

# Relational Neural Machines

Giuseppe Marra<sup>1</sup> and Michelangelo Diligenti<sup>2</sup> and Francesco Giannini<sup>2</sup> and Marco Gori<sup>2</sup> and Marco Maggini<sup>2</sup>

**Abstract.** Deep learning has been shown to achieve impressive results in several tasks where a large amount of training data is available. However, deep learning solely focuses on the accuracy of the predictions, neglecting the reasoning process leading to a decision, which is a major issue in life-critical applications. Probabilistic logic reasoning allows to exploit both statistical regularities and specific domain expertise to perform reasoning under uncertainty, but its scalability and brittle integration with the layers processing the sensory data have greatly limited its applications. For these reasons, combining deep architectures and probabilistic logic reasoning is a fundamental goal towards the development of intelligent agents operating in complex environments. This paper presents Relational Neural Machines, a novel framework allowing to jointly train the parameters of the learners and of a First-Order Logic based reasoner. A Relational Neural Machine is able to recover both classical learning from supervised data in case of pure sub-symbolic learning, and Markov Logic Networks in case of pure symbolic reasoning, while allowing to jointly train and perform inference in hybrid learning tasks. Proper algorithmic solutions are devised to make learning and inference tractable in large-scale problems. The experiments show promising results in different relational tasks.

## 1 Introduction

In the last few years, the availability of a large amount of supervised data caused a significant improvement in the performances of sub-symbolic approaches like artificial neural networks. In particular, deep neural networks have achieved impressive results in several tasks, thanks to their ability to jointly learn the decision function and the data representation from the low-level perception inputs [13, 23]. However, the dependency on the amount and quality of training data is also a major limitation of this class of approaches. Standard neural networks can struggle to represent relational knowledge on different input patterns, or relevant output structures, which have been shown to bring significant benefits in many challenging applications like image segmentation tasks [3]. For this reason, several work has been done in the direction of learning and representing relations using embeddings [42, 17, 30, 8, 1] and in developing and injecting relational features into the learning process [33, 32, 37].

On the other hand, symbolic approaches [31, 4, 20] are generally based on probabilistic logic reasoners, and can express high-level relational dependencies in a certain domain of discourse and perform exact or approximate inference in presence of uncertainty. Markov Logic Networks (MLN) [34] and its variants like Probabilis-

tic Soft Logic [2] are relational undirected models, mapping First-Order Logic formulas to a Markov network, and allowing to train the parameters of the reasoner and perform inference under uncertainty.

Another related line of research studies hybrid approaches leveraging neural networks to learn the structure of the reasoning process like done, for instance, by Relational Restricted Boltzmann machines [19] and auto-encoding logic programs [9]. Similarly, Neural Markov Logic Networks [29] extend MLN by defining the potential functions as general neural networks which are trained together with the model parameters. Neural Theorem Prover [35, 36] is an end-to-end differentiable prover that shows state-of-the-art performances on some link prediction benchmarks by combining Prolog backward chain with a soft unification scheme. TensorLog [44, 18] is a recent framework to reuse the deep learning infrastructure of TensorFlow to perform probabilistic logical reasoning.

Whereas the previously discussed methods provide a large step forward in the definition of a flexible and data-driven reasoning process, they do not still allow to co-optimize the low-level learners processing the environmental data. Methods bridging the gap between symbolic and sub-symbolic levels are commonly referred as neuro-symbolic approaches [11, 19, 39]. An early attempt to integrate learning and reasoning is the work by Lippi et al. [24]. The main limitation of this work is that it was devised ad-hoc to solve a specific task in bioinformatics and it does not define a general methodology to apply it to other contexts.

A distillation method to inject logic knowledge into the network weights is proposed by Hu et al. [16]. The main limitation of this work is that the logic is distilled into the network weights but there is no guarantee that the logic will be properly generalized to the test cases. Deep Structured Models [3] define a schema to inject complex output structures into deep learners. The approach is general but it does not focus on logic knowledge but on imposing statistical structure on the output predictions. Hazan et al. [15] integrate learning and inference in Conditional Random Fields [41], but they also do not focus on logic reasoning. The Semantic Loss [43] allows to translate the desired output structure of a learner via the definition of a loss, which can also accommodate logic knowledge. However, the loss and the resulting reasoning process does not define a probabilistic reasoning process, thus limiting the flexibility of the approach. Deep ProbLog [26] is a neuro-symbolic approach, based on the probabilistic logic programming language ProbLog [4] and approximating the predicates via deep learning. This approach is very flexible but it is limited to cases where exact inference is possible, as it lacks a modular and scalable solution like the one proposed in this paper.

Deep Logic Models (DLM) [27] are instead capable of jointly training the sensory and reasoning layers in a single differentiable architecture, which is a major advantage with respect to related approaches like Semantic-based Regularization [5], Logic Tensor Net-

<sup>1</sup> Department of Information Engineering, University of Florence, Florence, email: g.marra@unifi.it

<sup>2</sup> Department of Information Engineering and Science, University of Siena, Siena, email: {diligimic,fgiannini,marco,maggini}@diism.unisi.it

works [6] or Neural Logic Machines [7]. However, DLM is based on a brittle stacking of the learning and reasoning modules, failing to provide a real tight integration on how low-level learner employs the supervised data. For this reason, DLM requires the employment of heuristics like training plans to make learning effective.

This paper presents Relational Neural Machines (RNM), a novel framework introducing fundamental improvements over previous state-of-the-art-models in terms of scalability and in the tightness of the connection between the trainer and the reasoner. A RNM is able to perfectly replicate the effectiveness of training from supervised data of standard deep architectures, while still co-training a reasoning module over the environment that is built during the learning process. The bonding is very general as any (deep) learner can be integrated and any output or input structure can be expressed. On the other hand, when restricted to pure symbolic reasoning, RNM can replicate the expressivity of Markov Logic Networks [34].

The outline of the paper is as follows. Section 2 presents the model and how it can be used to integrate logic and learning. Section 3 studies tractable approaches to perform inference and model training from supervised and unsupervised data. Section 4 shows the experimental evaluation of the proposed ideas on various datasets. Finally, Section 5 draws some conclusions and highlights some planned future work.

## 2 Model

A Relational Neural Machine establishes a probability distribution over a set of  $n$  output variables of interest  $\mathbf{y} = \{y_1, \dots, y_n\}$ , given a set of predictions made by one or multiple deep architectures, and the model parameters. In this paper the output variables are assumed to be binary, i.e.  $y_i = \{0, 1\}$ , but the model can be extended to deal with continuous values for regression tasks.

Unlike standard neural networks which compute the output via a simple forward pass, the output computation in an RNM can be decomposed into two stages: a *low-level* stage processing the input patterns, and a subsequent *semantic* stage, expressing constraints over the output and performing higher level reasoning. In this paper, it is assumed that there is a single network processing the input sensorial data, but the theory is trivially extended to any number of learners. The first stage processes  $D$  input patterns  $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_D\}$ , returning the values  $\mathbf{f}$  using the network with parameters  $\mathbf{w}$ . The higher layer takes as input  $\mathbf{f}$  and applies reasoning using a set of constraints, whose parameters are indicated as  $\lambda$ , then it returns the set of output variables  $\mathbf{y}$ .

A RNM model defines a conditional probability distribution in the exponential family defined as:

$$p(\mathbf{y}|\mathbf{f}, \lambda) = \frac{1}{Z} \exp \left( \sum_c \lambda_c \Phi_c(\mathbf{f}, \mathbf{y}) \right) \quad (1)$$

where  $Z$  is the partition function and the  $C$  potentials  $\Phi_c$  express some properties on the input and output variables. The parameters  $\lambda = \{\lambda_1, \dots, \lambda_C\}$  determine the strength of the potentials  $\Phi_c$ .

This model can express a vast range of typical learning tasks. We start reviewing how to express simple classification problems, before moving to general neuro-symbolic integration mixing learning and reasoning. A main advantage of RNMs is that they can jointly express and solve these use cases, which are typically been studied as stacked separate problems.

In a classical and pure supervised learning setup, the patterns are i.i.d., it is therefore possible to split the  $\mathbf{y}$ ,  $\mathbf{f}$  into disjoint sets group-

ing the variables of each pattern, forming separate cliques. Let us indicate as  $\mathbf{y}(x)$ ,  $\mathbf{f}(x)$  the portion of the output and function variables referring to the processing of an input pattern  $x$ . A single potential  $\Phi_0$  is needed to represent supervised learning, and this potential decomposes over the patterns as:

$$\Phi_0(\mathbf{y}, \mathbf{f}) = \sum_{x \in S} \phi_0(\mathbf{y}(x), \mathbf{f}(x)) . \quad (2)$$

where  $S \subseteq \mathbf{x}$  is the set of supervised patterns. This yields the distribution,

$$p_0(\mathbf{y}|\mathbf{f}, \lambda) = \frac{1}{Z} \exp \left( \sum_{x \in S} \phi_0(\mathbf{y}(x), \mathbf{f}(x)) \right) \quad (3)$$

**One-label classification.** The mutual exclusivity rule defines a zero probability to assignments where a pattern belongs to more than one class. The following potential is defined for an input pattern  $x$ :

$$\phi_0(\mathbf{y}(x), \mathbf{f}(x)) = \begin{cases} -\infty & \text{if } \exists i, j : y_i(x) = y_j(x) = 1, i \neq j \\ \mathbf{f}(x) \cdot \mathbf{y} & \text{otherwise} \end{cases}$$

When only the  $\Phi_0$  potential is used, each pattern corresponds to a set of outputs independent on the other pattern outputs given the  $\mathbf{f}$ , the partition function decomposes over the patterns such that:

$$p_0(\mathbf{y}|\mathbf{f}, \lambda) = \begin{cases} 0 & \text{if } \exists x, i, j : y_i(x) = y_j(x) = 1, i \neq j \\ \frac{\exp \left( \sum_{x \in S} \mathbf{f}(x) \cdot \mathbf{y}(x) \right)}{\sum_{\mathbf{y}} \prod_{x \in S} \exp(\mathbf{f}(x) \cdot \mathbf{y}(x))} = & \text{otherwise} \\ = \prod_{x \in S} \frac{\exp(\mathbf{f}(x) \cdot \mathbf{y}(x))}{\sum_{i \in Y} \exp(f_i(x) \cdot y_i(x))} = \\ = \prod_{x \in S} [\text{softmax}(\mathbf{f}(x) \cdot \mathbf{y}(x))]_{\mathcal{I}(\mathbf{y}(x))} \end{cases}$$

where  $\mathcal{I}(\mathbf{y}(x))$  is the index of the element equal to 1 in  $\mathbf{y}(x)$  and  $\text{softmax}(\mathbf{f}(x) \cdot \mathbf{y}(x))$  is the softmax distribution for  $\mathbf{f}(x) \cdot \mathbf{y}(x)$ . This result provides an elegant justification for the usage of the softmax output for networks used in one-label classification tasks.

**Multi-label.** The following potential is expressed for each input pattern:  $\phi_0(\mathbf{y}(x), \mathbf{f}(x)) = \mathbf{f}(x) \cdot \mathbf{y}(x)$ .

When plugging in the previously defined potential into the potential in Equation 2 and the result plugged into Equation 3, the partition function can be decomposed into one component for each pattern and class, since each pattern and classification output is independent on all the other classifications:

$$\begin{aligned} p_0(\mathbf{y}|\mathbf{f}, \lambda) &= \frac{\exp \left( \sum_{x \in S} \mathbf{f}(x) \cdot \mathbf{y}(x) \right)}{\exp \left( \sum_{x \in S} \sum_{\mathbf{y}(x)} \mathbf{f}(x) \cdot \mathbf{y}(x) \right)} = \\ &= \prod_{x \in S} \prod_{i \in Y} \frac{\exp(f_i(x) \cdot y_i(x))}{1 + \exp(f_i(x))} = \\ &= \prod_{x \in S} \left[ \prod_{i \in Y^+(x)} \sigma(f_i(x)) \cdot \prod_{i \in Y^-(x)} (1 - \sigma(f_i(x))) \right] \end{aligned}$$

where  $\sigma$  is the sigmoid function,  $Y^+(x)$ ,  $Y^-(x)$  are the set of positive and negative classes for pattern  $x$ . This result provides an elegant justification for the usage of a sigmoidal output layer for multi-label classification tasks.

## 2.1 Neuro-symbolic integration

The most interesting and general case is when the presented model is used to perform both learning and reasoning, which is a task referred in the literature as neuro-symbolic integration.

Image Correlations $\forall O \in \mathcal{O}$
$\forall i_1 \forall i_2 \text{ SameLoc}(i_1, i_2) \Rightarrow (O(i_1) \wedge O(i_2)) \vee (\neg O(i_1) \wedge \neg O(i_2))$
Knowledge Graphs
$\forall i O(i) \Rightarrow \text{Mammal}(i) \quad \forall O \in \{\text{Lion}, \text{Cat}, \text{Dog}, \text{Sheep}, \dots\}$
$\forall i \text{Mammal}(i) \Rightarrow \text{Animal}(i)$
$\forall i \text{Mammal}(i) \Rightarrow \text{Legs}(i) \wedge \text{Body}(i) \wedge \text{Tail}(i)$
General Knowledge
$\forall i \text{Lion}(i) \rightarrow \text{Savanna}(i) \vee \text{Zoo}(i)$
$\forall i \text{Wall}(i) \rightarrow \neg \text{Savanna}(i)$
Category Correlations
$\forall i \text{PolarBear}(i) \wedge \text{Lion}(i)$
$\forall i \text{Antelope}(i) \wedge \text{Lion}(i)$
Supervisions
$\text{Lion}(i_1), \text{Wall}(i_1), \dots$

**Table 1.** An example of the knowledge available to express an object detection task, where  $\mathcal{O}$  is the set of all possible objects (or predicates to be learned). Supervisions about the image  $i_1$  containing the objects *Lion*, and *Wall* is added together to the background knowledge.

The general model described in Equation 1 is materialized with one potential  $\Phi_0$  enforcing the consistency with the supervised data together with potentials representing the logic knowledge. Using a similar approach to Markov Logic Networks, a set of First-Order Logic (FOL) formulas is input to the system, and there is a potential  $\Phi_c$  for each formula. The general form of the conditional probability distribution becomes:

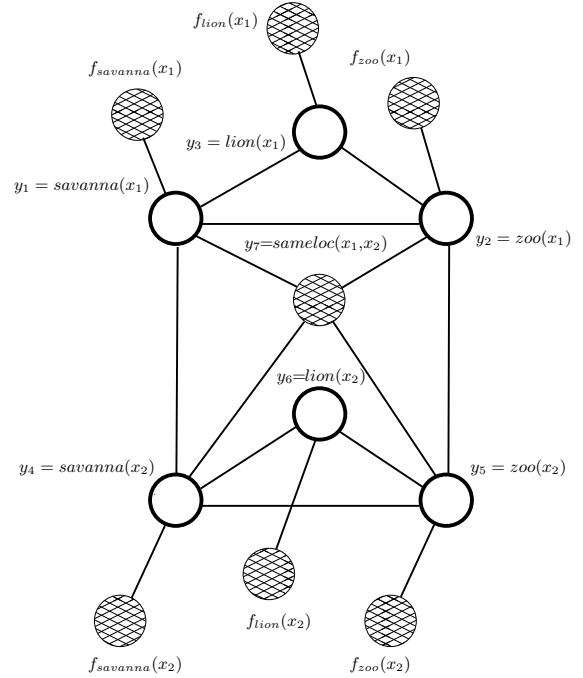
$$p(\mathbf{y}|\mathbf{f}, \lambda) = \frac{1}{Z} \exp \left( \sum_x \phi_0(\mathbf{f}(x), \mathbf{y}(x)) + \sum_c \lambda_c \Phi_c(\mathbf{y}) \right) \quad (4)$$

where it is assumed that some (or all) the predicates in a KB are unknown and need to be learned together with the  $\lambda$  parameters driving the reasoning process.

A grounded expression (the same applies to atom or predicate) is a FOL rule whose variables are assigned to specific constants. It is assumed that the undirected graphical model has the following structure, each grounded atom corresponds to a node in the graph, and all nodes connected by at least one rule are connected on the graph, so that there is one clique (and then potential) for each grounding  $g_c$  of the formula in  $\mathbf{y}$ . It is assumed that all the potentials resulting from the  $c$ -th formula share the same weight  $\lambda_c$ , therefore the potential  $\Phi_c$  is the sum over all groundings of  $\phi_c$  in the world  $\mathbf{y}$ , such that:  $\Phi_c(\mathbf{y}) = \sum_{\mathbf{y}_{c,g}} \phi_c(\mathbf{y}_{c,g})$  where  $\phi_c(g_c)$  assumes a value equal to 1 and 0 if the grounded formula holds true and false. This yields the probability distribution:

$$p(\mathbf{y}|\mathbf{f}, \lambda) = \frac{1}{Z} \exp \left( \sum_x \phi_0(\mathbf{f}(x), \mathbf{y}(x)) + \sum_c \lambda_c \sum_{\mathbf{y}_{c,g}} \phi_c(\mathbf{y}_{c,g}) \right)$$

This will allow to develop the data embeddings as part of training by enforcing the consistency between the reasoner and network outputs, while distilling the knowledge into the network weights.



**Figure 1.** The graphical model representing a RNM, where the output variables  $\mathbf{y}$  depend on the output of first stage  $\mathbf{f}$ , processing the inputs  $\{x_1, x_2\}$  instantiated for the rules  $\forall x \text{lion}(x) \Rightarrow \text{savanna}(x) \vee \text{zoo}(x)$ ,  $\forall x \forall x' \text{sameLoc}(x, x') \wedge \text{savanna}(x) \Rightarrow \text{savanna}(x')$ , and  $\forall x \forall x' \text{sameLoc}(x, x') \wedge \text{zoo}(x) \Rightarrow \text{zoo}(x')$ .

**Example.** It is required to train a classifier detecting the objects on images for a multi-object detection task in real world pictures. A knowledge graph may be available to describe hierarchical dependencies among the object classes, or object compositions. Pictures may be correlated by the locations where they have been shot. Table 1 shows the knowledge that could be used to express such a task, where the unknown predicates to be trained are indicated as the set  $\mathcal{O}$ . Other predicates like *SameLoc* may be known a priori based the meta-information attached to the images. Figure 1 shows the graphical model correlating the output variables  $\mathbf{y}$  and the  $\mathbf{f}$  for the inputs  $\{x_1, x_2\}$  instantiated for the rules  $\forall x \text{lion}(x) \Rightarrow \text{savanna}(x) \vee \text{zoo}(x)$ ,  $\forall x \forall x' \text{sameLoc}(x, x') \wedge \text{savanna}(x) \Rightarrow \text{savanna}(x')$ , and  $\forall x \forall x' \text{sameLoc}(x, x') \wedge \text{zoo}(x) \Rightarrow \text{zoo}(x')$ . The goal of the training process is to train the classifiers approximating the predicates, but also to establish the relevance of each rule. For example, the formula  $\forall x \text{Antelope}(x) \wedge \text{Lion}(x)$  is likely to be associated to a higher weight than  $\forall x \text{PolarBear}(x) \wedge \text{Lion}(x)$ , which are unlikely to correlate in the data.

**Logic Tensor Networks.** Logic Tensor Networks (LTN) [38] is a framework to learn neural networks under the constraints imposed by some prior knowledge expressed as a set of FOL clauses. As shown in this paragraph, LTN is a special case of a RNM, when the  $\lambda$  parameters are frozen. In particular, an LTN expresses each FOL rule via a continuous relaxation of a logic rule using fuzzy logic. The strength  $\lambda_c$  of the rule is assumed to be known a priori and not trained by the LTN. These rules provide a prior for the functions. Therefore, assuming the  $\lambda$  parameters are fixed, an LTN considers the following

distribution:

$$\begin{aligned} p(\mathbf{f}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{f}) \cdot p(\mathbf{f}) = \\ &= \frac{1}{Z_1} \exp \left( \sum_{x \in S} \phi_0(y(x), f(x)) \right) \cdot \\ &\quad \cdot \frac{1}{Z_2} \exp \left( \sum_c \lambda_c \Phi_c^s(\mathbf{f}) \right) \end{aligned}$$

where  $\Phi_c^s$  is the continuous relaxation of the  $c$ -th logic rule,  $p(\mathbf{y}|\mathbf{f})$  is used to express the fitting of the supervised data and the prior  $p(\mathbf{f})$  gives preference to the functions respecting the logic constraints. The parameters  $\mathbf{w}$  of the  $\mathbf{f}$  of an LTN can be optimized via gradient ascent by maximizing the likelihood of the training data.

**Semantic Based Regularization.** Semantic-Based Regularization (SBR) [5], defines a learning and reasoning framework which allows to train neural networks under the constraints imposed by the prior logic knowledge. The declarative language Lyrics [28] is available to provide a flexible and easy to use frontend for the SBR framework. At training time, SBR employs the knowledge like done by LTN, while SBR uses a continuous relaxation  $\Phi_c^s$  of the  $c$ -th logic rule and of the output vector at inference time. Therefore, SBR can also be seen as a special instance of a RNM, when the  $\lambda$  parameters are frozen and the continuous relaxation of the logic is used at test time. Both LTN and SBR have a major disadvantage over RNM, as they can not learn the weights of the reasoner, which are required to be known a priori. This is very unlikely to happen in most of the real world scenarios, where the strength of each rule must be co-trained with the learning system.

### 3 Learning and Inference

**Training.** A direct model optimization in RNM is intractable in most interesting cases, as the computation of the partition function requires a summation over all possible assignments of the output variables. However, if a partition function is assumed to be factorized into separate independent groups of variables  $\mathbf{y}_i \in \mathbf{y}$ , it holds that:

$$Z \approx \prod_i Z_i$$

A particularly interesting case is when it is assumed that the partition function factorizes over the potentials like done in piecewise likelihood [40]:

$$Z \approx \prod_c Z_c = \prod_c \left[ \sum_{\mathbf{y}'_c} \exp(\lambda_c \Phi_c(\mathbf{f}, \mathbf{y}'_c)) \right]$$

where  $\mathbf{y}_c$  is the subset of variables in  $\mathbf{y}$  that are involved in the computation of  $\Phi_c$ . We indicate as piecewise-local probability for the  $c$ -th constraint:

$$p_c^{PL}(\mathbf{y}_c|\mathbf{f}, \lambda) = \frac{\exp(\lambda_c \Phi_c(\mathbf{f}, \mathbf{y}_c))}{Z_c} = \frac{\exp(\lambda_c \Phi_c(\mathbf{f}, \mathbf{y}_c))}{\sum_{\mathbf{y}'_c} \exp(\lambda_c \Phi_c(\mathbf{f}, \mathbf{y}'_c))} \quad (5)$$

Under this assumption, the factors can be distributed over the product giving the following generalized piecewise likelihood:

$$p(\mathbf{y}|\mathbf{f}, \lambda) \approx \prod_c p(\mathbf{y}_c|\mathbf{y} \setminus \mathbf{y}_c, \mathbf{f}, \lambda) = \prod_c p_c^{PL}(\mathbf{y}_c|\mathbf{f}, \lambda)$$

If the variables in  $\mathbf{y}$  are binary, the computation of  $Z$  requires summation over all possible assignments which has  $O(2^{|\mathbf{y}|})$  complexity. Using the local decomposition this is reduced to  $O(|\mathbf{y}_c| \cdot 2^{n_c})$ , where  $\bar{c}$  is the index of the formula corresponding to the potential with the largest number  $n_{\bar{c}}$  of variables to ground. When a single potential involves too many variables, the pseudo-likelihood decomposition can be used, where each variable is factorized into a separate component with linear complexity with respect to the number of variables for each grounding with complexity:

$$p_c^{PPL}(\mathbf{y}_c|\mathbf{f}, \lambda) = \prod_{y_i \in \mathbf{y}_c} \frac{\lambda_c \Phi_c(\mathbf{f}, \mathbf{y}_c)}{\sum_{b \in \{0,1\}} \exp(\lambda_c \Phi_c(\mathbf{f}, [y_i = b, \mathbf{y}_c \setminus y_i]))}$$

where the factorization is performed with respect of the single variables, which has a cost proportional to  $n$ .

Assuming that the constraints are divided into two groups  $\{C_1, C_2\}$ , for which the local piecewise partitioning  $PL$  and the pseudo-likelihood  $PPL$  approximations are used, the distribution becomes:

$$p(\mathbf{y}|\mathbf{f}, \lambda) \approx \prod_{c \in C_1} p_c^{PL}(\mathbf{y}_c|\mathbf{f}, \lambda_c) \cdot \prod_{c \in C_2} p_c^{PPL}(\mathbf{y}_c|\mathbf{f}, \lambda_c)$$

If the  $c$ -th constraint is factorized using the  $PL$  partitioning, the derivatives of the log-likelihood with respect to the model potential weights are:

$$\begin{aligned} \frac{\partial \log p(\mathbf{y}|\mathbf{f}, \lambda)}{\partial \lambda_c} &\approx \Phi_c(\mathbf{f}, \mathbf{y}) - \sum_{\mathbf{y}'_c \in \mathbf{Y}_c} p_c^{PL}(\mathbf{y}'_c|\mathbf{f}, \lambda) \cdot \Phi_c(\mathbf{f}, \mathbf{y}'_c) = \\ &= \Phi_c(\mathbf{f}, \mathbf{y}) - E_{p_c^{PL}}[\Phi_c] \end{aligned} \quad (6)$$

and with respect to the learner parameters:

$$\begin{aligned} \frac{\partial \log p(\mathbf{y}|\mathbf{f}, \lambda)}{\partial w_k} &\approx \sum_i y_i \frac{\partial f_i}{\partial w_k} - E_{p_c^{PL}} \left[ \sum_i y_i \frac{\partial f_i}{\partial w_k} \right] = \\ &= \sum_i \frac{\partial f_i}{\partial w_k} [y_i - E_{y' \sim p_c^{PL}} [y'_i]] \end{aligned} \quad (7)$$

In the following of this section, it is assumed that all potentials are approximated using the piecewise local approximation to keep the notation simple, the extension to the pseudo likelihood is trivial, it is enough to replace the  $p_c^{PL}$  with the  $p_c^{PPL}$  in Equation 6.

**Training for neuro-symbolic integration.** An interesting case is when a potential represents the level of satisfaction of a logical constraint over its groundings in the world  $\mathbf{y}$ . In this case the predicates of the  $c$ -th formula are grounded with a set of  $\mathbf{x}_c$  groundings, and  $G_c(\mathbf{y})$  indicates the set of outputs for the grounded predicates in the world  $\mathbf{y}$ . Therefore, the potential is the sum over the grounded formulas:

$$\Phi_c(\mathbf{y}) = \sum_{\mathbf{y}_{c,g} \in G_c(\mathbf{y})} \phi_c(\mathbf{y}_{c,g}),$$

where  $\phi_c(\mathbf{y}_{c,g}) \in \{0, 1\}$  is the satisfaction of the formula (False or True) by the grounded predicates  $\mathbf{y}_{c,g}$ . Therefore, each grounding corresponds to a separate potential, even if they are all sharing the same weight. Assuming that each grounding of a formula is independent on all the others, then we can approximate the  $Z_c$  as:

$$Z_c \approx Z_c^{l|G_c(\mathbf{y})|} = \left( \sum_{\mathbf{y}'_{c,g}} \exp(\lambda_c \phi_c(\mathbf{y}'_{c,g})) \right)^{|G_c(\mathbf{y})|}$$

op \ t-norm	Product	Gödel	Lukaseiwicz
$x \wedge y$	$x \cdot y$	$\min(x, y)$	$\max(0, x + y - 1)$
$x \vee y$	$x + y - x \cdot y$	$\max(x, y)$	$\min(1, x + y)$
$\neg x$	$1 - x$	$1 - x$	$1 - x$
$x \Rightarrow y$	$x \leq y?1 : \frac{y}{x}$	$x \leq y?1 : y$	$\min(1, 1 - x + y)$

**Table 2.** The algebraic operations corresponding to primary logical connectives for the fundamental t-norm fuzzy logics.

where  $|G_c(\mathbf{y})|$  are the total number of groundings of the  $c$ -th formula in  $\mathbf{y}$  and each grounded formula shares the same local partition function  $Z_c^l$ .  $Z_c^l$  can be efficiently computed by pre-computing  $n_c^+$ ,  $n_c^-$ , indicating the number of possible different grounding assignments satisfying or not satisfying the  $c$ -th formula. Clearly, since for a formula with  $n_c$  atoms, there are  $2^{n_c}$  possible assignments, it holds that  $n_c^- + n_c^+ = 2^{n_c}$ , yielding:

$$\begin{aligned} Z_c^l &= \sum_{\mathbf{y}'_{c,g}} \exp(\lambda_c \phi_c(\mathbf{y}'_{c,g})) = \\ &= \underbrace{n_c^-}_{\text{each False evaluates to } e^0=1} + \underbrace{n_c^+ e^{\lambda_c}}_{\text{each True evaluates to } e^{\lambda_c}} = n_c^- + n_c^+ e^{\lambda_c} \end{aligned}$$

Using the piecewise local approximation for each grounding, the derivatives with respect of the model parameters become:

$$\begin{aligned} \frac{\partial \log p(\mathbf{y}|\mathbf{f}, \boldsymbol{\lambda})}{\partial \lambda_c} &\approx \sum_{\mathbf{y}_{c,g}} \left[ \phi_c(\mathbf{y}_{c,g}) - \prod_{\mathbf{y}'_{c,g}} p_{c,g}^{PL}(\mathbf{y}'_{c,g}) \phi_c(\mathbf{y}'_{c,g}) \right] = \\ &= \sum_{\mathbf{y}_{c,g}} \phi_c(\mathbf{y}_{c,g}) - |G_c(\mathbf{y})| \cdot \mathbb{E}_{p_{c,g}^{PL}}[\phi_c] \end{aligned}$$

Let us indicate as  $Avg(\Phi_c, \mathbf{y}) = \frac{1}{|G_c(\mathbf{y})|} \sum_{\mathbf{y}_{c,g} \in G_c(\mathbf{y})} \phi_c(\mathbf{y}_{c,g})$  the average satisfaction of the  $c$ -th constraint over the data training data, then the gradient is null when for all constraints:

$$Avg(\Phi_c, \mathbf{y}) = \mathbb{E}_{p_{c,g}^{PL}}[\phi_c] \quad (8)$$

The expected value of the satisfaction of the formula for a grounding,  $\mathbb{E}_{p_{c,g}^{PL}}[\phi_c]$  can be efficiently computed for a  $\{0, 1\}$ -valued  $\phi_c$  as:

$$\begin{aligned} \mathbb{E}_{p_{c,g}^{PL}}[\phi_c] &= \sum_{\mathbf{y}'_{c,g}} p_{c,g}^{PL}(\mathbf{y}'_{c,g}) \phi_c(\mathbf{y}'_{c,g}) = \\ &= \sum_{\mathbf{y}'_{c,g}} \frac{1}{Z_c^l} \exp(\lambda_c \phi_c(\mathbf{y}'_{c,g})) \phi_c(\mathbf{y}'_{c,g}) = \\ &= \frac{n_c^+ e^{\lambda_c}}{Z_c^l} = \frac{n_c^+ e^{\lambda_c}}{n_c^- + n_c^+ e^{\lambda_c}} \end{aligned}$$

yielding the following optimal assignment to the  $c$ -th parameter for a given assignment  $\mathbf{y}$ :

$$\lambda_c = \log \frac{Avg(\Phi_c, \mathbf{y})}{1 - Avg(\Phi_c, \mathbf{y})} - \log \frac{n_c^+}{n_c^-} \quad (9)$$

which shows that the log-likelihood is maximized by selecting a  $\lambda_c$  equal to difference between the log odds of the constraint satisfaction of the data and the log odds of the prior satisfaction of the constraint if all assignments are equally probable.

**Data:** Input data  $x$ , output variables  $\mathbf{y}$ , training and observed data  $\mathbf{y}^T$ , function models with weights  $\mathbf{w}$

**Result:** Trained model parameters  $\{\boldsymbol{\lambda}, \mathbf{w}\}$

Initialize  $i = 0$ ,  $\boldsymbol{\lambda} = \mathbf{0}$ , random  $\mathbf{w}$ ;

**while** not converged  $\wedge i < max\_iterations$  **do**

**Expectation:**

    Compute function outputs  $\mathbf{f}$  on  $x$  using weights  $\mathbf{w}$ ;

    Compute MAP solution using Equation 10

$\mathbf{y}^* = \operatorname{argmax}_{\mathbf{y}} \log p(\mathbf{y}|\mathbf{f}, \mathbf{y}^T, \boldsymbol{\lambda})$ ;

**Maximization:**

$\forall c$  compute constraint satisfaction  $Avg(\Phi_c, \mathbf{y} \cup \mathbf{y}^T)$ ;

$\forall c$  compute  $\lambda_c$  using Equation 9 on  $\mathbf{y} \cup \mathbf{y}^T$ ;

    Backpropagation with respect of the  $\mathbf{w}$  weights using derivatives from Equation 7;

    Update  $\mathbf{w}$ ;

  Set  $i=i+1$ ;

**end**

**Algorithm 1:** Iterative algorithm to train the function weights  $\mathbf{w}$  and the constraint weights  $\boldsymbol{\lambda}$ .

When the world is fully observed during training,  $\mathbf{y}^T$  indicates the training data assignments, then substituting  $\mathbf{y} = \mathbf{y}^T$  into Equation 9 returns the maximum likelihood assignment for the parameters.

When the world is not fully observed during training, an iterative EM schema can be used to marginalize over the unobserved data in the expectation step using the inference methodology as described in the next paragraph. Then, the average constraint satisfaction can be recomputed, and then the  $\boldsymbol{\lambda}$ ,  $\mathbf{w}$  parameters can be updated in the maximization step. This process is then iterated until convergence. Algorithm 1 reports the complete training algorithm for RNMs.

**Inference.** The MAP inference process searches the most probable assignment of the  $\mathbf{y}$  given the evidence and the fixed parameters  $\mathbf{w}$ ,  $\boldsymbol{\lambda}$ . The problem of finding the best assignment  $\mathbf{y}^*$  to the unobserved query variables given the evidence  $\mathbf{y}^e$  and current parameters can be stated as:

$$\mathbf{y}^* = \operatorname{argmax}_{\mathbf{y}'} \sum_c \lambda_c \Phi_c(\mathbf{f}, [\mathbf{y}', \mathbf{y}^e]) \quad (10)$$

where  $[\mathbf{y}', \mathbf{y}^e]$  indicates a full assignment to the  $\mathbf{y}$  variables, split into the query and evidence sets.

Gradient-based techniques can not be readily used to optimize the MAP problem stated by Equation 10, since the problem is discrete. A possible solution could be to relax the  $\mathbf{y}$  values into the  $[0, 1]$  interval and assume that each potential  $\Phi_c(\mathbf{f}, \mathbf{y})$  has a continuous surrogate  $\Phi_c^s(\mathbf{f}, \mathbf{y})$  which collapses into the original potential when the  $\mathbf{y}$  assume crisp values and is continuous with respect to each  $y_i$ . As described in the following, continuous surrogates are very appropriate to describe the potentials representing logic knowledge for neuro-symbolic integration and probabilistic logic reasoning.

When relaxing the potentials to accept continuous input variables, the MAP solution can be found by gradient-based techniques by computing the derivative with respect of each output variable:

$$\frac{\partial \log p(\mathbf{y}' | \mathbf{y}^e, \mathbf{f}, \boldsymbol{\lambda})}{\partial y_k} = f_k + \sum_c \lambda_c \frac{\partial \Phi_c^s(\mathbf{f}, [\mathbf{y}', \mathbf{y}^e])}{\partial y_k} \quad (11)$$

**T-Norm Fuzzy Logic.** Fuzzy logics extend the set of Boolean logic truth-values  $\{0, 1\}$  to the unit interval  $[0, 1]$  and, as a consequence, they can be exploited to convert Boolean logic expressions into continuous and differentiable ones. In particular, a t-norm fuzzy logic [14] is defined upon the choice of a certain t-norm [22]. A t-norm is a binary operation generalizing to continuous values the Boolean logic conjunction ( $\wedge$ ), while it recovers the classical AND when the variables assume the crisp values 0 (false) or 1 (true).

Throughout this paper, we assume that given a certain variable  $a$  assuming a continuous value  $a \in [0, 1]$ , its negation  $\neg a$  (also said strong negation) is evaluated as  $1 - a$ . Moreover, a t-norm and the strong negation allows the definition of additional logical connectives. For instance, the implication ( $\Rightarrow$ ) may be defined as the residuum of the t-norm, while the OR ( $\vee$ ) operator, also called t-conorm, may be defined according to the DeMorgan law with respect to the t-norm and the strong negation.

$$a \Rightarrow b = \sup\{c : a \wedge c \leq b\}; \quad a \vee b = \neg(\neg a \wedge \neg b).$$

Different t-norm fuzzy logics have been proposed in the literature. Table 2 reports the operations computed by different logic operators for the three fundamental continuous t-norms, i.e. Product, Gödel and Łukasiewicz logics. Furthermore, a fragment of the Łukasiewicz logic [12] has been recently proposed for translating logic inference into a differentiable optimization problem, since it defines a large class of clauses which are translated into convex functions.

An important role defining different ways to aggregate logical propositions on (possibly large) sets of domain variables is played by quantifiers. The universal quantifier ( $\forall$ ) and the existential quantifier ( $\exists$ ) express the fact that a clause should hold true over all or at least one grounding. Both the universal and existential quantifier are generally converted into real-valued functions according to different aggregation functions, e.g. the universal one as a t-norm and the existential one as a t-conorm over the groundings. When multiple universally or existentially quantified variables are present, the conversion is recursively performed from the outer to the inner variables as already stated. For example, consider the rule

$$\forall x A(x) \vee (B(x) \wedge \neg C(x))$$

where  $A, B, C$  are three unary predicates defined on the input set  $\{x_1, \dots, x_m\}$ . In this case, the output vector  $\mathbf{y}$  is defined as follows,

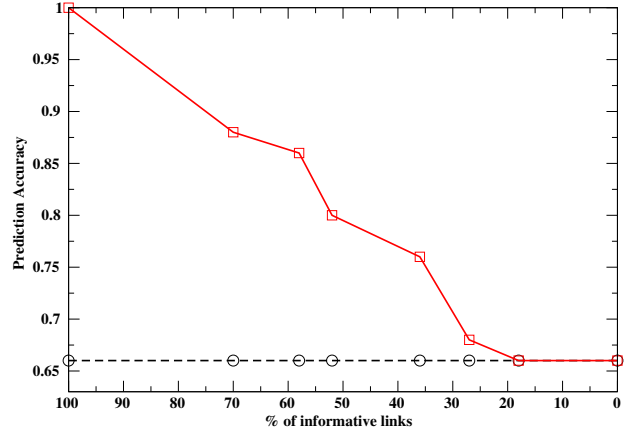
$$[y_A(x_1), \dots, y_A(x_m), y_B(x_1), \dots, y_B(x_m), y_C(x_1), \dots, y_C(x_m)]$$

where  $y_P(x_i)$  is the output of predicate  $P$  when grounded with  $x_i$ . The continuous surrogate for the FOL rule grounded over all patterns in the domain, in case of the product t-norm and universal quantifier converted with the arithmetic mean, is given by:

$$\Phi^s(\mathbf{y}) = \frac{1}{m} \sum_{i=1}^m y_A(x_i) + (1 - y_A(x_i)) \cdot y_B(x_i) \cdot (1 - y_C(x_i)).$$

## 4 Experiments

The proposed model has been experimentally evaluated on two different datasets where the relational structure on the output or input data may be exploited.



**Figure 2.** Prediction accuracy with respect to the percentage of links that correctly predict the next digit condition.

### 4.1 MNIST Following Pairs

This small toy task is designed to highlight the capability of RNMs to learn and employ soft rules that are holding only for a sub-portion of the whole dataset. The MNIST dataset contains images of handwritten digits, and this task assumes that additional relational logic knowledge is available to reason over the digits. In particular, given a certain subset of images, a binary predicate *link* between image pairs is considered. Given two images  $x, y$ , whose corresponding digits are denoted by  $i, j$ , a link between  $x$  and  $y$  is established if the second digit follows the first one, i.e.  $i = j + 1$ . However, it is assumed that the *link* predicate is noisy, therefore for  $i \neq j + 1$ , there is a given degree of probability that the *link*( $x, y$ ) is established anyway. The knowledge about the *link* predicate can be represented by the following FOL formula

$$\forall x \forall y \forall i \forall j \text{link}(x, y) \wedge \text{digit}(x, i) \wedge \text{digit}(y, j) \Rightarrow i = j + 1,$$

where *digit*( $x, i$ ) is a binary predicate indicating if a number  $i$  is the digit class of the image  $x$ . Since the *link* predicate holds true also for pairs of non-consecutive digits, the above rule is violated by a certain percentage of digit pairs. Therefore, the manifold established by the *link* predicate can help in driving the prediction, but the noisy links force the reasoner to be flexible about how to employ the knowledge.

The training set is created by randomly selecting 50 images from the MNIST dataset and by adding the *link* relation with an incremental degree of noise. For each degree of noise in the training set, we created an equally sized test set with the same degree of noise. A neural network with 100 hidden sigmoid neurons is used to process the input images.

Figure 2 reports a comparison between RNM and the baseline provided by the neural network varying the percentage of links that are predictive of a digit to follow another one. When the link predicate only holds for consecutive digit pairs, RNM is able to perfectly predict the images on the test set using this information. When the link becomes less informative (more noisy), RNM is still able to employ the rule as a soft suggestion. However, when the percentage of predictive links approaches 10%, the relation is not informative at all, as it does not add any information on top of the prior probability that two randomly picked up numbers follow each other. In this case, RNM is still able to detect that the formula is not useful, and only the supervised data is used to learn the predictions. As a result, the

predictive accuracy of RNM matches the one of the neural network. Finally, the time overhead of the method for this dataset is less than 20% over pure-supervised learning.

## 4.2 Document Classification on the Citeseer dataset.

The CiteSeer dataset [25] is a collection of 3312 scientific papers, each one assigned to one of the 6 classes: *AG*, *AI*, *DB*, *IR*, *ML* and *HCI*. The papers connect to each other by a citation network which contains 4732 links. Each paper in the dataset is described via its bag-of-words, e.g. a vector where the  $i$ -th element has a value equal to 1 or 0, depending on whether the  $i$ -th word in the vocabulary is present or not present in the document, respectively. The overall dictionary for this experiment contains 3703 unique words. The domain knowledge used for this task state that connected papers  $p_1, p_2$  tend to be about the same topic:

$$\begin{aligned} \forall p_1 \forall p_2 \text{ AG}(p_1) \wedge \text{Cite}(p_1, p_2) &\rightarrow \text{AG}(p_2) \\ \forall p_1 \forall p_2 \text{ AI}(p_1) \wedge \text{Cite}(p_1, p_2) &\rightarrow \text{AI}(p_2) \\ \forall p_1 \forall p_2 \text{ DB}(p_1) \wedge \text{Cite}(p_1, p_2) &\rightarrow \text{DB}(p_2) \\ \forall p_1 \forall p_2 \text{ IR}(p_1) \wedge \text{Cite}(p_1, p_2) &\rightarrow \text{IR}(p_2) \\ \forall p_1 \forall p_2 \text{ ML}(p_1) \wedge \text{Cite}(p_1, p_2) &\rightarrow \text{ML}(p_2) \\ \forall p_1 \forall p_2 \text{ HCI}(p_1) \wedge \text{Cite}(p_1, p_2) &\rightarrow \text{HCI}(p_2) \end{aligned}$$

where  $\text{Cite}(\cdot, \cdot)$  is an evidence predicate (e.g. its value over the groundings is known a-priori) determining whether a pattern cites another one. Different topics are differently closed with respect to other fields, and the above rules hold with different degrees.

A neural network with three hidden layers with 50 units and RELU activation functions and one output layer using the softmax activation is used for this task as baseline. RNM employs the same network but with no output layer as the output layer is computed as part of the inference process as shown in Section 2 for the one and multi-label classification cases. The Adam optimizer [21] is used to update the weights. The RNM framework develops the embeddings as part of training, when the logical knowledge is distilled into the network weights by enforcing the consistency between the reasoner and network outputs.

A variable portion of the data is sampled for training, of which 10% of this data is kept as validation set, while the remaining data is used as test set.

% training data	NN baseline	SBR	RNM
90	0.723	0.726	0.732
75	0.717	0.719	0.726
50	0.707	0.712	0.726
25	0.674	0.682	0.709
10	0.645	0.650	0.685

**Table 3.** Fully Observed Case. Results on the Citeseer dataset when using a subset of the supervised data for training using RNM, SBR and the baseline neural network.

**Fully Observed Case.** The train and test sets are kept separated, and all links between train and test papers are dropped, so that the train and test data are two separate worlds. Table 3 reports the result obtained by the baseline neural network, compared against the baseline model and SBR trained using the Lyrics framework as average over ten different samples of the train and test data. Since SBR can

not learn the weight of the rules, these are validated by selecting the best performing one on the validation set. RNM improves over the other methods for all tested configurations, thanks to its ability of selecting the best weights for each rule, exploiting the fact that each research community has a different intra-community citation rate.

% training data	NN baseline	SBR	RNM
90	0.726	0.780	0.780
75	0.708	0.764	0.766
50	0.695	0.747	0.753
25	0.667	0.729	0.735
10	0.640	0.703	0.708

**Table 4.** Partially Observed Case. Results on the Citeseer dataset when using a subset of the supervised data for training using RNM, SBR and the baseline neural network.

**Partially Observed Case.** This experiment assumes that the training, validation and test data are available at training time [10], even if only the training labels are used during the training process. This configuration models a real world scenario where a partial knowledge of a world is given, but it is required to perform inference over the unknown portion of the environment. In this transductive experiment, all Citeseer papers are supposed to be available in a single world together with the full citation network. Only a variable percentage of the supervised data is used during training. Therefore, the world is only partially observed at training time, and the EM schema described by Algorithm 1 must be used during training.

Table 4 reports the accuracy results obtained by the baseline neural network, compared against the baseline model and SBR. The SBR weights are validated by selecting the best performing one on the validation set. RNM improves over the other methods for all tested configurations, thanks to its ability of selecting the best weights for each rule.

The overhead of the method on this dataset is around 100% over pure-supervised learning, mostly because of the time needed to build the fully grounded networks. However, there is no significant overhead over other (less general and powerful) fully-grounded neuro-symbolic methods like LTN or SBR.

## 5 Conclusions and Future Work

This paper presented Relational Neural Machines a novel framework to provide a tight integration between learning from supervised data and logic reasoning, allowing to improve the quality of both modules processing the low-level input data and the high-level reasoning about the environment. The presented model provides significant advantages over previous work in terms of scalability and flexibility, while dropping any trade-off in exploiting the supervised data. The preliminary experimental results are promising, showing that the tighter integration between symbolic and a sub-symbolic levels helps in exploiting the input and output structures. As future work, we plan to undertake a larger experimental exploration of RNM on real world problems for more structured problems.

## ACKNOWLEDGEMENTS

This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 825619.

## REFERENCES

- [1] Miltiadis Allamanis, Pankajan Chanthirasegaran, Pushmeet Kohli, and Charles Sutton, ‘Learning continuous semantic representations of symbolic expressions’, in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 80–88. JMLR. org, (2017).
- [2] Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor, ‘Hinge-loss markov random fields and probabilistic soft logic’, *Journal of Machine Learning Research*, **18**, 1–67, (2017).
- [3] Liang-Chieh Chen, Alexander Schwing, Alan Yuille, and Raquel Urtasun, ‘Learning deep structured models’, in *International Conference on Machine Learning*, pp. 1785–1794, (2015).
- [4] Luc De Raedt, Angelika Kimmig, and Hannu Toivonen, ‘Problog: A probabilistic prolog and its application in link discovery’, in *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI’07*, pp. 2468–2473, San Francisco, CA, USA, (2007). Morgan Kaufmann Publishers Inc.
- [5] Michelangelo Diligenti, Marco Gori, and Claudio Sacca, ‘Semantic-based regularization for learning and inference’, *Artificial Intelligence*, **244**, 143–165, (2017).
- [6] I Donadello, L Serafini, and AS d’Avila Garcez, ‘Logic tensor networks for semantic image interpretation’, in *IJCAI International Joint Conference on Artificial Intelligence*, pp. 1596–1602, (2017).
- [7] Honghua Dong, Jiayuan Mao, Tian Lin, Chong Wang, Lihong Li, and Denny Zhou, ‘Neural logic machines’, in *International Conference on Learning Representations*, (2019).
- [8] Sebastijan Dumančić and Hendrik Blockeel, ‘Demystifying relational latent representations’, in *International Conference on Inductive Logic Programming*, pp. 63–77. Springer, (2017).
- [9] Sebastijan Dumančić, Tias Guns, Wannes Meert, and Hendrik Blockeel, ‘Learning relational representations with auto-encoding logic programs’, in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 6081–6087. AAAI Press, (2019).
- [10] Alexander Gammerman, Volodya Vovk, and Vladimir Vapnik, ‘Learning by transduction’, in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pp. 148–155. Morgan Kaufmann Publishers Inc., (1998).
- [11] Artur S d’Avila Garcez, Krysia B Broda, and Dov M Gabbay, *Neural-symbolic learning systems: foundations and applications*, Springer Science & Business Media, 2012.
- [12] Francesco Giannini, Michelangelo Diligenti, Marco Gori, and Marco Maggini, ‘On a convex logic fragment for learning and reasoning’, *IEEE Transactions on Fuzzy Systems*, (2018).
- [13] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio, *Deep learning*, volume 1, MIT press Cambridge, 2016.
- [14] Petr Hájek, *Metamathematics of fuzzy logic*, volume 4, Springer Science & Business Media, 2013.
- [15] Tamir Hazan, Alexander G Schwing, and Raquel Urtasun, ‘Blending learning and inference in conditional random fields’, *The Journal of Machine Learning Research*, **17**(1), 8305–8329, (2016).
- [16] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard H. Hovy, and Eric P. Xing, ‘Harnessing deep neural networks with logic rules’, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, (2016).
- [17] Shoaib Jameel and Steven Schockaert, ‘Entity embeddings with conceptual subspaces as a basis for plausible reasoning’, in *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pp. 1353–1361. IOS Press, (2016).
- [18] William W Cohen Fan Yang Kathryn and Rivard Mazaitis, ‘Tensorlog: Deep learning meets probabilistic databases’, *Journal of Artificial Intelligence Research*, **1**, 1–15, (2018).
- [19] Navdeep Kaur, Gautam Kunapuli, Tushar Khot, Kristian Kersting, William Cohen, and Sriraam Natarajan, ‘Relational restricted boltzmann machines: A probabilistic logic learning approach’, in *International Conference on Inductive Logic Programming*, pp. 94–111. Springer, (2017).
- [20] Angelika Kimmig, Stephen Bach, Matthias Broecheler, Bert Huang, and Lise Getoor, ‘A short introduction to probabilistic soft logic’, in *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, pp. 1–4, (2012).
- [21] Diederik P Kingma and Jimmy Ba, ‘Adam: A method for stochastic optimization’, *arXiv preprint arXiv:1412.6980*, (2014).
- [22] E.P. Klement, R. Mesiar, and E. Pap, *Triangular Norms*, Kluwer Academic Publisher, 2000.
- [23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, ‘Deep learning’, *nature*, **521**(7553), 436, (2015).
- [24] Marco Lippi and Paolo Frasconi, ‘Prediction of protein  $\beta$ -residue contacts by markov logic networks with grounding-specific weights’, *Bioinformatics*, **25**(18), 2326–2333, (2009).
- [25] Qing Lu and Lise Getoor, ‘Link-based classification’, in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 496–503, (2003).
- [26] Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt, ‘Deepproblog: Neural probabilistic logic programming’, *arXiv preprint arXiv:1805.10872*, (2018).
- [27] Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, and Marco Gori, ‘Integrating learning and reasoning with deep logic models’, in *Proceedings of the European Conference on Machine Learning*, (2019).
- [28] Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, and Marco Gori, ‘Lyrics: a general interface layer to integrate ai and deep learning’, *arXiv preprint arXiv:1903.07534*, (2019).
- [29] Giuseppe Marra and Ondřej Kuželka, ‘Neural markov logic networks’, *arXiv preprint arXiv:1905.13462*, (2019).
- [30] Pasquale Minervini, Luca Costabello, Emir Muñoz, Vít Nováček, and Pierre-Yves Vandenbussche, ‘Regularizing knowledge graph embeddings via equivalence and inversion axioms’, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 668–683. Springer, (2017).
- [31] Stephen Muggleton and Luc De Raedt, ‘Inductive logic programming: Theory and methods’, *The Journal of Logic Programming*, **19**, 629–679, (1994).
- [32] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio, ‘Holographic embeddings of knowledge graphs’, in *Thirtieth Aaai conference on artificial intelligence*, (2016).
- [33] Mathias Niepert, ‘Discriminative gaifman models’, in *Advances in Neural Information Processing Systems*, pp. 3405–3413, (2016).
- [34] Matthew Richardson and Pedro Domingos, ‘Markov logic networks’, *Machine learning*, **62**(1), 107–136, (2006).
- [35] Tim Rocktäschel and Sebastian Riedel, ‘Learning knowledge base inference with neural theorem provers’, in *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pp. 45–50, (2016).
- [36] Tim Rocktäschel and Sebastian Riedel, ‘End-to-end differentiable proving’, in *Advances in Neural Information Processing Systems*, pp. 3788–3800, (2017).
- [37] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap, ‘A simple neural network module for relational reasoning’, in *Advances in neural information processing systems*, pp. 4967–4976, (2017).
- [38] Luciano Serafini, Ivan Donadello, and Artur d’Avila Garcez, ‘Learning and reasoning in logic tensor networks: theory and application to semantic image interpretation’, in *Proceedings of the Symposium on Applied Computing*, pp. 125–130. ACM, (2017).
- [39] Gustav Sourek, Vojtech Aschenbrenner, Filip Zelezny, Steven Schockaert, and Ondrej Kuzelka, ‘Lifted relational neural networks: Efficient learning of latent relational structures’, *Journal of Artificial Intelligence Research*, **62**, 69–100, (2018).
- [40] Charles Sutton and Andrew McCallum, ‘Piecewise pseudolikelihood for efficient training of conditional random fields’, in *Proceedings of the 24th international conference on Machine learning*, pp. 863–870. ACM, (2007).
- [41] Charles Sutton, Andrew McCallum, et al., ‘An introduction to conditional random fields’, *Foundations and Trends® in Machine Learning*, **4**(4), 267–373, (2012).
- [42] Quan Wang, Bin Wang, and Li Guo, ‘Knowledge base completion using embeddings and rules’, in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, (2015).
- [43] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck, ‘A semantic loss function for deep learning with symbolic knowledge’, in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, (July 2018).
- [44] Fan Yang, Zhilin Yang, and William W Cohen, ‘Differentiable learning of logical rules for knowledge base reasoning’, in *Advances in Neural Information Processing Systems*, pp. 2319–2328, (2017).