Logic Tensor Networks for Semantic Image Interpretation

Ivan Donadello
Fondazione Bruno Kessler and
University of Trento
Trento, Italy
donadello@fbk.eu

Luciano Serafini
Fondazione Bruno Kessler
Via Sommarive 18, I-38123
Trento, Italy
serafini@fbk.eu

Artur d'Avila Garcez
City, University of London
Northampton Square
London EC1V 0HB, UK
a.garcez@city.ac.uk

Abstract

Semantic Image Interpretation (SII) is the task of extracting structured semantic descriptions from images. It is widely agreed that the combined use of visual data and background knowledge is of great importance for SII. Recently, Statistical Relational Learning (SRL) approaches have been developed for reasoning under uncertainty and learning in the presence of data and rich knowledge. Logic Tensor Networks (LTNs) are a SRL framework which integrates neural networks with first-order fuzzy logic to allow (i) efficient learning from noisy data in the presence of logical constraints, and (ii) reasoning with logical formulas describing general properties of the data. In this paper, we develop and apply LTNs to two of the main tasks of SII, namely, the classification of an image's bounding boxes and the detection of the relevant *part-of* relations between objects. To the best of our knowledge, this is the first successful application of SRL to such SII tasks. The proposed approach is evaluated on a standard image processing benchmark. Experiments show that background knowledge in the form of logical constraints can improve the performance of purely data-driven approaches, including the state-of-the-art Fast Region-based Convolutional Neural Networks (Fast R-CNN). Moreover, we show that the use of logical background knowledge adds robustness to the learning system when errors are present in the labels of the training data.

1 Introduction

Semantic Image Interpretation (SII) is the task of generating a structured semantic description of the content of an image. This structured description can be represented as a labelled directed graph. Each vertex corresponds to a bounding box of an object in the image, each edge represents a relation between pairs of objects. Vertices are labelled with a set of object types and edges are labelled with binary relations. Such a graph is also called a *scene graph* in [Krishna *et al.*, 2016].

A major obstacle to be overcome by SII is the so-called *semantic gap* [Neumann and Möller, 2008], that is, the lack of a direct correspondence between low-level features of the image and high-level semantic descriptions. To tackle this problem, a system for SII must learn the latent correlations that may exist between the numerical features that can be observed in an image and the semantic concepts associated with the objects. It is in this learning process that the availability of relational background knowledge can be of great help. Thus, recent SII systems have sought to combine, or even integrate, visual features obtained from data and symbolic knowledge in the form of logical axioms [Zhu *et al.*, 2014; Chen *et al.*, 2012; Donadello and Serafini, 2016].

The area of Statistical Relational Learning (SRL), or Statistical Artificial Intelligence (StarAI), seeks to combine data-driven learning, in the presence of uncertainty, with symbolic knowledge [Wang and Domingos, 2008; Bach *et al.*, 2015; Gutmann *et al.*, 2010; Diligenti *et al.*, 2015; Rocktaschel *et al.*, 2015; Ravkic *et al.*, 2015]. However, only very few SRL systems have been applied to SII tasks (c.f. Section 2) due to the high complexity associated with image learning. Most systems for solving SII tasks have been based, instead, on deep learning and neural network models. These, on the other hand, do not in general offer a well-founded way of learning from data in the presence of relational logical constraints, requiring the neural models to be highly engineered from scratch. In this paper, we develop and apply for the first time, the SRL framework called Logic Tensor Networks (LTNs) to computationally challenging SII tasks. LTNs combine learning in deep networks with relational logical constraints [Serafini and d'Avila Garcez, 2016]. It uses a First-order Logic (FOL) syntax interpreted in the real numbers, which is implemented as a deep tensor network. Logical terms are interpreted as feature vectors in a real-valued n-dimensional space. Function symbols are interpreted as real-valued functions, and predicate symbols as fuzzy logic relations. This syntax and semantics, called *real semantics*, allow LTNs to learn efficiently in hybrid domains, where elements are composed of both numerical and relational information. We argue, therefore, that LTNs are a good candidate for learning SII because they can express relational knowledge in FOL which serves as constraints on the data-driven learning within tensor networks. Being LTN a logic, it provides a notion of logical consequence, which forms the basis for learning within LTNs, which is defined as *best satisfiability*, c.f. Section 4. Solving the best satisfiability problem amounts to finding the latent correlations that may exist between a relational back-

ground knowledge and numerical data attributes. This formulation enables the specification of *learning as reasoning*, a unique characteristic of LTNs, which is seen as highly relevant for SII. This paper specifies SII within LTNs, evaluating it on two important tasks: (i) the classification of bounding boxes, and (ii) the detection of the *part-of* relation between any two bounding boxes. Both tasks are evaluated using the PASCAL-PART dataset [Chen *et al.*, 2014]. It is shown that LTNs improve the performance of the state-of-the-art object classifier Fast R-CNN [Girshick, 2015] on the bounding box classification task. LTNs also outperform a rule-based heuristic (which uses the *inclusion ratio* of two bounding boxes) in the detection of *part-of* relations between objects. Finally, LTNs are evaluated on their ability to handle errors, specifically misclassification of objects and part-of relations. Very large visual recognition datasets now exist which are noisy [Reed *et al.*, 2014], and it is important for learning systems to become robust to noise. LTNs were trained systematically on progressively noisier datasets, with results on both SII tasks showing that LTN's logical constraints are capable of adding robustness to the system, in the presence of errors in the labels of the training data. The paper is organized as follows: Section 2 contrasts the LTN approach with related work which integrate visual features and background knowledge for SII. Section 3 specifies LTNs in the context of SII. Section 4 defines the best satisfiability problem in this context, which enables the use of LTNs for SII. Section 5 describes in detail the comparative evaluations of LTNs on the SII tasks. Section 6 concludes the paper and discusses directions for future work.

2 Related Work

The idea of exploiting logical background knowledge to improve SII tasks dates back to the early days of AI. In what follows, we review the most recent results in the area in comparison with LTNs.

Logic-based approaches have used Description Logics (DL), where the basic components of the scene are all assumed to have been already discovered (e.g. simple object types or spatial relations). Then, with logical reasoning, new facts can be derived in the scene from these basic components [Neumann and Möller, 2008; Peraldi *et al.*, 2009]. Other logic-based approaches have used fuzzy DL to tackle uncertainty in the basic components [Hudelot *et al.*, 2008; Dasiopoulou *et al.*, 2009; Atif *et al.*, 2014]. These approaches have limited themselves to spatial relations or to refining the labels of the objects detected. In [Donadello and Serafini, 2016], the scene interpretation is created by combining image features with constraints defined using DL, but the method is tailored to the *part-of* relation and cannot be extended easily to account for other relations. LTNs, on the other hand, should be able to handle any semantic relation. In [Marszalek and Schmid, 2007; Forestier *et al.*, 2013], a symbolic Knowledge-base is used to improve object detection, but only the *subsumption* relation is explored and it is not possible to inject more complex knowledge using logical axioms.

A second group of approaches encodes background knowledge and visual features within *probabilistic graphical models*. In [Zhu *et al.*, 2014; Nyga *et al.*, 2014], visual features are combined with knowledge gathered from datasets, web resources or annotators, about object labels, properties (e.g., shape, colour, size) and affordances, using Markov Logic Networks (MLNs) [Richardson and Domingos, 2006] to predict facts in unseen images. Due to the specific knowledge-base schema adopted, the effectiveness of MLNs in this domain is evaluated only for Horn clauses, although the language of MLNs is more general. As a result, it is not easy to evaluate how the approach may perform with more complex axioms. In [Bach *et al.*, 2015], a probabilistic fuzzy logic is used, but not with real semantics. Clauses are weighted and universally-quantified formulas are instantiated, as done by MLNs. This is different from LTNs where the universally-quantified formulas are computed by using an aggregation operation, which avoids the need for instantiating all variables.

In other related work, [Chen *et al.*, 2012; Kulkarni *et al.*, 2011] encode background knowledge into a generic Conditional Random Field (CRF), where the nodes represent detected objects and the edges represent logical relationships between objects. The task is to find a correct labelling for this graph. In [Chen *et al.*, 2012], the edges encode logical constraints on a knowledge-base specified in DL. Although these ideas are close in spirit to the approach presented in this paper, they are not formalised as in LTNs, which use a deep tensor network and first-order logic, rather than CRFs or DL. In general, the logical theory behind the functions to be defined in the CRF is unclear. In [Kulkarni *et al.*, 2011], potential functions are defined as text priors such as co-occurrence of terms found in the image descriptions of Flickr.

In a final group of approaches, here called *language-priors*, background knowledge is taken from linguistic models [Ramanathan *et al.*, 2015; Lu *et al.*, 2016]. In [Ramanathan *et al.*, 2015], a neural network is built integrating visual features and a linguistic model to predict semantic relationships between bounding boxes. The linguistic model is a set of rules derived from WORDNET [Fellbaum, 1998], stating which types of semantic relationships occur between a subject and an object. In [Lu *et al.*, 2016], a similar neural network is proposed for the same task but with a more sophisticated language model, embedding in the same vector space triples of the form *subject-relation-object*, such that semantically similar triples are mapped closely together in the embedding space. In this way, even if no examples exist of some triples in the data, the relations can be inferred from similarity to more frequent triples. A drawback, however, is the possibility of inferring inconsistent triples, such as e.g. *man-eats-chair*, due to the embedding. LTNs avoid this problem with a logic-based approach (in the above example, with an axiom to the effect that chairs are not normally edible). LTNs can also handle exceptions, offering a system capable of dealing with crisp axioms and real-valued data, as specified in what follows.

3 Logic Tensor Networks

Let $\mathcal{L}$ be a first-order logic language, whose signature is composed of three disjoint sets $\mathcal{C}$, $\mathcal{F}$ and $\mathcal{P}$, denoting constants, functions and predicate symbols, respectively. For any function or predicate symbol s, let $\alpha(s)$ denote its arity. Logical formulas in $\mathcal{L}$ allow one to specify relational knowledge,

e.g. the atomic formula $\mathsf{partOf}(o_1, o_2)$, stating that object o_1 is a part of object o_2, the formulae $\forall xy(\mathsf{partOf}(x, y) \rightarrow \neg\mathsf{partOf}(y, x))$, stating that the relation partOf is asymmetric, or $\forall x(\mathsf{Cat}(x) \rightarrow \exists y(\mathsf{partOf}(x, y) \wedge \mathsf{Tail}(y)))$, stating that every cat should have a tail. In addition, exceptions are handled by allowing formulas to be interpreted in fuzzy logic, such that in the presence of an example of, say, a tailless cat, the above formula can be interpreted naturally as *normally, every cat has a tail*; this will be exemplified later.

Semantics of $\mathcal{L}$: We define the interpretation domain as a subset of $\mathbb{R}^n$, i.e. every object in the domain is associated with a n-dimensional vector of real numbers. Intuitively, this n-tuple represents n numerical features of an object, e.g. in the case of a person, their name in ASCII, height, weight, social security number, etc. Functions are interpreted as real-valued functions, and predicates are interpreted as fuzzy relations on real vectors. To emphasise the fact that we interpret symbols as real numbers, we use the term *grounding* instead of *interpretation*[1] in the following definition of semantics.

Definition 1 *Let $n \in \mathbb{N}$. An n-*grounding*, or simply* grounding*, $\mathcal{G}$ for a FOL $\mathcal{L}$ is a function defined on the signature of $\mathcal{L}$ satisfying the following conditions:*

1. $\mathcal{G}(c) \in \mathbb{R}^n$ *for every constant symbol* $c \in \mathcal{C}$;
2. $\mathcal{G}(f) \in \mathbb{R}^{n \cdot \alpha(f)} \longrightarrow \mathbb{R}^n$ *for every* $f \in \mathcal{F}$;
3. $\mathcal{G}(P) \in \mathbb{R}^{n \cdot \alpha(P)} \longrightarrow [0, 1]$ *for every* $P \in \mathcal{P}$.

Given a grounding $\mathcal{G}$, the semantics of closed terms and atomic formulas is defined as follows:

$$\mathcal{G}(f(t_1, \ldots, t_m)) = \mathcal{G}(f)(\mathcal{G}(t_1), \ldots, \mathcal{G}(t_m))$$
$$\mathcal{G}(P(t_1, \ldots, t_m)) = \mathcal{G}(P)(\mathcal{G}(t_1), \ldots, \mathcal{G}(t_m))$$

The semantics for connectives is defined according to fuzzy logic; using for instance the Lukasiewicz t-norm:[2]

$$\mathcal{G}(\neg\phi) = 1 - \mathcal{G}(\phi)$$
$$\mathcal{G}(\phi \wedge \psi) = \max(0, \mathcal{G}(\phi) + \mathcal{G}(\psi) - 1)$$
$$\mathcal{G}(\phi \vee \psi) = \min(1, \mathcal{G}(\phi) + \mathcal{G}(\psi))$$
$$\mathcal{G}(\phi \rightarrow \psi) = \min(1, 1 - \mathcal{G}(\phi) + \mathcal{G}(\psi))$$

The LTN semantics for $\forall$ is defined in [Serafini and d'Avila Garcez, 2016] using the min operator, that is, $\mathcal{G}(\forall x\phi(x)) = \min_{t \in term(\mathcal{L})} \mathcal{G}(\phi(t)))$, where $term(\mathcal{L})$ is the set of instantiated terms of $\mathcal{L}$. This, however, is inadequate for our purposes as it does not tolerate exceptions (the presence of a single exception to the universally-quantified formulae, such as e.g. a cat without a tail, would falsify the formulae). Instead, our aim in SII is that the more examples that satisfy a formulae $\phi(x)$, the higher the truth-value of $\forall x\phi(x)$ should be. To capture this, we use for the semantics of $\forall$ a *mean*-operator:

$$\mathcal{G}(\forall x\phi(x)) = \lim_{T \rightarrow term(\mathcal{L})} mean_p(\mathcal{G}(\phi(t)) \mid t \in T)$$

where $mean_p(x_1, \ldots, x_d) = \left(\frac{1}{d}\sum_{i=1}^{d} x_i^p\right)^{\frac{1}{p}}$ for $p \in \mathbb{Z}$.[3]

Finally, the classical semantics of $\exists$ is uniquely determined by the semantics of $\forall$, by making $\exists$ equivalent to $\neg\forall\neg$. This approach, however, has a drawback too when it comes to SII: if we adopt, for instance, the arithmetic mean for the semantic of $\forall$ then $\mathcal{G}(\forall x\phi(x)) = \mathcal{G}(\exists x\phi(x))$. Therefore, we shall interpret existential quantification via Skolemization: every formula of the form $\forall x_1, \ldots, x_n(\ldots \exists y\phi(x_1, \ldots, x_n, y))$ is rewritten as $\forall x_1, \ldots, x_n(\ldots \phi(x_1, \ldots, x_n, f(x_1, \ldots, x_n)))$, by introducing a new n-ary function symbol, called Skolem function. In this way, existential quantifiers can be eliminated from the language by introducing Skolem functions.

Formalizing SII in LTNs: To specify the SII problem, as defined in the introduction, we consider a signature $\Sigma_{\text{SII}} = \langle \mathcal{C}, \mathcal{F}, \mathcal{P} \rangle$, where $\mathcal{C} = \bigcup_{p \in Pics} b(p)$ is the set of identifiers for all the bounding boxes in all the images, $\mathcal{F} = \emptyset$, and $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}$, where $\mathcal{P}_1$ is a set of unary predicates, one for each object type, e.g. $\mathcal{P}_1 = \{\mathsf{Dog}, \mathsf{Cat}, \mathsf{Tail}, \mathsf{Muzzle}, \mathsf{Train}, \mathsf{Coach}, \ldots\}$, and $\mathcal{P}_2$ is a set of binary predicates representing relations between objects. Since in our experiments we focus on the *part-of* relation, $\mathcal{P}_2 = \{\mathsf{partOf}\}$. The FOL formulas based on this signature can specify (i) simple facts, e.g. the fact that bounding box b contains a cat, written $\mathsf{Cat}(b)$, the fact that b contains either a cat or a dog, written $\mathsf{Cat}(b) \vee \mathsf{Dog}(b)$, etc., and (ii) general rules such as $\forall x(\mathsf{Cat}(x) \rightarrow \exists y(\mathsf{partOf}(x, y) \wedge \mathsf{Tail}(y)))$.

A grounding for Σ_{SII} can be defined as follows: each constant b, denoting a bounding box, can be associated with a set of geometric features and a set of semantic features computed with a bounding box detector. Specifically, each bounding box is associated with *geometric features* describing the position and the dimension of the bounding box, and *semantic features* describing the classification score returned by the bounding box detector for each class. For example, for each bounding box $b \in \mathcal{C}$, $C_i \in \mathcal{P}_1$, $\mathcal{G}(b)$ is the $\mathbb{R}^{4+|\mathcal{P}_1|}$ vector:

$$\langle class(C_1, b), \ldots, class(C_{|\mathcal{P}_1|}, b), x_0(b), y_0(b), x_1(b), y_1(b) \rangle$$

where the last four elements are the coordinates of the top-left and bottom-right corners of b, and $class(C_i, b) \in [0, 1]$ is the classification score of the bounding box detector for b.

An example of groundings for predicates can be defined by taking a one-vs-all multi-classifier approach, as follows. First, define the following grounding for each class $C_i \in \mathcal{P}_1$ (below, $\mathbf{x} = \langle x_1, \ldots, x_{|\mathcal{P}_1|+4} \rangle$ is the vector corresponding to the grounding of a bounding box):

$$\mathcal{G}(C_i)(\mathbf{x}) = \begin{cases} 1 & \text{if } i = \text{argmax}_{1 \leq l \leq |\mathcal{P}_1|} x_l \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Then, a simple rule-based approach for defining a grounding for the partOf relation is based on the naïve assumption that the more a bounding box b is contained within a bounding box b', the higher the probability should be that b is part of b'. Accordingly, one can define $\mathcal{G}(\mathsf{partOf}(b, b'))$ as the inclusion ratio $ir(b, b')$ of bounding box b, with grounding $\mathbf{x}$, into bounding box b', with grounding $\mathbf{x}'$ (formally,

[1] In logic, the term *grounding* indicates the operation of replacing the variables of a term or formula with constants or terms that do not contain other variables. To avoid any confusion, we use the synonym *instantiation* for this purpose. It is worth noting that in LTN, differently from MLNs, the instantiation of every first order formula is not required.

[2] Examples of t-norms include Lukasiewicz, product and Gödel. The Lukasiewicz t-norm is $\mu_{Luk}(x, y) = \max(0, x+y-1)$, product t-norm is $\mu_{Pr}(x, y) = x \cdot y$, and Gödel t-norm is $\mu_{max}(x, y) = \min(x, y)$. See [Bergmann, 2008] for details.

[3] The popular mean operators, arithmetic, geometric and harmonic mean, are obtained by setting $p = 1, 2$, and -1, respectively.

$ir(b,b') = \frac{area(b \cap b')}{area(b)}$). A slightly more sophisticated rule-based grounding for partOf (used as baseline in the experiments to follow) takes into account also *type compatibilities* by multiplying the inclusion ratio by a factor w_{ij}. Hence, we define $\mathcal{G}(\mathsf{partOf}(b,b'))$ as follows:

$$\begin{cases} 1 & \text{if } ir(b,b') \cdot \max_{ij=1}^{|\mathcal{P}_1|}(w_{ij} \cdot x_i \cdot x'_j) \geq th_{ir} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

for some threshold th_{ir} (we use $th_{ir} > 0.5$), and with $w_{ij} = 1$ if C_i is a part of C_j, and 0 otherwise. Given the above grounding, we can compute the grounding of any atomic formula, e.g. $\mathsf{Cat}(b_1)$, $\mathsf{Dog}(b_2)$, $\mathsf{leg}(b_3)$, $\mathsf{partOf}(b_3,b_1)$, $\mathsf{partOf}(b_3,b_2)$, thus expressing the degree of truth of the formula. The rule-based groundings (Eqs. (1) and (2)) may not satisfy all the constraints to be imposed. E.g., the classification score may be wrong, a bounding box may include another one which is not in the part-of relation, etc. Furthermore, in many situations it is not possible to define a grounding a priori. Instead, groundings should be learned automatically from data, by optimizing the truth-values of the formulas in the background knowledge. This is discussed next.

4 Learning as Best Satisfiability

A partial grounding, denoted by $\hat{\mathcal{G}}$, is a grounding that is defined on a subset of the signature of $\mathcal{L}$. A grounding $\mathcal{G}$ is said to be a completion of $\hat{\mathcal{G}}$, if $\mathcal{G}$ is a grounding for $\mathcal{L}$ and coincides with $\hat{\mathcal{G}}$ on the symbols where $\hat{\mathcal{G}}$ is defined.

Definition 2 *A* grounded theory GT *is a pair $\langle \mathcal{K}, \hat{\mathcal{G}} \rangle$ with a set $\mathcal{K}$ of closed formulas and a partial grounding $\hat{\mathcal{G}}$.*

Definition 3 *A grounding $\mathcal{G}$ satisfies a GT $\langle \mathcal{K}, \hat{\mathcal{G}} \rangle$ if $\mathcal{G}$ completes $\hat{\mathcal{G}}$ and $\mathcal{G}(\phi) = 1$ for all $\phi \in \mathcal{K}$. A GT $\langle \mathcal{K}, \hat{\mathcal{G}} \rangle$ is* satisfiable *if there exists a grounding $\mathcal{G}$ that satisfies $\langle \mathcal{K}, \hat{\mathcal{G}} \rangle$.*

According to the previous definition, deciding the satisfiability of $\langle \mathcal{K}, \hat{\mathcal{G}} \rangle$ amounts to searching for a grounding $\hat{\mathcal{G}}$ such that *all* the formulas of $\mathcal{K}$ are mapped to 1. Differently from the classical satisfiability, when a GT is not satisfiable, we are interested in the best possible satisfaction that we can reach with a grounding. This is defined as follows.

Definition 4 *Let $\langle \mathcal{K}, \hat{\mathcal{G}} \rangle$ be a grounded theory. We define the* best satisfiability problem *as the problem of finding a grounding $\mathcal{G}^*$ that maximizes the truth-values of the conjunction of all clauses $cl \in \mathcal{K}$, i.e. $\mathcal{G}^* = \text{argmax}_{\hat{\mathcal{G}} \subseteq \mathcal{G} \in \mathbb{G}} \mathcal{G}(\bigwedge_{cl \in \mathcal{K}} cl)$.*

Grounding $\mathcal{G}^*$ captures the latent correlation between quantitative attributes of objects and their categorical/relational properties. Not all functions are suitable as a grounding; they should preserve some form of regularity. If $\mathcal{G}(\mathsf{Cat})(\mathbf{x}) \approx 1$ (the bounding box with feature vector $\mathbf{x}$ contains a cat) then for every $\mathbf{x}'$ close to $\mathbf{x}$ (i.e. for every bounding box with features similar to $\mathbf{x}$), one should have $\mathcal{G}(\mathsf{Cat})(\mathbf{x}') \approx 1$. In particular, we consider groundings of the following form.

Function symbols are grounded to linear transformations. If f is a m-ary function symbol, then $\mathcal{G}(f)$ is of the form:

$$\mathcal{G}(f)(\mathbf{v}) = M_f \mathbf{v} + N_f$$

where $\mathbf{v} = \langle \mathbf{v}_1^\intercal, \ldots, \mathbf{v}_m^\intercal \rangle^\intercal$ is the mn-ary vector obtained by concatenating each $\mathbf{v}_i$. The parameters for $\mathcal{G}(f)$ are the $n \times mn$ real matrix M_f and the n-vector N_f.

The grounding of an m-ary predicate P, namely $\mathcal{G}(P)$, is defined as a generalization of the neural tensor network (which has been shown effective at knowledge completion in the presence of simple logical constraints [Socher *et al.*, 2013]), as a function from $\mathbb{R}^{mn}$ to $[0,1]$, as follows:

$$\mathcal{G}(P)(\mathbf{v}) = \sigma\left(u_P^\intercal \tanh\left(\mathbf{v}^\intercal W_P^{[1:k]} \mathbf{v} + V_P \mathbf{v} + b_P\right)\right) \quad (3)$$

with σ the sigmoid function. The parameters for P are: $W_P^{[1:k]}$, a 3-D tensor in $\mathbb{R}^{k \times mn \times mn}$, $V_P \in \mathbb{R}^{k \times mn}$, $b_P \in \mathbb{R}^k$ and $u_P \in \mathbb{R}^k$. This last parameter performs a linear combination of the quadratic features given by the tensor product. With this encoding, the grounding (i.e. truth-value) of a clause can be determined by a neural network which first computes the grounding of the literals contained in the clause, and then combines them using the specific t-norm.

In what follows, we describe how a suitable GT can be built for SII. Let $Pics^t \subseteq Pics$ be a set of bounding boxes of images correctly labelled with the classes that they belong to, and let each pair of bounding boxes be correctly labelled with the part-of relation. In machine learning terminology, $Pics^t$ is a *training set* without noise. In real semantics, a training set can be represented by a theory $\mathcal{T}_{\text{expl}} = \langle \mathcal{K}_{\text{expl}}, \hat{\mathcal{G}} \rangle$, where $\mathcal{K}_{\text{expl}}$ contains the set of closed literals $C_i(b)$ (resp. $\neg C_i(b)$) and $\mathsf{partOf}(b,b')$ (resp. $\neg\mathsf{partOf}(b,b')$), for every bounding box b labelled (resp. not labelled) with C_i and for every pair of bounding boxes $\langle b, b' \rangle$ connected (resp$\neg\mathsf{partOf}(b,b')$. not connected) by the partOf relation. The partial grounding $\hat{\mathcal{G}}$ is defined on all bounding boxes of all the images in $Pics$ where both the semantic features $class(C_i, b)$ and the bounding box coordinates are computed by the Fast R-CNN object detector [Girshick, 2015]. $\hat{\mathcal{G}}$ is not defined for the predicate symbols in $\mathcal{P}$ and is to be learned. $\mathcal{T}_{\text{expl}}$ contains only assertional information about specific bounding boxes. This is the classical setting of machine learning where classifiers (i.e. the grounding of predicates) are inductively learned from positive examples (such as $\mathsf{partOf}(b,b')$) and negative examples ($\neg\mathsf{partOf}(b,b')$) of a classification. In this learning setting, mereological constraints such as "cats have no wheels" or "a tail is a part of a cat" are not taken into account. Examples of mereological constraints state, for instance, that the part-of relation is asymmetric ($\forall xy(\mathsf{partOf}(x,y) \rightarrow \neg\mathsf{partOf}(y,x))$), or lists the several parts of an object (e.g. $\forall xy(\mathsf{Cat}(x) \wedge \mathsf{partOf}(x,y) \rightarrow \mathsf{Tail}(y) \vee \mathsf{Muzzle}(y))$), or even, for simplicity, that every whole object cannot be part of another object (e.g. $\forall xy(\mathsf{Cat}(x) \rightarrow \neg\mathsf{partOf}(x,y))$) and every part object cannot be divided further into parts (e.g. $\forall xy(\mathsf{Tail}(x) \rightarrow \neg\mathsf{partOf}(y,x))$). This general knowledge is available from on-line resources, such as WordNet [Fellbaum, 1998], and can be retrieved by inheriting the meronymy relations for every concept corresponding to a whole object. A grounded theory that considers also mereological constraints as prior knowledge can be constructed by adding such axioms to $\mathcal{K}_{\text{expl}}$. More formally, we define $\mathcal{T}_{\text{prior}} = \langle \mathcal{K}_{\text{prior}}, \hat{\mathcal{G}} \rangle$, where $\mathcal{K}_{\text{prior}} = \mathcal{K}_{\text{expl}} + \mathcal{M}$, and $\mathcal{M}$ is

the set of mereological axioms. To check the role of $\mathcal{M}$, we evaluate both theories and then compare results.

5 Experimental Evaluation

We evaluate the performance of our approach for SII[4] on two tasks, namely, the classification of bounding boxes and the detection of partOf relations between pairs of bounding boxes. We chose the part-of relation because both data (the PASCAL-PART-dataset [Chen *et al.*, 2014]) and ontologies (WORDNET) are available on the part-of relation. In addition, part-of can be used to represent, via reification, a large class of relations [Guarino and Guizzardi, 2016] (e.g., the relation "a plant is lying on the table" can be reified in an object of type "lying event" whose parts are the plant and the table). However, it is worth noting that many other relations could have been included in this evaluation. The time complexity of LTN grows linearly with the number of axioms.

We also evaluate the robustness of our approach with respect to noisy data. It has been acknowledged by many that, with the vast growth in size of the training sets for visual recognition [Krishna *et al.*, 2016], many data annotations may be affected by noise such as missing or erroneous labels, non-localised objects, and disagreements between annotations, e.g. human annotators often mistake "part-of" for the "have" relation [Reed *et al.*, 2014].

We use the PASCAL-PART-dataset that contains 10103 images with bounding boxes annotated with object-types and the part-of relation defined between pairs of bounding boxes. Labels are divided into three main groups: animals, vehicles and indoor objects, with their corresponding parts and "part-of" label. Whole objects inside the same group can share parts. Whole objects of different groups do not share any parts. Labels for parts are very specific, e.g. "left lower leg". Thus, without loss of generality, we have merged the bounding boxes that referred to the same part into a single bounding box, e.g. bounding boxes labelled with "left lower leg" and "left upper leg" were merged into a single bounding box of type "leg". In this way, we have limited our experiments to a dataset with 20 labels for whole objects and 39 labels for parts. In addition, we have removed from the dataset any bounding boxes with height or width smaller than 6 pixels. The images were then split into a training set with 80%, and a test set with 20% of the images, maintaining the same proportion of the number of bounding boxes for each label.

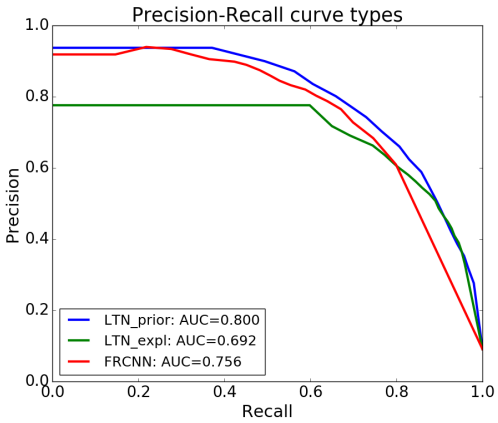
Object Type Classification and Detection of the Part-Of Relation: Given a set of bounding boxes detected by an object detector (we use Fast-RCNN), the task of object classification is to assign to each bounding box an object type. The task of Part-Of detection is to decide, given two bounding boxes, if the object contained in the first is a part of the object contained in the second. We use LTN to resolve both tasks simultaneously. This is important because a bounding box type and the part-of relation are not independent. Their dependencies are specified in LTN using background knowledge in the form of logical axioms.

To show the effect of the logical axioms, we train two LTNs: the first containing only training examples of object types and part-of relations ($\mathcal{T}_{\text{expl}}$), and the second containing also logical axioms about types and part-of ($\mathcal{T}_{\text{prior}}$). The LTNs were set up with tensor of $k = 6$ layers and a regularization parameter $\lambda = 10^{-10}$. We chose Lukasiewicz's T-norm ($\mu(a, b) = \max(0, a + b - 1)$) and use the harmonic mean as aggregation operator. We ran 1000 training epochs of the RMSProp learning algorithm available in TENSORFLOW^{TM}. We compare results with the Fast RCNN at object type classification (Eq.(1)), and the *inclusion ratio* ir baseline (Eq.(2)) at the part-of detection task.[5] If ir is larger than a given threshold th (in our experiments, $th = 0.7$) then the bounding boxes are said to be in the partOf relation. Every bounding box b is classified into $C \in \mathcal{P}_1$ if $\mathcal{G}(C(b)) \geq th$. With this, a bounding box can be classified into more than one class. For each class, precision and recall are calculated in the usual way. Results for indoor objects are shown in Figure 1 where AUC is the area under the precision-recall curve. The results show that, for both object types and the part-of relation, the LTN trained with prior knowledge given by mereological axioms has better performance than the LTN trained with examples only. Moreover, prior knowledge allows LTN to improve the performance of the Fast R-CNN (FRCNN) object detector. Notice that the LTN is trained using the Fast R-CNN results as features. FRCNN assigns a bounding box to a class if the values of the corresponding semantic features exceed th. This is local to the specific semantic features. If such local features are very discriminative (which is the case in our experiments) then very good levels of precision can be achieved. Differently from FRCNN, LTNs make a global choice which takes into consideration all (semantic and geometric) features together. This should offer robustness to the LTN classifier at the price of a drop in precision. The logical axioms compensate this drop. For the other object types (animals and vehicles), LTN has results comparable to FRCNN: FRCNN beats $\mathcal{T}_{\text{prior}}$ by 0.05 and 0.037 AUC, respectively, for animals and vehicles. Finally, we have performed an initial experiment on *small data*, on the assumption that the LTN axioms should be able to compensate a reduction in training data. By removing 50% of the training data for indoor objects, a similar performance to $\mathcal{T}_{\text{prior}}$ with the full training set can be achieved: 0.767 AUC for object types and 0.623 AUC for the part-of relation, which shows an improvement in performance.

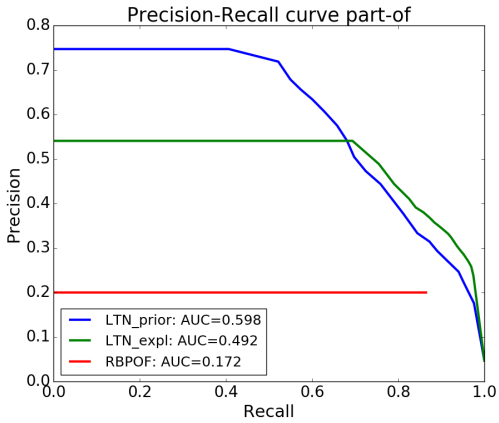
Robustness to Noisy Training Data: In this evaluation, we show that logical axioms improve the robustness of LTNs in the presence of errors in the labels of the training data. We have added an increasing amount of noise to the PASCAL-PART-dataset training data, and measured how performance degrades in the presence and absence of axioms. For $k \in \{10, 20, 30, 40\}$, we randomly select $k\%$ of the bounding boxes in the training data, and randomly change their classification labels. In addition, we randomly select $k\%$ of pairs of bounding boxes, and flip the value of the part-of relation's label. For each value of k, we train LTNs $\mathcal{T}^k_{\text{expl}}$ and $\mathcal{T}^k_{\text{prior}}$

[4] LTN has been implemented as a Google TENSORFLOW^{TM} library. Code, partOf ontology, and dataset are available at `https://gitlab.fbk.eu/donadello/LTN_IJCAI17`

[5] A direct comparison with [Chen *et al.*, 2012] is not possible because their code was not available.



(a) LTNs with prior knowledge improve the performance of Fast R-CNN on object type classification, achieving an Area Under the Curve (AUC) of 0.800 in comparison with 0.756.



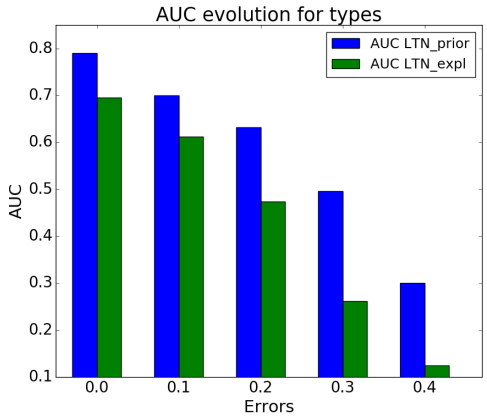
(b) LTNs with prior knowledge outperform the rule-based approach of Eq.2 in the detection of part-of relations, achieving AUC of 0.598 in comparison with 0.172.

Figure 1: Precision-recall curves for indoor objects type classification and the partOf relation between objects.

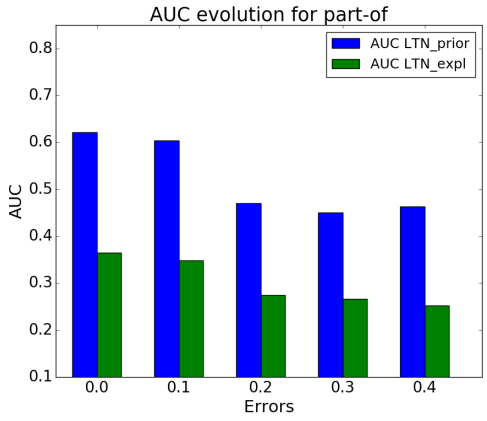
and evaluate results on both SII tasks as done before. As expected, adding too much noise to training labels leads to a large drop in performance. Figure 2 shows the AUC measures for indoor objects with increasing error k. Each pair of bars indicates the AUC of $\mathcal{T}^k_{\text{prior}}, \mathcal{T}^k_{\text{expl}}$, for a given $k\%$ of errors. Results indicate that the LTN axioms offer robustness to noise: in addition to the expected overall drop in performance, an increasing gap can be seen between the drop in performance of the LTN trained with examples only and the LTN trained including background knowledge.

6 Conclusion and Future Work

SII systems are required to address the semantic gap problem: combining visual low-level features with high-level concepts. We argue that the problem can be addressed by the integration of numerical and logical representations in deep learning. LTNs learn from numerical data and logical constraints, enabling approximate reasoning on unseen data to predict new facts. In this paper, LTNs were shown to improve on state-of-the-art method Fast R-CNN for bounding box classification, and to outperform a rule-based method at learning part-of relations in the PASCAL-PART-dataset. Moreover, LTNs were evaluated on how to handle noisy data through the systematic creation of training sets with errors in the labels. Results indicate that relational knowledge can add robustness to neural systems. As future work, we shall apply LTNs to larger datasets such as VISUAL GENOME, and continue to compare the various instances of LTN with SRL, deep learning and other neural-symbolic approaches on such challenging visual intelligence tasks.



(a) Object types



(b) Part-of predicate

Figure 2: AUCs for indoor object types and part-of relation with increasing noise in the labels of the training data. The drop in performance is noticeably smaller for the LTN trained with background knowledge.

References

[Atif *et al.*, 2014] J. Atif, C. Hudelot, and I Bloch. Explanatory reasoning for image understanding using formal concept analysis and description logics. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 44(5):552–570, May 2014.

[Bach *et al.*, 2015] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *CoRR*, abs/1505.04406, 2015.

[Bergmann, 2008] M. Bergmann. *An Introduction to Many-Valued and Fuzzy Logic: Semantics, Algebras, and Derivation Systems*. Cambridge University Press, 2008.

[Chen *et al.*, 2012] Na Chen, Qian-Yi Zhou, and Viktor Prasanna. Understanding web images by object relation network. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 291–300, New York, NY, USA, 2012. ACM.

[Chen *et al.*, 2014] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014.

[Dasiopoulou *et al.*, 2009] S. Dasiopoulou, Y. Kompatsiaris, and M.l G. Strintzis. Applying fuzzy dls in the extraction of image semantics. *J. Data Semantics*, 14:105–132, 2009.

[Diligenti *et al.*, 2015] Michelangelo Diligenti, Marco Gori, and Claudio Saccà. Semantic-based regularization for learning and inference. *Artificial Intelligence*, 2015.

[Donadello and Serafini, 2016] I. Donadello and L. Serafini. Integration of numeric and symbolic information for semantic image interpretation. *Intelligenza Artificiale*, 10(1):33–47, 2016.

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

[Forestier *et al.*, 2013] G. Forestier, C. Wemmert, and A. Puissant. Coastal image interpretation using background knowledge and semantics. *Computers & Geosciences*, 54:88–96, 2013.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015.

[Guarino and Guizzardi, 2016] N. Guarino and G. Guizzardi. On the reification of relationships. In *24th Italian Symp. on Advanced Database Sys.*, pages 350–357, 2016.

[Gutmann *et al.*, 2010] B. Gutmann, M. Jaeger, and L. De Raedt. Extending problog with continuous distributions. In *Proc. ILP*, pages 76–91. Springer, 2010.

[Hudelot *et al.*, 2008] C. Hudelot, J. Atif, and I. Bloch. Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems*, 159(15):1929–1951, 2008.

[Krishna *et al.*, 2016] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-L. Li, D. Shamma, M. Bernstein, and Li F.-F. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[Kulkarni *et al.*, 2011] G.h Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In *CVPR*, 2011.

[Lu *et al.*, 2016] C. Lu, R. Krishna, M. Bernstein, and L Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016.

[Marszalek and Schmid, 2007] Marcin Marszalek and Cordelia Schmid. Semantic Hierarchies for Visual Object Recognition. In *CVPR*, 2007.

[Neumann and Möller, 2008] Bernd Neumann and Ralf Möller. On scene interpretation with description logics. *Image and Vision Computing*, 26(1):82 – 101, 2008. Cognitive Vision-Special Issue.

[Nyga *et al.*, 2014] D. Nyga, F. Balint-Benczedi, and M. Beetz. Pr2 looking at things-ensemble learning for unstructured information processing with markov logic networks. In *IEEE Intl. Conf.on Robotics and Automation*, pages 3916–3923, 2014.

[Peraldi *et al.*, 2009] I. S. Espinosa Peraldi, A. Kaya, and R. Möller. Formalizing multimedia interpretation based on abduction over description logic aboxes. In *Proc. of the 22nd Intl. Workshop on Description Logics*, volume 477 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.

[Ramanathan *et al.*, 2015] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rosenberg, and L. Fei-Fei. Learning semantic relationships for better action retrieval in images. In *CVPR*, 2015.

[Ravkic *et al.*, 2015] Irma Ravkic, Jan Ramon, and Jesse Davis. Learning relational dependency networks in hybrid domains. *Machine Learning*, 100(2-3):217–254, 2015.

[Reed *et al.*, 2014] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *CoRR*, abs/1412.6596, 2014.

[Richardson and Domingos, 2006] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

[Rocktaschel *et al.*, 2015] T. Rocktaschel, S. Singh, and S. Riedel. Injecting logical background knowledge into embeddings for relation extraction. In *NAACL*, 2015.

[Serafini and d'Avila Garcez, 2016] L. Serafini and A. S. d'Avila Garcez. Learning and reasoning with logic tensor networks. In *Proc. AI*IA*, pages 334–348, 2016.

[Socher *et al.*, 2013] R. Socher, D. Chen, C. D. Manning, and A. Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, pages 926–934, 2013.

[Wang and Domingos, 2008] Jue Wang and Pedro M Domingos. Hybrid markov logic networks. In *AAAI*, volume 8, pages 1106–1111, 2008.

[Zhu *et al.*, 2014] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, pages 408–424. 2014.