Towards a Conditional and Multi-preferential Approach to Explainability of Neural Network Models in Computational Logic (Extended Abstract)

Mario Alviano[1], Francesco Bartoli[2], Marco Botta[2], Roberto Esposito[2], Laura Giordano[3], Valentina Gliozzi[2] and Daniele Theseider Dupré[3]

[1] *Università della Calabria, Italy*
[2] *Università di Torino, Italy*
[3] *Università del Piemonte Orientale, Italy*

Abstract

This short paper reports on 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 a general one; it has first been proposed for Self-Organising Maps (SOMs), and exploited for MultiLayer Perceptrons (MLPs) in the verification of properties of a network by model-checking. An MLPs can be regarded as a (fuzzy) conditional knowledge base (KB), in which the synaptic connections correspond to weighted conditionals. Reasoners for many-valued weighted conditional KBs are under development based on Answer Set solving to deal with entailment and model-checking.

Keywords

Preferential Description Logics, Typicality, Neural Networks, Explainability

1. Introduction

In this short paper we report on an approach to exploit the logic of commonsense reasoning for the explainability of some neural network models. We also report on preliminary experiments in the verification of properties of feedforward neural networks by model checking.

Preferential approaches to commonsense reasoning (e.g., [1]) have their roots in conditional logics [2, 3], and have been more recently extended to Description Logics (DLs), to deal with defeasible reasoning 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 defeasible DLs. Among these, the concept-wise multi-preferential semantics [11], which allows to account for preferences with respect to different concepts. It has been introduced first as a semantics of ranked knowledge bases in a lightweight

3rd Italian Workshop on Explainable Artificial Intelligence (XAI.it 2022), November 30, 2022

mario.alviano@unical.it (M. Alviano); 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)

CEUR Workshop Proceedings (CEUR-WS.org)

description logic (DL) and then for weighted conditional DL knowledge bases, and proposed as a semantics for some neural network models [12, 13, 14].

We have considered both an unsupervised model, Self-organising maps (SOMs) [15], which are considered 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 aim was to capture in a semantic interpretation the behavior of the network after training. Considering a domain of input stimuli presented to a 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 $C_1, \ldots, C_n$ 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].

For MLPs, the relationship between the logic of commonsense reasoning and deep neural networks is even stronger, as the 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 *φ-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 the (finitely) *many-valued* case. An ASP-based approach has been proposed for reasoning with weighted conditional KBs under φ-coherent entailment [21], and Datalog with weakly stratified negation has been used for developing a model-checking approach for MLPs 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. The preliminary results can be the basis for further solutions for the multi-valued φ-coherent entailment, which exploit state of the art ASP solving, including custom propagation based on the *clingo* API [24] and fuzzy ASP solving [25], in the verification of properties of neural networks.

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, in the direction of a trustworthy and explainable AI [26, 27, 28]. The possibility of combining learned knowledge with elicited knowledge in the same formalism is also a step towards neuro-symbolic integration.

2. The concept-wise multi-preferential semantics

The idea underlying the multi-preferential semantics is that, for two domain elements x and y and two concepts, e.g., $Horse$ and $Zebra$, x can be regarded as being more typical than y as a horse ($x <_{Horse} y$), while x could be less typical than y as a zebra ($y <_{Zebra} x$).

This idea has been exploited in the definition of *concept-wise* multi-preferential interpretations [11] for a description logic with typicality concepts (e.g., $\mathbf{T}(Horse)$, representing the class of typical horses), and defeasible inclusions (e.g., $\mathbf{T}(Horse) \sqsubseteq Tall$, meaning that "normally horses are tall"). Typicality inclusions $\mathbf{T}(C) \sqsubseteq D$ correspond to Kraus-Lehmann-Magidor (KLM) conditionals $C \mid\!\sim D$ [1].

Concept-wise multi-preferential interpretations are defined by adding to standard DL interpretations (pairs $I = \langle \Delta, \cdot^I \rangle$, 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 typicality of individuals in Δ with respect to such concepts. Each $<_{C_i}$ is a modular and well-founded strict partial order on Δ, like preferences in KLM rational models.

The preference relations are used to define the meaning of typicality concepts. 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. In the fuzzy case [13], the preference relation $<_C$ of a concept C is induced by the fuzzy interpretation C^I of the concept, a function mapping each domain element in Δ to a value in $[0, 1]$, that is $x <_C y$ iff $C^I(x) > C^I(y)$.

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). Once the SOM has learned to categorize, to assess category generalization, Gliozzi and Plunkett [29] 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 stimulus x from 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 [14, 12], by letting $x <_{C_i} y$ if and only if $d(x, C_i) > d(y, C_i)$. Based on the assumption that the abstraction process in the SOM identifies 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.

The notion of generalization degree introduced by Gliozzi and Plunkett [29] can be used to define a fuzzy multi-preferential interpretation of SOMs. 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 [29].

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 [30].

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 Δ 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 Δ based on the activity values, $y_k(v)$, of neuron k for each input $v \in \Delta$. In this way, a fuzzy multi-preferential interpretation of the network can be constructed over the domain Δ.

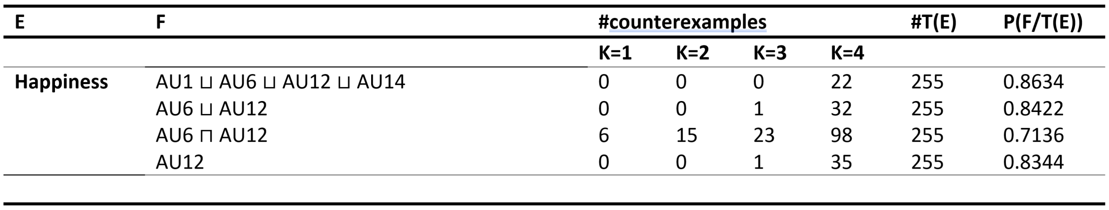
In a fuzzy multi-preferential interpretation, the activation value $y_k(x)$ of neuron k for a stimulus x in the network (assumed to be in the interval $[0, 1]$) is taken to be the degree of membership of x in concept C_k. The interpretation of boolean concepts is defined by fuzzy combination functions, as usual in fuzzy DLs [31, 32]. 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 Δ is finite, as in this case). Let us call $\mathcal{M}^{f,\Delta}_{\mathcal{N}}$ the fuzzy multi-preferential interpretation built from network $\mathcal{N}$ over a domain Δ. 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 as well be of interest.

A Datalog-based approach has been developed [22], which builds a multi-valued preferential interpretation $\mathcal{M}^{f,\Delta}_{\mathcal{N},n}$ of a trained feedforward network $\mathcal{N}$ and, then, verifies the properties of the network for post-hoc explanation. A multi-valued truth space $\mathcal{C}_n = \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}\}$ is considered, for some $n \geq 1$.

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) [33], which involves Action Units (AUs), i.e., facial muscle contractions. From the RAF-DB [34] data set, we selected the subset of the images that were labelled using only one emotion in the set $\{suprise, fear, happiness, anger\}$. A processed dataset containing 5 975 images was input to OpenFace 2.0; the output intensities of AUs were rescaled in order to make their distribution conformant to the expected one in case AUs were recognized by humans [33]. 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 relations between such AUs and emotions, studied by psychologists [35], have been used as a reference for formulae to be verified on neural networks trained to learn such relations. The model checking approach was applied, using the Clingo ASP solver as Datalog engine, taking as set of input stimuli Δ 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 (form 1 to n), 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 [30].

The typicality inclusions relate instances with a high degree of membership in the output class (the one for the output node) with combinations of AU values. In this case, the results can be compared with expectations from domain experts [35]; in general, they can be used to point out knowledge the network has learned, where the attention on *typical* instances of the



E	F	#counterexamples				#T(E)	P(F/T(E))
		K=1	K=2	K=3	K=4		
Happiness	AU1 ⊔ AU6 ⊔ AU12 ⊔ AU14	0	0	0	22	255	0.8634
	AU6 ⊔ AU12	0	0	1	32	255	0.8422
	AU6 ⊓ 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

output class may be useful to concentrate on cases that are far from borderline. The probability measure provides complementary information. Fuzzy DL inclusions may also include fuzzy modifiers (*very*, *slightly*, etc.), which have been considered in fuzzy DLs [36] (e.g., are slightly happy people instances of $au6 \sqcup au12$ with a degree $\geq 2/5$?).

Concerning Table 1, for example, the formula $\mathbf{T}(happiness) \sqsubseteq au1 \sqcup au6 \sqcup au12 \sqcup au14 \geq 3/5$ holds for all individuals, while $\mathbf{T}(happiness) \sqsubseteq au12 \geq 3/5$ (where $au12$ is the activation of the lip corner puller muscle, that is, smiling) 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 $\mathbf{T}(happiness) \sqsubseteq au12 \geq 4/5$.

5. MultiLayer Perceptrons as Weighted conditional knowledge bases

The fuzzy multi-preferential interpretation $\mathcal{M}_{\mathcal{N}}^{f,\Delta}$, built from a network $\mathcal{N}$ for a given set of input stimuli (a domain Δ) as 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 $K^{\mathcal{N}}$ [13].

The weighted conditional knowledge base $K^{\mathcal{N}}$ contains, 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 can 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 $K^{\mathcal{N}}$ based on a fuzzy multipreferential semantics, and specifically based on the notions of *coherent* [13], *faithful* [17] and φ*-coherent* [18, 37] (fuzzy) multi-preferential semantics.

In general a weighted conditional KB $K^{\mathcal{N}}$ [13], besides a set of weighted conditional inclusions, also contains a TBox and an ABox as in standard (and in fuzzy) description logics. Multipreferential semantics for weighted conditional KBs have been defined through a semantic closure construction in the spirit of Lehmann's lexicographic closure [38] and Kern-Isberner's c-representations [39], but adopting a concept-wise approach, so that different preference relations are defined.

Specifically, a coherent multi-preferential model of a weighted KB is defined as a fuzzy interpretation $I = \langle \Delta, \cdot^I \rangle$, which satisfies all DL axioms in TBox and ABox, as well as a coherence condition which requires that each preference relation $<_{C_i}$, induced from the fuzzy interpretation over the domain Δ, is coherent with the the weights $W_i(x)$ of all domain individuals x with respect to concept C_i. For each distinguished concept C_i, and 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: $W_i(x) = \sum_h w_h^i \; D_{i,h}^I(x)$.

For instance, in the φ*-coherence* semantics a function $\varphi_i : \mathbb{R} \rightarrow [0,1]$ is associated to each distinguished concepts C_i. An interpretation $I = \langle \Delta, \cdot^I \rangle$ is φ*-coherent* if, for all concepts $C_i \in \mathcal{C}$ and $x \in \Delta$,

$$C_i^I(x) = \varphi_i(\sum_h w_h^i \; D_{i,h}^I(x))$$

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.

Once a trained neural network can be seen as a weighted defeasible KB $K^{\mathcal{N}}$, φ-entailment can then be used to prove properties of the network for post-hoc explanation. Some preliminary experiments have been done based on finitely many-valued Gödel description logic with typicality $G_n\mathcal{LC}\mathbf{T}$ [21], by defining an ASP encoding of entailment. As a proof of concept, in [21] the entailment approach has been experimented for the weighted $G_n\mathcal{LC}\mathbf{T}$ KBs corresponding to two of the trained multilayer feedforward network for the MONK's problems ([40]).

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 requires advanced solving techniques for reasoning about large networks, which may include state of the art ASP solving, and fuzzy ASP solving [25], as well as other techniques.

6. 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 up 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 [22]).

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 Δ 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. For MLPs, given the associated weighted KB, properties of single individuals can as well be verified through entailment (by instance checking, in DL terminology), and an interesting direction of investigation is the study of counterfactual explanation [49].

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: We thank the anonymous referees for their helpful suggestions. This research was partially supported by the Università del Piemonte Orientale, by the LAIA lab (part of the SILA labs), and by INDAM-GNCS Project 2022 'LESLIE".

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-

ence 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.

[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, 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, Theory Pract. Log. Program. 22 (2022) 589–605. doi: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, 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] M. Gebser, R. Kaminski, B. Kaufmann, M. Ostrowski, T. Schaub, P. Wanko, Theory solving made easy with Clingo 5, in: Technical Commun. of the 32nd International Conference on Logic Programming, ICLP 2016 TCs, October 16-21, 2016, New York City, USA, 2016.

[25] M. Alviano, R. Peñaloza, Fuzzy answer set computation via satisfiability modulo theories, Theory Pract. Log. Program. 15 (2015) 588–603.

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

[27] 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.

[28] 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.

[29] V. Gliozzi, K. Plunkett, Grounding bayesian accounts of numerosity and variability effects in a similarity-based framework: the case of self-organising maps, Cogn. Sci. 31 (2019).

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

[31] G. Stoilos, G. B. Stamou, V. Tzouvaras, J. Z. Pan, I. Horrocks, Fuzzy OWL: uncertainty and the semantic web, in: OWLED*05 Workshop, volume 188 of *CEUR Workshop Proc.*, 2005.

[32] T. Lukasiewicz, U. Straccia, Managing uncertainty and vagueness in description logics for the Semantic Web, J. Web Semant. 6 (2008) 291–308.

[33] P. Ekman, W. Friesen, J. Hager, Facial Action Coding System, Research Nexus, 2002.

[34] 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.

[35] B. Waller, J. C. Jr., A. Burrows, Selection for universal facial emotion, Emotion 8 (2008) 435–439.

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

[37] 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.

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

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

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

[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.

[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.

[49] R. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020, ACM, 2020, pp. 607–617.