the typical  $C$ -elements to be identified, provided the interpretation is well-founded (an assumption which clearly holds when the domain  $\Delta$  is finite, as in this case). Let us call  $\mathcal{M}_{\mathcal{N}}^{f, \Delta}$  the fuzzy multi-preferential interpretation built from network  $\mathcal{N}$  over a domain  $\Delta$ .

As for SOMs, logical properties of the network (including fuzzy typicality inclusions) can then be verified by model checking over such an interpretation. Evaluating properties involving hidden units might be as well of interest. We refer to the typicality properties considered in the verification examples in Sections 6.1 and 6.2.

The fuzzy multi-preferential interpretation  $\mathcal{M}_{\mathcal{N}}^{f,\Delta}$  described above can be 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. Let us introduce a notion of weighted conditional KB.

## 5. Weighted conditional knowledge bases and MultiLayer Perceptrons

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

A weighted  $\mathcal{ALC}$  knowledge base contains, besides standard inclusion axioms (Tbox  $\mathcal{T}$ ) and assertions (Abox  $\mathcal{A}$ ), a set  $\mathcal{C} = \{C_1, \dots, C_k\}$  of distinguished  $\mathcal{ALC}$  concepts and, for each  $C_i$ , a set of weighted typicality inclusions of the form  $\mathbf{T}(C_i) \sqsubseteq D_j$ , with a positive or negative weight  $w_{i,j}$  (a real number). In the fuzzy case,  $\mathcal{T}$  and  $\mathcal{A}$  contain fuzzy axioms [31].

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

$$\begin{array}{ll} (d_1) \mathbf{T}(Bird) \sqsubseteq Fly, +20 & (d_2) \mathbf{T}(Bird) \sqsubseteq \exists has\_Wings.\top, +50 \\ (d_3) \mathbf{T}(Bird) \sqsubseteq \exists has\_Feathers.\top, +50 & (d_4) \mathbf{T}(Penguin) \sqsubseteq Fly, -70 \\ (d_5) \mathbf{T}(Penguin) \sqsubseteq Black, +50 & (d_6) \mathbf{T}(Penguin) \sqsubseteq Grey, +10 \end{array}$$

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.

In the two-valued case, a semantics for weighted  $\mathcal{ALC}$  knowledge bases can be defined with a semantic closure construction in the spirit of Lehmann's lexicographic closure [32], but more similar to Kern-Isberner's semantics of c-representations [33, 34], 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  $\Delta$ , 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 multi-preferential 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. Specifically, let  $\mathcal{T}_{C_i} = \{(d_h^i, w_h^i)\}$  be the set of all weighted typicality inclusions  $d_h^i = \mathbf{T}(C_i) \sqsubseteq D_{i,h}$  for the distinguished concept  $C_i$ , for each domain element  $x \in \Delta$ , the weight  $W_i(x)$  of  $x$  wrt  $C_i$  in a fuzzy interpretation  $I = \langle \Delta, \cdot^I \rangle$  is the sum:  $W_i(x) = \sum_h w_h^i D_{i,h}^I(x)$ .

To guarantee that the preferences determined from the knowledge base are coherent with the fuzzy interpretation of concepts, some different semantic constructions have been considered, namely the notions of *coherent* [13], *faithful* [17] and  $\varphi$ -*coherent* [18, 35] (fuzzy) multi-preferential semantics. Specifically, for coherent interpretations we require that:

$$C_i^I(x) < C_i^I(y) \iff W_i(x) < W_i(y)$$

The notion of  $\varphi$ -coherence of a fuzzy interpretation  $I$  wrt a KB exploits a function  $\varphi$  from  $\mathbb{R}$  to the interval  $[0, 1]$ , i.e.,  $\varphi : \mathbb{R} \rightarrow [0, 1]$ . By slightly generalizing the fuzzy multi-preferential semantics introduced as a gradual argumentation semantics in [18], we assume that different functions  $\varphi_i$  are associated to the distinguished concepts  $C_i$ .

An interpretation  $I = \langle \Delta, \cdot^I \rangle$  is  $\varphi$ -coherent if, for all concepts  $C_i \in \mathcal{C}$  and  $x \in \Delta$ ,

$$C_i^I(x) = \varphi_i\left(\sum_h w_h^i D_{i,h}^I(x)\right) \quad (1)$$

where  $\mathcal{T}_{C_i} = \{(\mathbf{T}(C_i) \sqsubseteq D_{i,h}, w_h^i)\}$  is the set of weighted conditionals for  $C_i$ . A  $\varphi$ -coherent model of knowledge base  $K$ , is defined as a fuzzy interpretation  $I$  satisfying TBox  $\mathcal{T}$ , ABox  $\mathcal{A}$  and the  $\varphi$ -coherence condition (1).

As usual in preferential semantics, we restrict to canonical models, which are large enough to contain a domain element for any possible valuation of concepts which is present in some  $\varphi$ -coherent model of  $K$ . Based on a notion of  $\varphi$ -coherent canonical model of a weighted knowledge base [21], a notion of  $\varphi$ -coherent entailment can be defined as expected.

A mapping of a neural network  $\mathcal{N}$  to a conditional KB  $K^{\mathcal{N}}$  can be defined in a simple way [13], associating a concept name  $C_i$  with each unit  $i$  in the network and by introducing, for each synaptic connection from neuron  $h$  to neuron  $i$  with weight  $w_{ih}$ , a conditional  $\mathbf{T}(C_i) \sqsubseteq C_h$  with weight  $w_h^i = w_{ih}$  in  $K^{\mathcal{N}}$ . If we assume that the activation functions  $\varphi_i$  of the units in the network  $\mathcal{N}$  return values in the interval  $[0, 1]$ , then the solutions of equations (1) characterize the stationary states of MLPs, where  $C_i^I(x)$  corresponds to the activation of neuron  $i$  for some input stimulus  $x$ , each  $D_{i,h}^I(x)$  corresponds to an input signal  $x_h$  to neuron  $i$ , and  $\sum_h w_h^i D_{i,h}^I(x)$  corresponds to the induced local field of neuron  $i$  [16].

Let us consider the fuzzy multi-preferential interpretation  $\mathcal{M}_{\mathcal{N}}^{f,\Delta}$  built from  $\mathcal{N}$  over a domain  $\Delta$  of input stimuli, as described in Section 4, and assume that a concept  $C_k$  is introduced in the language for each unit  $k$ . It has been proven [13] that the interpretation  $\mathcal{M}_{\mathcal{N}}^{f,\Delta}$  is a coherent fuzzy multi-preferential 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 then satisfied by  $\mathcal{M}_{\mathcal{N}}^{f,\Delta}$ , for any choice of the domain  $\Delta$ .

## 6. ASP and Datalog for reasoning about neural networks in the many-valued case: from entailment to model-checking

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.

Undecidability results for fuzzy DLs with general inclusion axioms [19, 20] have motivated the investigation of many-valued approximations of fuzzy multi-preferential entailment. The semantics above have been reconsidered in the finitely many-valued case. In [21] an ASP-based approach has been exploited for reasoning with weighted conditional KBs under  $\varphi$ -coherent

entailment. Datalog with weakly stratified negation has been used for developing a model-checking approach for MLPs, still in the many-valued case [22, 23]. Both the entailment and the model-checking approaches have been experimented in the verification of properties of some trained multilayer feedforward networks.

### 6.1. The entailment approach

Reasoning on weighted KBs associated to neural networks, based on a multi-valued truth space  $\mathcal{C}_n = \{0, \frac{1}{n}, \dots, \frac{n-1}{n}, \frac{n}{n}\}$ , for an integer  $n \geq 1$ , requires introducing, for each activation function  $\varphi$ , a function  $\varphi_n$  which approximates  $\varphi(x)$  to the nearest value in  $\mathcal{C}_n$ . A notion of  $\varphi_n$ -coherence is defined (the analog of  $\varphi$ -coherence in Sec. 5), and the corresponding  $\varphi_n$ -coherent entailment, i.e., satisfaction in all  $\varphi_n$ -coherent models.

In particular, we consider the entailment of a typicality inclusion such as  $\mathbf{T}(C) \sqsubseteq D \geq \alpha$  from a weighted knowledge base  $K$  in the finitely many-valued Gödel description logic with typicality  $G_n\mathcal{LCT}$ , introduced in [21] for the boolean fragment  $\mathcal{LC}$  of  $\mathcal{ALC}$ . Such a verification can be formulated as a problem of computing *preferred answer sets* of an ASP program, considering a single distinguished domain element  $aux_C$ , intended to represent a typical  $C$ -element, and selecting, as preferred answer sets, the ones maximizing the membership of  $aux_C$  in concept  $C$ . Answer sets maximizing the membership of  $aux_C$  in concept  $C$  can be selected with an *asprin* preference program, and represent those inputs stimuli associated with typical  $C$ -elements. For all typical  $C$ -elements it is verified that membership in concept  $D$  is greater than  $\alpha$ .

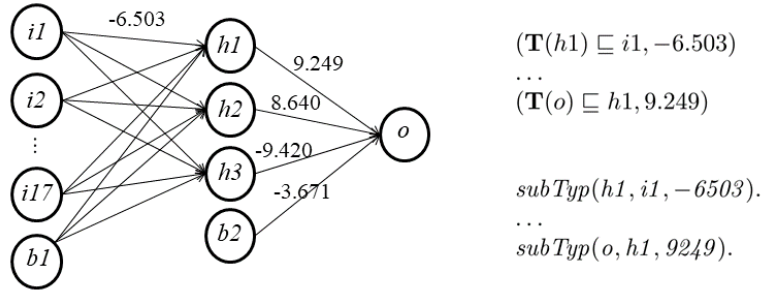
As a proof of concept, in [21] the entailment approach has been experimented for the weighted  $G_n\mathcal{LCT}$  KBs corresponding to two of the trained multilayer feedforward network for the MONK's problems ([36]). The networks have 17 non-independent binary inputs, corresponding to values of 6 inputs having 2 to 4 possible values; such inputs are features of a robot, e.g., head shape and body shape being round, square or octagon, and jacket color being red, yellow, green or blue. The network for problem 1 has 3 hidden units ( $h1, h2, h3$ ) and an output unit ( $o$ ); the one for problem 3 has 2 hidden units.

For example, in the first problem, the trained network learned to classify inputs satisfying a formula  $F1 \equiv jacket\_color\_red \text{ or } head\_shape = body\_shape$  which, in terms of the classes  $i1, \dots, i17$  corresponding to the binary inputs, is:  $F1 \equiv i12 \sqcup (i1 \sqcap i4) \sqcup (i2 \sqcap i5) \sqcup (i3 \sqcap i6)$ .

For instance, the formula  $\mathbf{T}(o) \sqsubseteq F1 \geq 1$  can be verified for e.g.  $n = 5$ , where  $o$  is the concept name associated with the output unit. That is, the  $G_5\mathcal{LCT}$  knowledge base entails that the typical  $o$ -elements satisfy  $F1$ . Stronger variants of  $F1$  have also been considered, to check that the network learned  $F1$  but not such variants. The following formulae have been verified for hidden nodes  $h1, h2, h3$ :  $\mathbf{T}(h1) \sqsubseteq i12 \sqcup (\neg i1 \sqcap \neg i4) \geq 1$ ,  $\mathbf{T}(h2) \sqsubseteq i12 \sqcup (\neg i3 \sqcap \neg i6) \geq 1$ ,  $\mathbf{T}(h3) \sqsubseteq \neg i12 \sqcup (i2 \sqcup i5) \geq 1$ .

### 6.2. The model-checking approach

Based on the general idea of using model-checking for verifying the properties of a neural network, as described in Section 4 for MLPs, in [22] we have developed a Datalog-based approach



**Figure 1:** The network for MONK’s problem 1, with some of the weights after training, two of the corresponding typicality inclusions and their ASP representation [21].

which builds a multi-valued preferential interpretation of a trained feedforward network  $\mathcal{N}$  and, then, verifies the properties of the network for post-hoc explanation.

The Datalog encoding contains a component  $\Pi(\mathcal{N}, \Delta, n)$  which is intended to build a (single) many-valued, preferential interpretation with truth degrees in  $\mathcal{C}_n$ , and a component associated to the formulae to be checked. We exploited Datalog with weakly stratified negation. The model checking approach has been experimented in the verification of properties of neural networks for the recognition of basic emotions using the Facial Action Coding System (FACS) [37].

The RAF-DB [38] data set contains almost 30000 images labeled with basic emotions or combinations of two emotions. It was used as input to OpenFace 2.0 [39], which detects a subset of the Action Units (AUs) in [37], i.e., facial muscle contractions. The relations between such AUs and emotions, studied by psychologists [40], can be used as a reference for formulae to be verified on neural networks trained to learn such relations.

From the original dataset, we selected the subset of the images that were labelled using only one emotion in the set  $\{surprise, fear, happiness, anger\}$ . The dataset was highly unbalanced and we preprocessed the data by subsampling the larger classes and augmenting the minority ones using standard data-augmentation techniques. The processed dataset contains 5 975 images (the number of images was 4 283 before augmentation). The images were input to OpenFace 2.0; the output intensities were rescaled in order to make their distribution conformant to the expected one in case AUs were recognized by humans [37]. The resulting AUs were used as input to a neural network trained to classify its input as an instance of the four emotions. The neural network model we used is a fully connected feed forward neural network with three hidden layers having 1 800, 1 200, and 600 nodes (all hidden layers use RELU activation functions, while the softmax function is used in the output layer).

The model checking approach was applied, using the Clingo ASP solver as Datalog engine, taking as set of input stimuli  $\Delta$  the test set, containing 1194 images, and  $n = 5$  (given that AU intensities, when assigned by humans, are on a scale of five values). Table 1 reports some results for the verification of typicality inclusions  $\mathbf{T}(E) \sqsubseteq F \geq k/n$ , with the number of typical individuals for the emotion  $E$ , the number of counterexamples for different values of  $k$ , as well as the value of the conditional probabilities  $p(F/\mathbf{T}(E))$  of concept  $F$  given concept  $\mathbf{T}(E)$ , based on Zadeh’s probability of fuzzy events [28]. The approach is the one adopted to develop a probabilistic interpretation of SOMs after training, starting from a fuzzy interpretation



E	F	#counterexamples				#T(E)	P(F/T(E))
		K=1	K=2	K=3	K=4		
Happiness	AU1 $\sqcup$ AU6 $\sqcup$ AU12 $\sqcup$ AU14	0	0	0	22	255	0.8634
	AU6 $\sqcup$ AU12	0	0	1	32	255	0.8422
	AU6 $\sqcap$ AU12	6	15	23	98	255	0.7136
	AU12	0	0	1	35	255	0.8344

**Table 1**  
Results for checking formulae on the test set

[14]. It exploits a recent characterization of the continuous t-norms compatible with Zadeh’s probability of fuzzy events ( $P_Z$ -compatible t-norms) by Montes et al. [29]. To compute the conditional probabilities, we have assumed a uniform probability distribution over  $\Delta$ . Note that also typicality concepts can occur in conditional probabilities.

For example,  $\mathbf{T}(happiness) \sqsubseteq au12 \geq 3/5$  (where  $au12$  is the activation of the lip corner puller muscle, that is, smiling) does not hold as it has 1 counterexample out of 255 instances of  $\mathbf{T}(happiness)$ . The value of  $P(au12/\mathbf{T}(happiness))$  is larger than  $4/5$ , even though there are 35 counterexamples for  $\mathbf{T}(happiness) \sqsubseteq au12 \geq 4/5$ .

### 6.3. Further considerations

For the emotion recognition problem in Section 6.2, in the entailment approach the grounding size of the ASP program only allowed to deal with a network using boolean inputs for the 17 AUs considered, a layer of 8 hidden units, and a single output for deciding membership to a single emotion. For happiness, in particular, with  $n=9$  (i.e. 10 discrete values) the formula  $T(happiness) \sqsubseteq au6 \sqcup au12 \geq 1$  was found to have 4 counterexamples among the  $2^{17}$  combinations of boolean inputs, 1446 being instances of  $T(happiness)$ . Interestingly enough, such 4 combinations do not occur in the data set (indeed, only a small fraction of  $2^{17}$ , i.e., 131072 combinations may occur in a few thousand images).

As expected, the model-checking approach outperforms the entailment approach. In fact, the model checking approach considers a subset of all the possible inputs to the network, and the verification problem is polynomial in time in the size of the domain  $\Delta$  and in the size of the formula to be verified [22]. On the other hand, all the possible combinations of the values of all units (including hidden ones) need to be considered in the entailment-based approach. This was the reason for limiting the size of the network (and, specifically, the number of units in the hidden layers). Note, the entailment approach has been developed for general weighted conditional knowledge bases, which are not required to be acyclic, while in the experimentation we have considered feedforward networks.

A multilayer network can be seen as a set of weighted defeasible inclusions in a simple description logic (only including boolean concepts). However, a weighted conditional knowledge base can be more general. It can be defined for several DLs including roles (as it has been done, for instance, for  $\mathcal{EL}$  [13] and for  $\mathcal{ALC}$  [17]), and it allows for general inclusions axioms and assertions. The combination of defeasible inclusions with strict (or fuzzy) inclusions and assertions in a weighted KB allows for the combination of the knowledge acquired from the

network and symbolic knowledge in the same formalism. In the entailment based approach this has been exploited in several ways, by adding constraints on the possible inputs through ABox and TBox axioms (e.g., to exclude combinations of input values).

## 7. Conclusions

Conditional logics of commonsense reasoning can be used for interpreting and verifying the knowledge learned by a neural network for post-hoc explanation and, for MLPs, a trained network can itself be seen as a conditional knowledge base.

Much work has been devoted to the combination of neural networks and symbolic reasoning (e.g., the work by d’Avila Garcez et al. [41, 42, 43] and Setzu et al. [44]), as well as to the definition of new computational models [45, 46, 47, 48]. The work summarized in this paper 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.

Using a multi-preferential logic for the verification of typicality properties of a neural network by model-checking is a general (*model agnostic*) approach. It can be used for SOMs, as in [12, 14], by exploiting a notion of *distance* of a stimulus from a category to define a preferential structure, as well as for MLPs, by exploiting units activity to build a fuzzy preferential interpretation. Given the simplicity of the approach, a similar construction can be adapted to other network models and learning approaches, and used in applications combining different network models (as in the mentioned experiment to the recognition of basic emotions).

Both the model-checking approach and the entailment-based approach are *global* approaches (see, e.g., [44] for the notions of local and global approaches), as they consider the behavior of the network over a set  $\Delta$  of input stimuli. Indeed, the evaluation of typicality inclusions considers all the individuals in the domain to establish preference relations among them, with respect to different aspects. However, properties of single individuals can as well be verified (by instance checking, in DL terminology).

The model-checking approach does not require to consider the activity of all units, but only of the units involved in the property to be verified. In the entailment-based approach, on the other hand, all units are considered. This limits its range of applicability to simple networks.

The entailment-based approach is based on the idea of regarding a multilayer network as weighted conditional knowledge base, and is specific for this network model. For MLPs, it has been proven that, in the fuzzy case, the interpretation built for model-checking is indeed a model of the weighted conditional KB corresponding to the network [13]. Whether it is possible to extend the logical encoding of MLPs as weighted KBs to other neural network models is a subject for future investigation. The development of a temporal extension of this formalism to capture the transient behavior of MLPs is also an interesting direction to extend this work.

**Acknowledgement:** This research is partially supported by Università del Piemonte Orientale and by INDAM-GNCS Project 2022 “Logiche non-classiche per tool intelligenti ed explainable”.

## References

- [1] S. Kraus, D. Lehmann, M. Magidor, Nonmonotonic reasoning, preferential models and cumulative logics, *Artificial Intelligence* 44 (1990) 167–207.
- [2] D. Lewis, *Counterfactuals*, Basil Blackwell Ltd, 1973.
- [3] D. Nute, *Topics in conditional logic*, Reidel, Dordrecht (1980).
- [4] L. Giordano, V. Gliozzi, N. Olivetti, G. L. Pozzato, Preferential Description Logics, in: *LPAR 2007*, volume 4790 of *LNAI*, Springer, Yerevan, Armenia, 2007, pp. 257–272.
- [5] K. Britz, J. Heidema, T. Meyer, Semantic preferential subsumption, in: G. Brewka, J. Lang (Eds.), *KR 2008*, AAAI Press, Sidney, Australia, 2008, pp. 476–484.
- [6] G. Casini, T. A. Meyer, I. Varzinczak, Contextual conditional reasoning, in: *AAAI-21, Virtual Event*, February 2-9, 2021, AAAI Press, 2021, pp. 6254–6261.
- [7] L. Giordano, V. Gliozzi, A reconstruction of multipreference closure, *Artif. Intell.* 290 (2021).
- [8] G. Casini, U. Straccia, Rational Closure for Defeasible Description Logics, in: T. Janhunen, I. Niemelä (Eds.), *JELIA 2010*, volume 6341 of *LNCS*, Springer, Helsinki, 2010, pp. 77–90.
- [9] G. Casini, T. Meyer, K. Moodley, R. Nortje, Relevant closure: A new form of defeasible reasoning for description logics, in: *JELIA 2014*, LNCS 8761, Springer, 2014, pp. 92–106.
- [10] L. Giordano, V. Gliozzi, N. Olivetti, G. L. Pozzato, Semantic characterization of rational closure: From propositional logic to description logics, *Art. Int.* 226 (2015) 1–33.
- [11] L. Giordano, D. Theseider Dupré, An ASP approach for reasoning in a concept-aware multipreferential lightweight DL, *TPLP* 10(5) (2020) 751–766.
- [12] L. Giordano, V. Gliozzi, D. Theseider Dupré, On a plausible concept-wise multipreference semantics and its relations with self-organising maps, in: F. Calimeri, S. Perri, E. Zumpano (Eds.), *CILC 2020*, Rende, IT, Oct. 13-15, 2020, volume 2710 of *CEUR*, 2020, pp. 127–140.
- [13] L. Giordano, D. Theseider Dupré, 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*, Springer, 2021, pp. 225–242.
- [14] L. Giordano, V. Gliozzi, D. T. Dupré, A conditional, a fuzzy and a probabilistic interpretation of self-organizing maps, *J. Log. Comput.* 32 (2022) 178–205.
- [15] T. Kohonen, M. Schroeder, T. Huang (Eds.), *Self-Organizing Maps*, Third Edition, Springer Series in Information Sciences, Springer, 2001.
- [16] S. Haykin, *Neural Networks - A Comprehensive Foundation*, Pearson, 1999.
- [17] L. Giordano, On the KLM properties of a fuzzy DL with Typicality, in: *Proc. ECSQARU 2021*, Prague, Sept. 21-24, 2021, volume 12897 of *LNCS*, Springer, 2021, pp. 557–571.
- [18] L. Giordano, From weighted conditionals of multilayer perceptrons to a gradual argumentation semantics, in: *5th Workshop on Advances in Argumentation in Artif. Intell.*, 2021, Milan, Italy, Nov. 29, volume 3086 of *CEUR Workshop Proc.*, 2021. URL: <http://ceur-ws.org/Vol-3086/paper8.pdf>.
- [19] M. Cerami, U. Straccia, On the undecidability of fuzzy description logics with gcis with lukasiewicz t-norm, *CoRR* abs/1107.4212 (2011). URL: <http://arxiv.org/abs/1107.4212>.
- [20] S. Borgwardt, R. Peñaloza, Undecidability of fuzzy description logics, in: G. Brewka, T. Eiter, S. A. McIlraith (Eds.), *Proc. KR 2012*, Rome, Italy, June 10-14, 2012, AAAI Press, 2012.

- [21] L. Giordano, D. Theseider Dupré, An ASP approach for reasoning on neural networks under a finitely many-valued semantics for weighted conditional knowledge bases (2022). To appear in TPLP, <https://doi.org/10.1017/S1471068422000163>.
- [22] F. Bartoli, M. Botta, R. Esposito, L. Giordano, D. Theseider Dupré, 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: 4th International Workshop on the Resurgence of Datalog in Academia and Industry*, September 5, 2022, Genova - Nervi, Italy, volume 3203 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 54–67. URL: <http://ceur-ws.org/Vol-3203/paper4.pdf>.
- [23] F. Bartoli, A Typicality-based Interpretation of Neural Networks: an Experiment on Facial Emotion Recognition, Master Thesis in Stochastics and Data Science, University of Torino, 2022.
- [24] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), *IEEE Access* 6 (2018) 52138–52160.
- [25] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM Comput. Surv.* 51 (2019) 93:1–93:42.
- [26] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, *Inf. Fusion* 58 (2020) 82–115.
- [27] V. Gliozzi, 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 (2019).
- [28] L. Zadeh, Probability measures of fuzzy events, *J.Math.Anal.Appl* 23 (1968) 421–427.
- [29] I. Montes, J. Hernández, D. Martinetti, S. Montes, Characterization of continuous t-norms compatible with zadeh’s probability of fuzzy events, *Fuzzy Sets Syst.* 228 (2013) 29–43.
- [30] G. Stoilos, G. B. Stamou, V. Tzouvaras, J. Z. Pan, I. Horrocks, Fuzzy OWL: uncertainty and the semantic web, in: *OWLED\*05 Workshop on OWL Galway, Ireland, Nov 11-12, 2005*, volume 188 of *CEUR Workshop Proc.*, 2005.
- [31] T. Lukasiewicz, U. Straccia, Managing uncertainty and vagueness in description logics for the semantic web, *J. Web Semant.* 6 (2008) 291–308.
- [32] D. J. Lehmann, Another perspective on default reasoning, *Ann. Math. Artif. Intell.* 15 (1995) 61–82.
- [33] G. Kern-Isberner, Conditionals in Nonmonotonic Reasoning and Belief Revision - Considering Conditionals as Agents, volume 2087 of *LNCS*, Springer, 2001.
- [34] G. Kern-Isberner, C. Eichhorn, Structural inference from conditional knowledge bases, *Stud Logica* 102 (2014) 751–769.
- [35] L. Giordano, From weighted conditionals with typicality to a gradual argumentation semantics and back, in: *Proc. 20th International Workshop on Non-Monotonic Reasoning, NMR 2022, Part of FLoC 2022, Haifa, Israel, August 7-9, 2022*, volume 3197 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 127–138.
- [36] Thrun, S. et al., A Performance Comparison of Different Learning Algorithms, Technical Report CMU-CS-91-197, Carnegie Mellon University, 1991.
- [37] P. Ekman, W. Friesen, J. Hager, Facial Action Coding System, Research Nexus, 2002.

- [38] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 2584–2593.
- [39] T. Baltrusaitis, A. Zadeh, Y. C. Lim, L. Morency, Openface 2.0: Facial behavior analysis toolkit, in: 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, IEEE Computer Society, 2018, pp. 59–66.
- [40] B. Waller, J. C. Jr., A. Burrows, Selection for universal facial emotion, *Emotion* 8 (2008) 435–439.
- [41] A. S. d’Avila Garcez, K. Broda, D. M. Gabbay, Symbolic knowledge extraction from trained neural networks: A sound approach, *Artif. Intell.* 125 (2001) 155–207.
- [42] A. S. d’Avila Garcez, L. C. Lamb, D. M. Gabbay, *Neural-Symbolic Cognitive Reasoning, Cognitive Technologies*, Springer, 2009.
- [43] A. S. d’Avila Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, S. N. Tran, Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning, *FLAP* 6 (2019) 611–632.
- [44] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, GlocalX - from local to global explanations of black box AI models, *Artif. Intell.* 294 (2021) 103457. doi:10.1016/j.artint.2021.103457.
- [45] L. C. Lamb, A. S. d’Avila Garcez, M. Gori, M. O. R. Prates, P. H. C. Avelar, M. Y. Vardi, Graph neural networks meet neural-symbolic computing: A survey and perspective, in: C. Bessiere (Ed.), *Proc. IJCAI 2020*, ijcai.org, 2020, pp. 4877–4884.
- [46] L. Serafini, A. S. d’Avila Garcez, Learning and reasoning with logic tensor networks, in: XVth Int. Conf. of the Italian Association for Artificial Intelligence, AI\*IA 2016, Genova, Italy, Nov 29 - Dec 1, volume 10037 of *LNCS*, Springer, 2016, pp. 334–348.
- [47] P. Hohenecker, T. Lukasiewicz, Ontology reasoning with deep neural networks, *J. Artif. Intell. Res.* 68 (2020) 503–540.
- [48] D. Le-Phuoc, T. Eiter, A. Le-Tuan, A scalable reasoning and learning approach for neural-symbolic stream fusion, in: *AAAI 2021*, February 2-9, AAAI Press, 2021, pp. 4996–5005.