

Verifying Properties of a MultiLayer Network for the Recognition of Basic Emotions in a Conditional DL with Typicality (Extended Abstract)

Mario Alviano¹, Francesco Bartoli², Marco Botta², Roberto Esposito², Laura Giordano³ and Daniele Theseider Dupré³

¹DEMACS, Università della Calabria, Via Bucci 30/B, 87036 Rende (CS), Italy

²Dipartimento di Informatica, Università di Torino, Corso Svizzera 185, 10149 Torino, Italy

³DISIT, Università del Piemonte Orientale, Viale Michel 11, 15121 Alessandria, Italy

Abstract

The extended abstract (an abridged version of [1]) reports about our work investigating the relationships between a multi-preferential semantics for defeasible reasoning in knowledge representation and a multilayer neural network model. Weighted knowledge bases for a simple description logic with typicality are considered under a (many-valued) “concept-wise” multipreference semantics. The semantics is used to provide a preferential interpretation of MultiLayer Perceptrons (MLPs). A model checking and an entailment based approach are exploited in the verification of properties of neural networks for the recognition of basic emotions.

Keywords


Description Logics, Preferential and Conditional reasoning, Typicality, Explainability

Preferential approaches to commonsense reasoning (e.g., [2, 3, 4, 5, 6, 7, 8, 9]) have their roots in conditional logics [10, 11], and have been 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 [12, 13, 14] and closure constructions (e.g., [15, 16, 17, 18, 19]) have been proposed for defeasible DLs. Among these, the concept-wise multi-preferential semantics, which allows to account for preferences with respect to different concepts. It has been first introduced as a semantics of ranked \mathcal{EL}^\perp knowledge bases (KBs) [20], and then for weighted conditional DL knowledge bases [21], and has been proposed as a semantics for the post-hoc verification of some neural network models [22, 1].


The idea underlying the multi-preferential semantics is that, for two domain elements *spirit* and *buddy* and two concepts, e.g., *Horse* and *Zebra*, *spirit* might be more typical than *buddy* as a horse ($spirit <_{Horse} buddy$), while less typical than *buddy* as a zebra ($buddy <_{Zebra} spirit$).

As for the DLs with typicality based on a single preference relation (e.g., [14, 17]), a typicality operator **T** is introduced in the language to single out the typical instances **T**(*C*) of a concept *C*. 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) preference relations $<_{C_1}, \dots, <_{C_n}$ for a set of distinguished concepts C_1, \dots, C_n , to represent the typicality of individuals in Δ relative to the concepts. Each $<_{C_i}$ is assumed to be a modular and well-founded strict partial order on Δ , like preferences in Kraus, Lehmann and Magidor’s (KLM) ranked models [4].

The preference relations are used to define the meaning of typicality concepts. In the two-valued case, the concept **T**(*C_i*) is interpreted as the set of all $<_{C_i}$ -minimal elements (*x* is an instance of **T**(*Horse*) if there is no other instance of *Horse* preferred to *x* with respect to $<_{Horse}$). The typicality operator cannot be nested and it allows to define defeasible inclusions (typicality inclusions) of the form **T**(*C*) \sqsubseteq *D*, whose intended meaning is that “the typical *C*’s are *D*’s” or “normally *C*’s are *D*’s”, which correspond to KLM conditionals [4]. Defeasible inclusions, unlike strict inclusions $C \sqsubseteq D$, may

 DL 2024: 37th International Workshop on Description Logics, June 18–21, 2024, Bergen, Norway

 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); dtd@uniupo.it (D. Theseider Dupré)

 0000-0002-2052-2063 (M. Alviano); 0000-0001-9445-7770 (L. Giordano)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

have exceptions.

In *weighted defeasible knowledge bases* (KBs) typicality inclusions come with a weight. A concept C_i can be associated with a set of typicality inclusions (conditionals) of the form $\mathbf{T}(C_i) \sqsubseteq D_{j,i}$, with a weight w_{ij} , representing the *prototypical properties* of concept C_i . The weight w_{ij} is a real number representing the plausibility or implausibility of the property $D_{j,i}$ for members of C_i . For instance, one may want to represent a situation in which horses are normally tall and run fast, it is very plausible that they have a tail, but implausible that they have stripes. In a weighted KB these defeasible properties of horses may be represented as:

$$\begin{aligned} \mathbf{T}(\text{Horse}) &\sqsubseteq \text{Tall}, 4.5 & \mathbf{T}(\text{Horse}) &\sqsubseteq \text{RunFast}, 4.2 \\ \mathbf{T}(\text{Horse}) &\sqsubseteq \exists \text{has_Tail}.\top, 9.7 & \mathbf{T}(\text{Horse}) &\sqsubseteq \exists \text{has_Stripes}.\top, -20 \end{aligned}$$

where negative weights represent implausible properties. The defeasible Tbox above can be used to define an ordering among domain elements, comparing their typicality as horses. For instance, assuming that Spirit is tall, has tail, no stripes and does not run fast, while Buddy is tall, has tail, runs fast and has stripes, we can expect that $\text{spirit} <_{\text{Horse}} \text{buddy}$. In our approach such features (such as, being tall or having a tail) are as well represented as concepts in the DL.

In the two valued case, the preference relations $<_{C_i}$ can be constructed from the KB by defining the weight $W_{C_i}(x)$ of a domain element x with respect to a concept C_i , by summing up the weights of the typicality inclusions for C_i satisfied by x . The preference relations are then induced from such weights as: $x <_{C_i} y$ iff $W_{C_i}(x) > W_{C_i}(y)$. In the example: Spirit satisfies the first and the third default, hence $W_{\text{Horse}}(\text{spirit}) = 14.2$, while Buddy satisfies all the defaults, hence, $W_{\text{Horse}}(\text{buddy}) = -1.6$. As $W_{\text{Horse}}(\text{spirit}) > W_{\text{Horse}}(\text{buddy})$ then $\text{spirit} <_{\text{Horse}} \text{buddy}$. The semantic construction is in the spirit of other semantics for conditionals [23, 9, 24], but it adopts multiple preferences.

Note that the interpretation of a typicality concept $\mathbf{T}(C)$, for an arbitrary C (e.g., $\mathbf{T}(\text{Student} \sqcap \text{Employee})$) would require the definition of a preference $<_C$ for each C , or the definition of a global preference relation $<$. In [20], e.g., a global preference $<$ is defined based on a (modified) Pareto-combination of preferences $<_{C_i}$. An alternative route is to move to a fuzzy interpretation of concepts, and define $<_C$ based on the fuzzy interpretation of C .

Fuzzy and many-valued DLs are well studied in the literature (see, for instance, [25, 26, 27, 28, 29]). In fuzzy DLs, the idea is that a concept C is interpreted as a function $C^I : \Delta \rightarrow [0, 1]$ mapping each domain element to a value in the unit interval $[0, 1]$. Then, for a domain element $x \in \Delta$, $C^I(x)$ is regarded the *degree of membership* of x in concept C . In the fuzzy case [21, 1], the preference relation $<_C$ of any concept C is induced by the fuzzy interpretation C^I of concept C : $x <_C y$ iff $C^I(x) > C^I(y)$. In a *non-crisp interpretation of typicality* [1], the fuzzy interpretation of typicality concepts $\mathbf{T}(C)$ in an interpretation I is defined as: $(\mathbf{T}(C))^I(x) = C^I(x)$, if there is no $y \in \Delta$ such that $y <_C x$; $(\mathbf{T}(C))^I(x) = 0$, otherwise. This choice has some impact on the (KLM) properties of entailment. When $(\mathbf{T}(C))^I(x) > 0$, we say that x is a typical C -element in I (and all typical C -elements have the same membership degree in C).

As in the two-valued case, besides usual fuzzy DL axioms, a weighted KB includes a defeasible TBox, a set of weighted typicality inclusions $\mathbf{T}(C_i) \sqsubseteq D_{j,i}$, with weight w_{ij} , for each distinguished concept C_i . The definition of $W_{C_i}(x)$ in a fuzzy interpretation I is defined by considering the degree to which x satisfies the properties (being tall, running fast, etc.). The *weight* $W_{C_i}(x)$ of x wrt C_i in an interpretation $I = \langle \Delta, \cdot^I \rangle$ is defined as follows: $W_{C_i}(x) = \sum_h w_{ih} D_{i,h}^I(x)$, if $C_i^I(x) > 0$; $W_{C_i}(x) = -\infty$, otherwise.

The models of a KB are required to satisfy further properties beyond satisfying fuzzy DL axioms [30], by enforcing that the membership degree $C^I(x)$ of x in C is aligned with the weight $W_{C_i}(x)$ in I . For instance, in *coherent models* [21] of a KB, we require that $x <_{C_i} y$ iff $W_{C_i}(x) > W_{C_i}(y)$. *Faithful models* [31] exploit a slightly weaker condition, while the stronger notion of φ -coherence of a fuzzy interpretation I wrt a KB exploits a monotonically non-decreasing function $\varphi : \mathbb{R} \rightarrow [0, 1]$. I is φ -coherent with respect to a weighted KB if: for all $C_i \in \mathcal{C}$ and $x \in \Delta$, $C_i^I(x) = \varphi(W_{C_i}(x))$.

A mapping of a multilayer network to a conditional KB can be defined in a simple way [21, 1], by 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_{ih} . If we assume that φ is the *activation function* of all units in the network (having value in the unit interval

$[0, 1]$), then the φ -coherent semantics characterizes unit activation: $C_i^I(x)$ corresponds to the activation of unit i for some input stimulus x . The semantics can also consider multiple functions φ_i to represent the activation functions of different units. φ -coherent interpretations capture the *stationary states* of the network, both for MLPs and for recurrent networks, which allow for feedback cycles (a weighted KB can indeed have cycles).

Since a multilayer network can be regarded as a conditional KB, *entailment* in the conditional logic can be used for the verification of *conditional properties* of the network for *post-hoc verification*. Undecidability results for fuzzy DLs with general inclusion axioms [32, 29] have led to considering a *finitely-valued version of φ -coherent semantics*, which provides an *approximation* of the fuzzy semantics [1], by taking $\mathcal{C}_n = \{0, \frac{1}{n}, \dots, \frac{n-1}{n}, 1\}$, for $n \geq 1$, as the truth space. For the boolean fragment, in the finitely-valued case, an ASP-based approach has been proposed for defeasible reasoning under φ -coherent entailment [33]. Complexity results have been investigated, as well as the scalability of different encodings of entailment in ASP, by taking advantage of custom propagators, weak constraints and weight constraints [34].

In [1] we consider both the *entailment based* approach and a *model checking* approach in the verification of conditional properties of some trained multilayer feedforward networks for the recognition of basic emotions, using the Facial Action Coding System (FACS) [35] and the RAF-DB [36] data set, containing almost 30000 images labeled with basic emotions or combinations of two emotions. The images were input to OpenFace 2.0 [37], which detects a subset of the Action Units (AUs) in [35], corresponding to facial muscle contractions; The AUs were used as input layer of an MLP, trained to recognize four emotions. The relations between such AUs and emotions, studied by psychologists [38], have been used as a reference for formulae to be verified.

The model checking approach exploits the behavior of the network \mathcal{N} over a set Δ of input exemplars (e.g., the test set), to construct a single multi-preferential interpretation $I_{\mathcal{N}}$ with domain Δ , considering only some units of interest (e.g., input and output units). For such units i , the associated concept C_i is interpreted by letting $C_i^{I_{\mathcal{N}}}(x)$ be the activity of unit i for input x . Graded conditional properties of the form $\mathbf{T}(E) \sqsubseteq F \geq l$ (as well as strict properties $E \sqsubseteq F \geq l$) can then be checked in $I_{\mathcal{N}}$. Verifying the satisfiability of an inclusion in the interpretation $I_{\mathcal{N}}$ requires polynomial time in the size of $I_{\mathcal{N}}$ and of the formula.

The entailment based approach has been experimented for a binary classification task, for the class *happiness* vs other emotions. A set of 8 835 images was used. The OpenFace output intensities were rescaled in order to make their distribution conformant to the expected one in case AUs are recognized by humans [35]. The resulting 17 AUs were used as input units of a fully connected feed forward NN, with two hidden layers of 50 and 25 nodes, using the logistic activation function for all layers. The F1 score of the trained network was 0.831. Verification has been performed taking \mathcal{C}_5 as the truth value space (given that a scale of five values, plus absence, is used by humans for AU intensities), and using minimum t-norm, the associated t-conorm, and standard involutive negation. With truth space \mathcal{C}_5 and 17 AUs as input units, the size of the search space for the solver was 6^{17} , i.e., more than 10^{13} . The weighted conditional knowledge base associated to the network contains 2 201 weighted typicality inclusions. The version of the solver in [34] based on weight constraints and order encoding was used.

Let us consider the two graded inclusion axioms: (a) $\mathbf{T}(\text{happiness}) \sqsubseteq \text{au1} \sqcup \text{au6} \sqcup \text{au12} \sqcup \text{au14} \geq k/5$ and (b) $\mathbf{T}(\text{happiness}) \sqsubseteq \text{au6} \sqcup \text{au12} \geq k/5$. The model checking approach, applied to the test set (2 651 individuals with 390 instances of $\mathbf{T}(\text{happiness})$), finds that both formulae hold for $k = 3$ and do not hold for $k = 4$.

In the entailment approach, the solver finds in seconds that (a) is not entailed for $k = 4$, and in minutes that it is entailed for $k = 1$, while for $k = 2, 3$, it does not provide a result in hours. On a variant of the experiment, using as inputs AU intensities that are not rescaled, the solver finds in seconds that (a) is not entailed for $k = 2$, and in minutes that it is entailed for $k = 1$. The graded inclusion axiom (b) is entailed for $k = 1$ and not for $k = 3$. In the latter case, then, a counterexample is found by entailment, whose search space includes all possible combinations of input vectors, while it is not found by model checking on the test set. The co-existence of strict and defeasible inclusions in weighted KBs also allows for combining empirical knowledge with elicited knowledge for reasoning and for post-hoc

verification. A different experiment in the verification of properties of a network trained to classify its input as an instance of four emotions *surprise*, *fear*, *happiness*, *anger*, is also reported in [1].

While the model-checking approach does not require to consider the activity of all units to build a preferential interpretation of a network, in the entailment-based approach all units are considered. Also, the model-checking approach, based on the conditional multi-preferential semantics, is a general (*model agnostic*) approach, which may be suitable to explain different network models (and was first considered for SOMs [22]). On the other hand, the entailment-based approach is specific for MLPs. Both approaches are *global* ones (see, e.g., [39]), as they consider the behavior of the network over a set Δ of input stimuli. We refer to [1] for detailed results, discussion and related work on this conditional approach to explainability.

Acknowledgments

The work was partially supported by the INDAM-GNCS Project 2024 “LCXAI: Logica Computazionale per eXplainable Artificial Intelligence”.

References

- [1] M. Alviano, F. Bartoli, M. Botta, R. Esposito, L. Giordano, D. Theseider Dupré, A preferential interpretation of multilayer perceptrons in a conditional logic with typicality, *Int. Journal of Approximate Reasoning* 164 (2024). URL: <https://doi.org/10.1016/j.ijar.2023.109065>.
- [2] J. Delgrande, A first-order conditional logic for prototypical properties, *Artificial Intelligence* 33 (1987) 105–130.
- [3] D. Makinson, General theory of cumulative inference, in: *Non-Monotonic Reasoning*, 2nd International Workshop, Grassau, FRG, June 13-15, 1988, Proceedings, 1988, pp. 1–18.
- [4] S. Kraus, D. Lehmann, M. Magidor, Nonmonotonic reasoning, preferential models and cumulative logics, *Artificial Intelligence* 44 (1990) 167–207.
- [5] J. Pearl, System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning, in: *TARK’90*, Pacific Grove, CA, USA, 1990, pp. 121–135.
- [6] D. Lehmann, M. Magidor, What does a conditional knowledge base entail?, *Artificial Intelligence* 55 (1992) 1–60.
- [7] S. Benferhat, C. Cayrol, D. Dubois, J. Lang, H. Prade, Inconsistency management and prioritized syntax-based entailment, in: *Proc. IJCAI’93*, Chambéry, 1993, pp. 640–647.
- [8] R. Booth, J. B. Paris, A note on the rational closure of knowledge bases with both positive and negative knowledge, *Journal of Logic, Language and Information* 7 (1998) 165–190. doi:10.1023/A:1008261123028.
- [9] G. Kern-Isberner, Conditionals in Nonmonotonic Reasoning and Belief Revision - Considering Conditionals as Agents, volume 2087 of *LNCS*, Springer, 2001.
- [10] D. Lewis, *Counterfactuals*, Basil Blackwell Ltd, 1973.
- [11] D. Nute, *Topics in conditional logic*, Reidel, Dordrecht (1980).
- [12] 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.
- [13] 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.
- [14] L. Giordano, V. Gliozzi, N. Olivetti, G. L. Pozzato, ALC+T: a preferential extension of Description Logics, *Fundamenta Informaticae* 96 (2009) 1–32.
- [15] 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.
- [16] 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.

- [17] 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.
- [18] M. Pensel, A. Turhan, Reasoning in the defeasible description logic EL_{\perp} - computing standard inferences under rational and relevant semantics, *Int. J. Approx. Reasoning* 103 (2018) 28–70.
- [19] 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.
- [20] L. Giordano, D. Theseider Dupré, An ASP approach for reasoning in a concept-aware multipreferential lightweight DL, *TPLP* 10(5) (2020) 751–766.
- [21] 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.
- [22] L. Giordano, V. Gliozzi, D. Theseider Dupré, A conditional, a fuzzy and a probabilistic interpretation of self-organizing maps, *J. Log. Comput.* 32 (2022) 178–205.
- [23] D. J. Lehmann, Another perspective on default reasoning, *Ann. Math. Artif. Intell.* 15 (1995) 61–82.
- [24] E. Weydert, System JLZ - rational default reasoning by minimal ranking constructions, *Journal of Applied Logic* 1 (2003) 273–308.
- [25] U. Straccia, Towards a fuzzy description logic for the semantic web (preliminary report), in: *ESWC 2005*, Heraklion, Crete, May 29 - June 1, 2005, volume 3532 of *LNCS*, Springer, 2005, pp. 167–181.
- [26] 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.
- [27] T. Lukasiewicz, U. Straccia, Description logic programs under probabilistic uncertainty and fuzzy vagueness, *Int. J. Approx. Reason.* 50 (2009) 837–853.
- [28] A. García-Cerdaña, E. Armengol, F. Esteva, Fuzzy description logics and t-norm based fuzzy logics, *Int. J. Approx. Reason.* 51 (2010) 632–655.
- [29] 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.
- [30] T. Lukasiewicz, U. Straccia, Managing uncertainty and vagueness in description logics for the Semantic Web, *J. Web Semant.* 6 (2008) 291–308.
- [31] 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.
- [32] M. Cerami, U. Straccia, On the (un)decidability of fuzzy description logics under Łukasiewicz t-norm, *Inf. Sci.* 227 (2013) 1–21. URL: <https://doi.org/10.1016/j.ins.2012.11.019>.
- [33] 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.
- [34] M. Alviano, L. Giordano, D. Theseider Dupré, Complexity and scalability of defeasible reasoning in many-valued weighted knowledge bases, in: *Logics in Artificial Intelligence - 18th European Conference, JELIA 2023*, Dresden, Germany, September 20-22, 2023, *Proceedings*, volume 14281 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 481–497. doi:10.1007/978-3-031-43619-2_33.
- [35] P. Ekman, W. Friesen, J. Hager, Facial Action Coding System, Research Nexus, 2002.
- [36] 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.
- [37] 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.
- [38] B. Waller, J. C. Jr., A. Burrows, Selection for universal facial emotion, *Emotion* 8 (2008) 435–439.
- [39] 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.