Model Checking Verification of MultiLayer Perceptrons in Datalog: a Many-valued Approach with Typicality

Francesco Bartoli[1], Marco Botta[1], Roberto Esposito[1], Laura Giordano[2] and Daniele Theseider Dupré[2]

[1]Dipartimento di Informatica, Università di Torino, Italy

[2]DISIT - Università del Piemonte Orientale, Alessandria, Italy

Abstract

Description logics with typicality have been considered under a "concept-wise" multi-preferential semantics as the basis of a logical interpretation of MultiLayer Perceptrons (MLPs). In this paper we exploit a Datalog-based approach to prove logical properties of a trained network by model checking, starting from its input/output behavior, building a many-valued preferential model for the verification of typicality properties. The model is also used for providing a probabilistic account of MLPs, exploiting typicality concepts and Zadeh's probability of fuzzy events. We report about some experiments to the verification of properties of neural networks for the recognition of basic emotions. This work is a step in the direction of verifying and interpreting knowledge learned by a neural network, to achieve a trustworthy and explainable AI.

Keywords

Description Logic, Typicality, Neural Networks, Explainability

1. Introduction

Preferential approaches to common sense reasoning [1, 2, 3, 4, 5], have been extended to description logics (DLs) to deal with inheritance with exceptions in ontologies, by allowing for non-strict inclusions, called *typicality or defeasible inclusions*, with different preferential semantics, e.g., [6, 7] and [8, 9], and closure constructions, e.g, [10, 11, 12, 9].

In recent work, a concept-wise multipreference semantics has been proposed [13] as a semantics for ranked Description Logic (DL) knowledge bases (KBs), i.e. knowledge bases in which defeasible or typicality inclusions of the form $\mathbf{T}(C) \sqsubseteq D$ (meaning "the typical C's are D's" or "normally C's are D's"), stemming from KLM conditionals [1, 3], are given a rank,

Datalog 2.0 2022: 4th International Workshop on the Resurgence of Datalog in Academia and Industry, September 05, 2022, Genova - Nervi, Italy

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)

http://informatica.unito.it/persone/marco.botta/ (M. Botta); http://informatica.unito.it/persone/roberto.esposito (R. Esposito); http://people.unipmn.it/laura.giordano/ (L. Giordano); 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)

representing their strength, where $\mathbf{T}$ is a *typicality operator* [6], that singles out the typical instances of concept C.

The multi-preferential semantics has been extended [14] to weighted knowledge bases, in which typicality inclusions have a real (positive or negative) weight, representing plausibility or implausibility. The semantics has been exploited to provide a preferential interpretation to Multilayer Perceptrons (MLPs, [15]), an approach previously considered [16, 17] for self-organising maps (SOMs, [18]). In both cases, considering the domain of all input stimuli presented to the network during training (or in the generalization phase), one can build a semantic interpretation of the network as a multi-preferential interpretation, where preferences are associated to concepts. This allows properties of a neural network to be verified by *model checking* over a fuzzy preferential interpretation. A MLP can as well be regarded as a weighted conditional knowledge base [14] (based on a fuzzy concept-wise preferential semantics) by interpreting synaptic connections as conditional implications. Specifically, the notions of *coherent* [14], *faithful* [19] and φ*-coherent* [20] models of a weighted $\mathcal{ALC}$ knowledge base have been considered for fuzzy $\mathcal{ALC}$ with typicality.

In previous work, proof methods for reasoning with weighted conditional KBs have been studied, in the two-valued case [21] for $\mathcal{EL}^{\perp}$ KBs, and in the finitely many-valued case [22], providing an approximation of fuzzy φ-coherent entailment for the boolean fragment. Finitely many-valued DLs are well-studied in the literature [23, 24, 25, 26]. In particular, for the boolean fragment $\mathcal{LC}$ of $\mathcal{ALC}$ (which does not contain roles, and then neither universal nor existential restrictions), the finitely many-valued Gödel and Łukasiewicz description logics, $G_n\mathcal{LC}$ and Ł$_n\mathcal{LC}$, have been extended with a typicality operator, and a semantic closure construction, based on φ_n-coherent interpretations, has been introduced to deal with weighted KBs. ASP and *asprin* [27] have been exploited for deciding φ_n-coherent entailment, a many-valued approximation of φ-coherence entailment, and the ASP encoding is used to prove that the problem is in Π_2^p [22].

In this paper, we investigate a Datalog-based approach to model checking for verifying the logical properties of a neural network, by constructing a preferential interpretation of the network starting from its input/output behavior over a set of input stimuli. We exploit the activations of units for those stimuli, to define a many-valued interpretation of the concepts associated with those units (namely, the units of interest for verification).

More specifically, we exploit Datalog with weakly stratified negation [28] in the construction of the model of a neural network $\mathcal{N}$ over a given domain Δ of input stimuli. This allows the verification in polynomial time of typicality properties of the form $\mathbf{T}(C) \sqsubseteq D\theta\alpha$, for $\theta \in \{\geq, \leq, >, <\}$ and $\alpha \in [0, 1]$, as well as to evaluate the conditional probabilities $P(D|C)$, based on Zadeh's probability of fuzzy events [29], also including occurrences of typicality concepts $\mathbf{T}(C)$.

We conclude the paper by reporting about some experiments to the verification of properties of neural networks for the recognition of basic emotions using the Facial Action Coding System (FACS) [30].

2. A fuzzy and a finitely many-valued description logic

Fuzzy description logics have been widely studied in the literature for representing vagueness in DLs [31, 32, 33], based on the idea that concepts and roles can be interpreted as fuzzy sets and fuzzy relations. In fuzzy logic, formulas have a truth degree from a truth space $\mathcal{S}$, usually $[0, 1]$, as in Mathematical Fuzzy Logic [34] or $\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}\}$, for an integer $n \geq 1$. The finitely many-valued case is also well studied for DLs [23, 24, 25, 26]; in the following, we will also consider a *finitely many-valued* extension of the boolean fragment of $\mathcal{ALC}$ with typicality.

Let $\mathcal{LC}$ be the fragment of $\mathcal{ALC}$ with no roles, N_C be a set of concept names and N_I a set of individual names. The set of $\mathcal{LC}$ *concepts* can be defined inductively as follows: (i) $A \in N_C$, $\top$ and $\bot$ are concepts; (ii) if C and D are concepts, then $C \sqcap D$, $C \sqcup D$, $\neg C$ are concepts.

A *fuzzy interpretation* for $\mathcal{LC}$ is a pair $I = \langle \Delta, \cdot^I \rangle$ where Δ is a non-empty domain and $\cdot^I$ is *fuzzy interpretation function* that assigns to each concept name $A \in N_C$ a function $A^I : \Delta \rightarrow [0, 1]$, and to each individual name $a \in N_I$ an element $a^I \in \Delta$. A domain element $x \in \Delta$ belongs to the extension of A to some degree in $[0, 1]$, i.e., A^I is a fuzzy set.

The interpretation function $\cdot^I$ is extended to complex concepts as follows:

$$\top^I(x) = 1 \qquad \bot^I(x) = 0 \qquad (\neg C)^I(x) = \ominus C^I(x)$$
$$(C \sqcap D)^I(x) = C^I(x) \otimes D^I(x)$$
$$(C \sqcup D)^I(x) = C^I(x) \oplus D^I(x)$$

where $x \in \Delta$ and $\otimes$, $\oplus$, $\rhd$ and $\ominus$ are arbitrary but fixed t-norm, s-norm, implication function, and negation function, chosen among the combination functions of various fuzzy logics (we refer to [32] for details). In particular, in Gödel logic $a \otimes b = min\{a, b\}$, $a \oplus b = max\{a, b\}$, $a \rhd b = 1$ *if* $a \leq b$ *and* b *otherwise*; $\ominus a = 1$ *if* $a = 0$ *and* 0 *otherwise.* In Łukasiewicz logic, $a \otimes b = max\{a + b - 1, 0\}$, $a \oplus b = min\{a + b, 1\}$, $a \rhd b = min\{1 - a + b, 1\}$ and $\ominus a = 1 - a$.

The interpretation function $\cdot^I$ is also extended to non-fuzzy axioms (i.e., to strict inclusions and assertions of an $\mathcal{LC}$ knowledge base) as follows:

$$(C \sqsubseteq D)^I = inf_{x \in \Delta} C^I(x) \rhd D^I(x)$$
$$(C(a))^I = C^I(a^I)$$

A *fuzzy* $\mathcal{LC}$ *knowledge base* K is a pair $(\mathcal{T}, \mathcal{A})$ where $\mathcal{T}$ is a fuzzy TBox and $\mathcal{A}$ a fuzzy ABox. A fuzzy TBox is a set of *fuzzy concept inclusions* of the form $C \sqsubseteq D\ \theta\ \alpha$, where $C \sqsubseteq D$ is an $\mathcal{LC}$ concept inclusion axiom, $\theta \in \{\geq, \leq, >, <\}$ and $\alpha \in [0, 1]$. A fuzzy ABox $\mathcal{A}$ is a set of *fuzzy assertions* of the form $C(a)\ \theta\alpha$ where C is an $\mathcal{LC}$ concept, $a \in N_I$, $\theta \in \{\geq, \leq, >, <\}$ and $\alpha \in [0, 1]$. Following Bobillo and Straccia [35], we assume that fuzzy interpretations are *witnessed*, i.e., the sup and inf are attained at some point of the involved domain.

Definition 1 (Satisfiability and entailment). *A fuzzy interpretation I satisfies a fuzzy $\mathcal{LC}$ axiom E (denoted $I \models E$), as follows:*
- I satisfies a fuzzy $\mathcal{LC}$ inclusion axiom $C \sqsubseteq D\ \theta\ \alpha$ if $(C \sqsubseteq D)^I \theta\ \alpha$;
- I satisfies a fuzzy $\mathcal{LC}$ assertion $C(a)\theta\alpha$ if $C^I(a^I)\theta\ \alpha$,
for $\theta \in \{\geq, \leq, >, <\}$.

Given a fuzzy KB $K = (\mathcal{T}, \mathcal{A})$, a fuzzy interpretation I satisfies $\mathcal{T}$ (resp. $\mathcal{A}$) if I satisfies all fuzzy inclusions in $\mathcal{T}$ (resp. all fuzzy assertions in $\mathcal{A}$). A fuzzy interpretation I is a model *of K if I satisfies $\mathcal{T}$ and $\mathcal{A}$. A fuzzy axiom E is entailed by a fuzzy knowledge base K, written $K \models E$, if for all models $I = \langle\Delta, \cdot^I\rangle$ of K, I satisfies E.*

In the finitely many-valued case, we assume the truth space to be $\mathcal{C}_n = \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}\}$, for an integer $n \geq 1$. A *finitely many-valued interpretation* for $\mathcal{LC}$ is a pair $I = \langle\Delta, \cdot^I\rangle$ where: Δ is a non-empty domain and $\cdot^I$ is an *interpretation function* that assigns to each $a \in N_I$ a value $a^I \in \Delta$, and to each $A \in N_C$ a function $A^I : \Delta \rightarrow \mathcal{C}_n$. In particular, in [22] we have considered two finitely many-valued cases based on $\mathcal{ALC}$, the finitely many-valued Łukasiewicz description logic Ł$_n\mathcal{ALC}$ and the finitely many-valued Gödel description logic $G_n\mathcal{ALC}$, extended with a standard involutive negation $\ominus a = 1 - a$. Such logics are defined along the lines of the finitely many-valued Łukasiewicz description logic $\mathcal{SROIQ}$ [24], the fuzzy extension of the descrption logic $\mathcal{SROIQ}$ that joins Gödel and Zadeh fuzzy logics (called $GZ\,\mathcal{SROIQ}$) [25], and the logic $\mathcal{ALC}^*(\mathcal{S})$ [23]. In the following we will focus on the $\mathcal{LC}$ fragment $G_n\mathcal{LC}$ of $G_n\mathcal{ALC}$. For $G_n\mathcal{LC}$ the interpretation function $\cdot^I$ is extended to complex concepts and fuzzy axioms as above, and we assume the interpretation of negated concepts exploits involutive negation, i.e., $(\neg C)^I(x) = \ominus C^I(x) = 1 - C^I(x)$. The notions of knowledge base, satisfiability and entailment are defined as above.

3. Fuzzy $\mathcal{LC}$ with typicality and φ-coherent models

Let us consider now fuzzy $\mathcal{LC}$ with typicality $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$, following the approach for $\mathcal{ALC}$ in [14, 19], as well as the finite many-valued case. The idea is similar to the extension of $\mathcal{ALC}$ with typicality in the two-valued case [6], but the degree of membership of domain individuals in a concept C is used to identify the typical elements of C. The extension allows for the definition of *typicality concepts* of the form $\mathbf{T}(C)$, corresponding to the set of most typical C-elements.

Note that, in a fuzzy interpretation $I = \langle\Delta, \cdot^I\rangle$, the degree of membership $C^I(x)$ of x in a concept C induces a preference relation $<_C$ on Δ:

$$x <_C y \quad \text{iff } \mathrm{C^I(x) > C^I(y)}$$

For a witnessed fuzzy $\mathcal{LC}$ interpretation I, each preference relation $<_C$ has the properties of preference relations in KLM-style ranked interpretations [3], that is, $<_C$ is a modular and well-founded strict partial order. Similarly for a finitely many-valued $\mathcal{LC}$ interpretation I. Each relation $<_C$ has the properties of a preference relation in KLM rational interpretations, also called ranked interpretations. It captures the relative typicality of domain elements wrt concept C and may then be used to identify the *typical C-elements*. Let $C^I_{>0} = \{x \in \Delta \mid C^I(x) > 0\}$. One can provide a (crisp) interpretation of typicality concepts $\mathbf{T}(C)$ in an interpretation I as follows:

$$(\mathbf{T}(C))^I(x) = \begin{cases} 1 & \text{if } x \in min_{<_C}(C^I_{>0}) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $min_<(S) = \{u : u \in S \text{ and } \nexists z \in S \text{ s.t. } z < u\}$. When $(\mathbf{T}(C))^I(x) = 1$, x is said to be a typical C-element in I. Let us denote with $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ the extension of fuzzy $\mathcal{LC}$ with typicality, and with $G_n\mathcal{LC}\mathbf{T}$ the extension of $G_n\mathcal{LC}$ with typicality.

Definition 2 ($\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ interpretation). *A $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ interpretation $I = \langle \Delta, \cdot^I \rangle$ is a fuzzy $\mathcal{LC}$ interpretation, extended by interpreting typicality concepts as in (1).*

In a similar way, we can define a $G_n\mathcal{LC}\mathbf{T}$ interpretation: a many-valued interpretation $I = \langle \Delta, \cdot^I \rangle$ implicitly defines a multi-preferential interpretation, where any concept C is associated to a preference relation $<_C$. The notions of *model* of an $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ (resp., a $G_n\mathcal{LC}\mathbf{T}$) KB, and of $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ (resp., $G_n\mathcal{LC}\mathbf{T}$) *entailment* are defined similarly as for fuzzy $\mathcal{LC}$ knowledge bases (see Section 2).

In [14, 19] a notion of *weighted $\mathcal{ALC}^{\mathbf{F}}\mathbf{T}$ knowledge base* has been considered (and similarly for the boolean fragment and the many-valued case [22]), as a tuple $\langle \mathcal{T}, \mathcal{T}_{C_1}, \ldots, \mathcal{T}_{C_k}, \mathcal{A} \rangle$, where $\mathcal{T}$ is a set of inclusion axioms, $\mathcal{A}$ is a set of assertions and $\mathcal{T}_{C_i} = \{(d^i_h, w^i_h)\}$ is a set of all weighted typicality inclusions $d^i_h = \mathbf{T}(C_i) \sqsubseteq D_{i,h}$ for C_i, indexed by h, where each inclusion d^i_h has weight w^i_h, a real number, and C_i and $D_{i,h}$ are $\mathcal{LC}$ concepts.

Some different fuzzy semantics (a coherent [14], a faithful [19] and a φ-coherent semantics [20]) have been considered for weighted knowledge bases, and have been exploited to provide a semantic characterization of multilayer perceptrons as weighted knowledge bases. Based on a notion of φ-coherent entailment and its approximation to the finitely many-valued case, an ASP based approach has been proposed for the verification of typicality properties of a weighted conditional knowledge bases [22] in the logics $G_n\mathcal{LC}\mathbf{T}$ and $\text{Ł}_n\mathcal{LC}\mathbf{T}$. More precisely, an algorithm for deciding entailment of typicality inclusions $\mathbf{T}(C) \sqsubseteq D \; \theta \; \alpha$ from a weighted knowledge base in $G_n\mathcal{LC}\mathbf{T}$ (or in $\text{Ł}_n\mathcal{LC}\mathbf{T}$) has been developed by exploiting an ASP encoding and the *asprin* framework for answer set preferences [27].

In this paper, we consider a single interpretation which can be built over the domain of input stimuli, by exploiting the activity of units, for the different inputs. We will see that such a model can be constructed using Datalog with negation and that the Datalog program can be used for proving properties of the network by model checking.

4. A fuzzy preferential model of a network $\mathcal{N}$

The idea from [14] is that a fuzzy multi-preferential interpretation can be associated to a network $\mathcal{N}$, based on the activity of the network over a set of input stimuli Δ. Fuzzy and typicality properties of the network can then be verified by model checking over such an interpretation, and used for post-hoc explanation.

Here, we consider a trained feedforward network $\mathcal{N}$, and associate a concept name $C_i \in N_C$ to the *units of interest* i in $\mathcal{N}$ for property verification. They may include input, output or hidden units. We construct a multi-preference interpretation over a (finite) *domain* Δ of input stimuli. For instance, the input vectors considered for training and/or generalization, or a subset of it. We assume the activation of units to be in the interval $[0, 1]$.

Assume Δ is a finite set. Following [14], we associate to $\mathcal{N}$ and Δ a fuzzy multi-preferential interpretation as follows.

Definition 3. *The* fuzzy multi-preferential interpretation of a network $\mathcal{N}$ *over the domain Δ, is the $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ interpretation $I^{\Delta}_{\mathcal{N}} = \langle \Delta, \cdot^I \rangle$ where the interpretation function $\cdot^I$ satisfies condition*

$C_k^I(x) = y_k(x)$, for all concept names $C_k \in N_C$ and $x \in \Delta$, where $y_k(x)$ is the output signal of unit k, for input vector x.

As we have seen above, the $\mathcal{LC}^{\mathbf{F}}\mathbf{T}$ interpretation $I_{\mathcal{N}}^{\Delta}$ is a multi-preferential interpretation, as the fuzzy interpretation of concepts induces a preference relation associated to each concept, i.e., to each unit. It has been proven that this interpretation is actually a model of the network [14], when the network is regarded as a weighted knowledge base, and under some conditions on the activation functions of units. It allows the set of typical instances of a concept C_k to be identified in the obvious way, by selecting the input stimuli $x \in \Delta$ with the highest activity values $y_k(x)$, for unit k. For instance, according to the semantics of typicality concepts, the verification of an inclusion $\mathbf{T}(C_h) \sqsubseteq D \geq \alpha$ over model $I_{\mathcal{N}}^{\Delta}$ would require to identify typical C_h-elements and to check whether their membership degree in concept D is greater or equal than α, according to the choice of the t-norm, s-norm, and negation functions.

In the next section we propose a Datalog-based approach to construct the many-valued approximation $I_{\mathcal{N},n}^{\Delta}$ of model $I_{\mathcal{N}}^{\Delta}$, and to verify concept inclusions, and typicality inclusions, over such a model.

5. Model checking of a neural network in Datalog

We construct a many-valued interpretation $I_{\mathcal{N},n}^{\Delta}$ of a network $\mathcal{N}$ over a domain Δ, by restricting to the truth space $\mathcal{C}_n$, and approximating values $v \in [0,1]$ to the nearest value in $\mathcal{C}_n$ as follows:

$$[v]^n = \begin{cases} 0 & \text{if } v \leq \frac{1}{2n} \\ \frac{i}{n} & \text{if } \frac{2i-1}{2n} < v \leq \frac{2i+1}{2n}, \text{ for } 0 < i < n \\ 1 & \text{if } \frac{2n-1}{2n} < v \end{cases} \tag{2}$$

In the following, we will focus on the verification of properties of the network $\mathcal{N}$ in the logic $G_n\mathcal{LC}\mathbf{T}$, then building a $G_n\mathcal{LC}\mathbf{T}$ interpretation $I_{\mathcal{N},n}^{\Delta}$. The same approach can be used for verifying properties of the network in $Ł_n\mathcal{LC}\mathbf{T}$, with minor differences in Datalog encoding. First let us define the interpretation $I_{\mathcal{N},n}^{\Delta}$.

Definition 4. *The* many-valued interpretation $I_{\mathcal{N},n}^{\Delta} = \langle \Delta, \cdot^I \rangle$ of a network $\mathcal{N}$ *over the domain Δ, is a $G_n\mathcal{LC}\mathbf{T}$ interpretation such that function $\cdot^I$ satisfies, for all concept names $C_k \in N_C$ and domain elements $x \in \Delta$, the condition $C_k^I(x) = [y_k(x)]^n$, where $y_k(x)$ is the output signal of unit k, for input vector x.*

The verification that the network satisfies an inclusion of the form $C \sqsubseteq D \geq \alpha$, where C and D are concepts built from the concept names $C_i \in N_C$, possibly containing typicality concepts, can be done by checking whether the inclusion $C \sqsubseteq D \geq \alpha$ is satisfied in the model $I_{\mathcal{N},n}^{\Delta}$. As a special case, one can verify inclusions of the form $\mathbf{T}(C) \sqsubseteq D \geq \alpha$. Such formulae are interesting, e.g., in case C is associated to an output unit and D is a boolean combination of input units, to check whether inputs that are classified as Cs with highest degree, satisfy D with at least degree α.

In the following we describe a Datalog encoding of the model checking problem. The encoding contains a component $\Pi(\mathcal{N}, \Delta, n)$ which describes the interpretation $I^{\Delta}_{\mathcal{N},n}$, and a component associated to the formula or the formulae to be checked.

The program is defined in such a way that its unique stable model, corresponding to the well-founded model of the program, also corresponds to model $I^{\Delta}_{\mathcal{N},n}$. The main features of the program $\Pi(\mathcal{N}, \Delta, n)$ are the following.

The activation of the relevant units in $\mathcal{N}$ for each input stimulus $x \in \Delta$ is represented as follows. Each activation $y_i(x)$ is approximated to the nearest value $[y_i(x)]^n$ in $\mathcal{C}_n$ and transformed to an integer $v_i = [y_i(x)]^n \times n$. Each input stimulus $x \in \Delta$ is associated a number h, and a corresponding constant h in the program. A preprocessing phase will introduce in the program $\Pi(\mathcal{N}, \Delta, n)$ an atom $individual(h, v_1, \ldots, v_m)$ for each input stimulus x in Δ with number h, providing the tuple with all the (approximated) activation values for x of the units of interest (where $v_i = [y_i(x)]^n \times n$).

The valuation is encoded by a set of atoms of the form $inst(x, A, v)$, meaning that $\frac{v}{n} \in C_n$ is the degree of membership of x in A; $val(0..n)$ asserts that $0..n$ are the possible values, representing $\mathcal{C}_n$. For each concept name A_i associated to a unit of interest, the rule:

$$inst(X, A'_i, V_i) \qquad \leftarrow \qquad val(V_i), individual(X, V_1, \ldots, V_m).$$

where A'_i is the constant representing A_i, associates to each input stimulus x a membership degree $\frac{V_i}{n} \in \mathcal{C}_n$ in concept A_i. A rule

$$ind(X) \leftarrow individual(X, V_1, \ldots, V_m)$$

identifies individuals.

Formulae (concepts and concept inclusions) are represented using, for boolean concepts, terms such as $and(C', D')$ for $C \sqcap D$, where C' and D' represent C and D, and $t(C')$ for $\mathbf{T}(C)$. Function symbols are used as syntactic sugar, as the grounding of rules is finite.

The valuation is extended to boolean concepts C, and, similarly, to concept inclusions, defining a predicate $eval(C', X, V)$. As in [22] the definition of the *eval* predicate depends on the choice of the combination functions. For example, the rule:

$$\begin{aligned} eval(and(A, B), I, V) \leftarrow\ & ind(I), conc(and(A, B)), \\ & eval(A, I, V1), eval(B, I, V2), val(V1), \\ & val(V2), min(V1, V2, V). \end{aligned}$$

evaluates conjunctions, using a suitably defined min as combination function; $conc$ is used to make the instantiation of such rules finite, defining formulas of interest, that are the formulas to be verified and their subformulae.

Typical C-elements and the extension of $eval$ to typicality concepts can be defined using weakly stratified negation:

$$\begin{aligned} typical(X, C) \leftarrow\ & conc(t(C)), \\ & eval(C, X, N), N! = 0, \\ & hasmaxval(C, N). \end{aligned}$$

$$
\begin{aligned}
hasmaxval(C, n) \leftarrow\ & conc(t(C)), someval(C, n).\\
hasmaxval(C, M) \leftarrow\ & val(M), M < n,\\
& conc(t(C)), someval(C, M),\\
& not\ hasval\_geq(C, M + 1).\\
someval(C, M) \leftarrow\ & ind(Y), conc(t(C)),\\
& eval(C, Y, M).\\
hasval\_geq(C, M) \leftarrow\ & val(M), conc(t(C)),\\
& someval(C, M).\\
hasval\_geq(C, M) \leftarrow\ & val(M), conc(t(C)),\\
& M! = 0, M < n,\\
& hasval\_geq(C, M + 1).\\
eval(t(A), I, n) \leftarrow & ind(I), conc(t(A)),\\
& typical(I, A).\\
eval(t(A), I, 0) \leftarrow & ind(I), conc(t(A)),\\
& not\ typical(I, A).
\end{aligned}
$$

One or more formulae can be verified using, e.g., the following rules, relying on assertions $formula(Name, impl(C', D'), Val)$, to verify $C \sqsubseteq D \geq Val$, where $impl(C', D')$ represents $C \sqsubseteq D$, and the inclusion is given a (unique) $Name$; then either $ok(Name)$ or $notok(Name)$ will be derived, and, in the latter case, $notok/2$ points out the counterexamples:

$$
\begin{aligned}
conc(C) \leftarrow\ & formula(Name, C, Val).\\
notok(X, Fname) \leftarrow\ & formula(Fname, F, Val),\\
& ind(X), eval(F, X, V), V < Val.\\
notok(Fname) \leftarrow\ & notok(X, Fname).\\
ok(Fname) \leftarrow\ & fname(Fname),\\
& not\ notok(Fname).\\
fname(Name) \leftarrow\ & formula(Name, F, Val).
\end{aligned}
$$

The soundness and completeness of the Datalog encoding of the model checking problem in $I^{\Delta}_{\mathcal{N},n}$, can be proven along the same lines of the one-to-one correspondence between $G_n\mathcal{LC}\mathbf{T}$ models of a knowledge base and the answer sets of its ASP encoding [22] (Lemma 1 in the supplementary material). While in [22] the answer sets of the program capture the models of the conditional knowledge base associated to the network, here program $\Pi(\mathcal{N}, \Delta, n)$ has a unique weakly perfect model [28], corresponding to the interpretation $I^{\Delta}_{\mathcal{N},n}$ (as well as a unique stable model).

The size of the Datalog program $\Pi(\mathcal{N}, \Delta, n)$ is linear in $|\Delta| \times |N_C| \times n)$, where $|\Delta|$ is the size of the domain of input stimuli considered and $|N_C|$ is the number of concepts (units) which are of interest for the verification. It is easy to prove that the verification of a typicality inclusion $T(C) \sqsubseteq D \geq \alpha$ in $G_n\mathcal{LC}\mathbf{T}$ is $O(|\Delta| \times (|C| + |D|) \times n)$.

6. Typicality concepts and the probability of fuzzy events

For the properties $\mathbf{T}(C) \sqsubseteq D \geq \alpha$, especially in case they do not hold for all stimuli, the conditional probability of D given $\mathbf{T}(C)$, can be evaluated (and then compared with α) based on Zadeh's probability of fuzzy events. In particular, based on 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. [36], a probabilistic interpretation of SOMs has been provided [17], starting from a fuzzy model of SOMs after training. The same approach has been considered as well for MLPs [37]. In this section we consider as well typicality concepts, which do not require a special treatment, except considering their semantics, as for all other concepts.

Assuming a discrete probability distribution p over the domain Δ of a fuzzy interpretation $I = \langle \Delta, \cdot^I \rangle$, the probability of the fuzzy set C^I, for each DL concept C, can be defined as: $P(C^I) = \sum_{d \in \Delta} C^I(d)\ p(d)$. Let us consider the specific interpretation $I = I^{\Delta}_{\mathcal{N}}$ built from the trained network $\mathcal{N}$ over a set of input stimuli Δ. In the following we will simply write $P(C)$, rather than $P(C^I)$.

Following Smets [38], we let the conditional probability of a fuzzy event C given the fuzzy event D be $P(C \mid D) = P(D \sqcap C)/P(D)$ (provided $P(D) > 0$). As observed by Dubois and Prade [39], this generalizes both conditional probability and the fuzzy inclusion index advocated by Kosko [40]. Specifically, under the assumption that the probability distribution p is uniform over the set Δ of input stimuli, then $P(D|C) = M((D \sqcap C)^I)/M(C^I)$, where $M(C^I) = \sum_{x \in \Delta} C^I(x)$ is the *size* of the fuzzy event C^I in the interpretation I.

Note that, for a concept name $C_k \in N_C$, associated to a unit k, and a domain element $x \in \Delta$, it holds that $P(C_k|\{x\}) = C_k(x)$ [17, 37] (where $\{x\}$ stands for the crisp concept containing only x), which can be interpreted as a subjective probability that x is an instance of C_k [41], i.e., the *degree of belief* that x is an instance of concept C_k.

Computing conditional probabilities requires computing the size of the involved fuzzy sets. We have extended our rule based approach to compute conditional probabilities $P(D|\mathbf{T}(C))$ over the many-valued approximation $I^{\Delta}_{\mathcal{N},n}$ of model $I^{\Delta}_{\mathcal{N}}$, where D and C may be boolean concepts. In this case, the size of $(\mathbf{T}(C))^I$ coincides with the number of typical C elements, and the size of $(D \sqcap \mathbf{T}(C))^I$ can be computed as $\sum_{x \in (\mathbf{T}(C))^I} D^I(x)$. This allows for the verification of conditional constraints of the form $P(D|\mathbf{T}(C))\theta\alpha$ over the model $I^{\Delta}_{\mathcal{N},n}$. The computation can be performed using aggregates as follows:

$$
\begin{array}{l}
\mathit{numtyp}(N, C) \text{ :- } \mathit{conc}(t(C)),\\
\qquad N = \#\mathit{count}\{X : \mathit{typical}(X, C)\}.\\
\mathit{fuzzysetcondprob}(\mathit{Name}, P) \text{ :-}\\
\qquad \mathit{formula}(\mathit{Name}, \mathit{impl}(t(C), D), K),\\
\qquad \mathit{numtyp}(N, C),\\
\qquad W = \#\mathit{sum}\{V, X : \mathit{val}(V), \mathit{ind}(X),\\
\qquad\qquad \mathit{typical}(X, C), \mathit{eval}(D, X, V)\},\\
\qquad P = (k * W)/N.
\end{array}
$$

where k is, e.g., 1000, to get the decimal part of the result in 3 digits. This use of aggregates satisfies weak stratification restrictions [42] and the program has a unique stable model.

We report the results of an experimentation in the next section.

7. Recognizing basic emotions: an experimentation

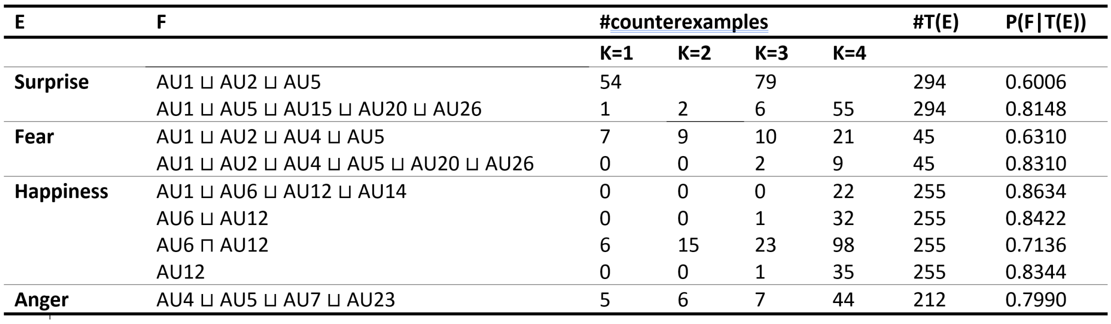
In this section, we report about experiments on the verification of properties of neural networks for the recognition of basic emotions using the Facial Action Coding System (FACS) [30].

The RAF-DB [43] data set contains almost 30000 images labeled with basic emotions or combinations of two emotions. The data set was used as input to OpenFace 2.0 [44], which detects a subset of the Action Units (AUs) in [30], i.e., facial muscle contractions. The relations between such AUs and emotions, studied by psychologists [45], 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 with only one emotion in the set { suprise, fear, happiness, anger }. The dataset is highly unbalanced and this can affect the training of the neural network model; then we preprocessed the data by subsampling the larger classes and augmenting the minority ones using standard data-augmentation techniques (e.g., rotations, flipping, etc.). 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 [30]. 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 network was trained using the Adam [46] optimizer with an initial learning rate η set to 0.003, and parameters $\beta_1 = 0.895$, $\beta_2 = 0.99$, and $\epsilon = 10^{-7}$. The network has been trained for 150 epochs with a batch size of 128 examples. All hyper-parameters have been tuned on a separated validation set.

The model checking approach in section 5 was applied, using the Clingo ASP solver as Datalog engine, taking, as set Δ of input stimuli, the test set used in the learning phase, containing 1194 images, and $n = 5$ (given that AU intensities, when assigned by humans, are on a scale of five values). Formulae of the form $\mathbf{T}(E) \sqsubseteq F \geq k/5$ were checked, where E is an emotion and F is a combination of AUs, using table 1 in [45] as a reference. Table 1 reports the results, with the number of typical individuals for the emotion, the number of counterexamples for different values of k, and the value of $P(F|\mathbf{T}(E))$.

For example, the formula $\mathbf{T}(happiness) \sqsubseteq au1 \sqcup au6 \sqcup au12 \sqcup au14 \geq 3/5$ holds for all individuals, as well as $\mathbf{T}(happiness) \sqsubseteq au12 \geq 2/5$, 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 $\mathbf{T}(happiness) \sqsubseteq au12 \geq 4/5$.



E	F	#counterexamples				#T(E)	P(F|T(E))
		K=1	K=2	K=3	K=4		
Surprise	AU1 ⊔ AU2 ⊔ AU5	54		79		294	0.6006
	AU1 ⊔ AU5 ⊔ AU15 ⊔ AU20 ⊔ AU26	1	2	6	55	294	0.8148
Fear	AU1 ⊔ AU2 ⊔ AU4 ⊔ AU5	7	9	10	21	45	0.6310
	AU1 ⊔ AU2 ⊔ AU4 ⊔ AU5 ⊔ AU20 ⊔ AU26	0	0	2	9	45	0.8310
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
Anger	AU4 ⊔ AU5 ⊔ AU7 ⊔ AU23	5	6	7	44	212	0.7990

Table 1
Results for checking formulae on the test set

8. Conclusions

In this paper we have described a Datalog approach to evaluate properties of trained MLPs by model checking. The approach exploits a finitely many-valued approximation of a semantics for fuzzy description logics with typicality.

As a proof of concept, the proposed approach has been experimented for checking properties of a trained neural network for the recognition of basic emotions using the Facial Action Coding System (FACS) [30].

This work is a step in the direction of verifying and interpreting knowledge learned by a neural network, in order to achieve a trustworthy and explainable AI. In the case study, there were expectations, to be verified, on the input-output relation (emotions and AUs); in other cases, less knowledge could be available in advance, so that the results could turn out to be more useful, even though more difficult to find.

Interpreting knowledge learned by a neural network in a logical form also opens the possibility of combining empirical knowledge with elicited knowledge, e.g., in the form of strict inclusions and definitions.

We refer to the surveys by Garcez et al. [47] and by Guidotti et al. [48] for an outline of current directions on the explanation of neural models and on the combination of neural networks and symbolic reasoning.

For future work, it would be interesting to investigate whether Fuzzy answer set programming (FASP) via satisfiability modulo theories (SMT) [49], can be used for MLPs property verification.

Acknowledgement: This research is partially supported 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] 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.

[3] D. Lehmann, M. Magidor, What does a conditional knowledge base entail?, Artificial Intelligence 55 (1992) 1–60.

[4] 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.

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

[6] 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.

[7] L. Giordano, V. Gliozzi, N. Olivetti, G. L. Pozzato, A NonMonotonic Description Logic for Reasoning About Typicality, Artif. Intell. 195 (2013) 165–202. doi:10.1016/j.artint.2012.10.004.

[8] 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.

[9] K. Britz, G. Casini, T. Meyer, K. Moodley, U. Sattler, I. Varzinczak, Principles of KLM-style defeasible description logics, ACM Trans. Comput. Log. 22 (2021) 1:1–1:46.

[10] 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.

[11] G. Casini, U. Straccia, Defeasible inheritance-based description logics, Journal of Artificial Intelligence Research (JAIR) 48 (2013) 415–473.

[12] 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.

[13] L. Giordano, D. Theseider Dupré, An ASP approach for reasoning in a concept-aware multipreferential lightweight DL, TPLP 10(5) (2020) 751–766.

[14] 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.

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

[16] 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.

[17] 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.

[18] T. Kohonen, M. Schroeder, T. Huang (Eds.), Self-Organizing Maps, Third Edition, Springer Series in Information Sciences, Springer, 2001.

[19] 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.

[20] 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.

[21] L. Giordano, D. Theseider Dupré, Weighted conditional $\mathrm{EL}^{\perp}$ knowledge bases with integer weights: an ASP approach, in: Proc. 37th Int. Conf. on Logic Programming, ICLP 2021 (Technical Communications), Porto, Sept. 20-27, 2021, volume 345 of *EPTCS*, 2021, pp.

70–76. URL: https://doi.org/10.4204/EPTCS.345.19.

[22] L. Giordano, D. Theseider Dupré, An ASP approach for reasoning on neural networks under a finitely many-valued semantics for weighted conditional knowledge bases, TPLP 22 (2022) 589–605.

[23] A. García-Cerdaña, E. Armengol, F. Esteva, Fuzzy description logics and t-norm based fuzzy logics, Int. J. Approx. Reason. 51 (2010) 632–655. doi:10.1016/j.ijar.2010.01.001.

[24] F. Bobillo, U. Straccia, Reasoning with the finitely many-valued Łukasiewicz fuzzy Description Logic SROIQ, Inf. Sci. 181 (2011) 758–778. doi:10.1016/j.ins.2010.10.020.

[25] F. Bobillo, M. Delgado, J. Gómez-Romero, U. Straccia, Joining Gödel and Zadeh Fuzzy Logics in Fuzzy Description Logics, Int. J. Uncertain. Fuzziness Knowl. Based Syst. 20 (2012) 475–508. doi:10.1142/S0218488512500249.

[26] S. Borgwardt, R. Peñaloza, The complexity of lattice-based fuzzy description logics, J. Data Semant. 2 (2013) 1–19.

[27] G. Brewka, J. P. Delgrande, J. Romero, T. Schaub, asprin: Customizing answer set preferences without a headache, in: Proc. AAAI 2015, 2015, pp. 1467–1474.

[28] H. Przymusinska, T. C. Przymusinski, Weakly perfect model semantics for logic programs, in: Logic Programming, Proceedings of the Fifth International Conference and Symposium, Seattle, Washington, USA, August 15-19, 1988 (2 Volumes), MIT Press, 1988, pp. 1106–1120.

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

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

[31] 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.

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

[33] 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.

[34] P. Cintula, P. Hájek, C. Noguera (Eds.), Handbook of Mathematical Fuzzy Logic, volume 37-38, College Publications, 2011.

[35] F. Bobillo, U. Straccia, Reasoning within fuzzy OWL 2 EL revisited, Fuzzy Sets Syst. 351 (2018) 1–40.

[36] 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.

[37] L. Giordano, D. Theseider Dupré, Weighted defeasible knowledge bases and a multipreference semantics for a deep neural network model, CoRR abs/2012.13421 (2021). Technical Report, https://arxiv.org/abs/2012.13421v2.

[38] P. Smets, Probability of a fuzzy event: An axiomatic approach, Fuzzy Sets and Systems 7 (1982) 153–164.

[39] D. Dubois, H. Prade, Fuzzy sets and probability: misunderstandings, bridges and gaps, in: [Proceedings 1993] Second IEEE International Conference on Fuzzy Systems, 1993, pp. 1059–1068 vol.2. doi:10.1109/FUZZY.1993.327367.

[40] B. Kosko, Neural networks and fuzzy systems: a dynamical systems approach to machine intelligence, Prentice Hall, 1992.

[41] N. Friedman, J. Y. Halpern, D. Koller, First-order conditional logic for default reasoning revisited, ACM TOCL, ACM Press 1 (2000) 175–207.

[42] K. A. Ross, Modular stratification and magic sets for Datalog programs with negation, J. ACM 41 (1994) 1216–1266.

[43] 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.

[44] 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.

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

[46] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[47] 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.

[48] 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.

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