

# DeepProbLog: Neural Probabilistic Logic Programming<sup>☆</sup>

Robin Manhaeve<sup>a,\*</sup>, Sebastijan Dumančić<sup>a</sup>, Angelika Kimmig<sup>b</sup>, Thomas Demeester<sup>c,1</sup>, Luc De Raedt<sup>a,1</sup>

<sup>a</sup>*KU Leuven*

<sup>b</sup>*Cardiff University*

<sup>c</sup>*Ghent University - imec*

---

## Abstract

We introduce DeepProbLog, a neural probabilistic logic programming language that incorporates deep learning by means of neural predicates. We show how existing inference and learning techniques of the underlying probabilistic logic programming language ProbLog can be adapted for the new language. We theoretically and experimentally demonstrate that DeepProbLog supports (i) both symbolic and subsymbolic representations and inference, (ii) program induction, (iii) probabilistic (logic) programming, and (iv) (deep) learning from examples. To the best of our knowledge, this work is the first to propose a framework where general-purpose neural networks and expressive probabilistic-logical modeling and reasoning are integrated in a way that exploits the full expressiveness and strengths of both worlds and can be trained end-to-end based on examples.

*Keywords:* logic, probability, neural networks, probabilistic logic programming, neuro-symbolic integration, learning and reasoning

---

## 1. Introduction

Many tasks in AI can be divided into roughly two categories: those that require low-level perception, and those that require high-level reasoning. At the same time, there is a growing consensus that being capable of tackling both types of tasks is essential to achieve true (artificial) intelligence [2]. Deep learning is empowering a new generation of intelligent systems that excel at low-level perception, where it is used to interpret images, text and speech with unprecedented accuracy. The success of deep learning has caused a lot of excitement and has also created the impression that deep learning can solve any problem in

---

<sup>☆</sup>This is an extended and revised version of work previously published at NeurIPS 2018 [1].

<sup>\*</sup>Corresponding author

*Email address:* `robin.manhaeve@cs.kuleuven.be` (Robin Manhaeve)

<sup>1</sup>Joint last authors.

artificial intelligence. However, there is a growing awareness of the limitations of deep learning: deep learning requires large amounts of (the right kind of) data to train the network, it provides neither justifications nor explanations, and the models are black-boxes that can neither be understood nor modified by domain experts. Although there have been attempts to demonstrate reasoning-like behaviour with deep learning [3], their current reasoning abilities are nowhere close to what is possible with typical high-level reasoning approaches. The two most prominent frameworks for reasoning are logic and probability theory. While in the past, these were studied by separate communities in artificial intelligence, many researchers are working towards their integration, and aim at combining probability with logic and statistical learning; cf. the areas of statistical relational artificial intelligence [4, 5] and probabilistic logic programming [6].

The abilities of deep learning and statistical relational artificial intelligence approaches are complementary. While deep learning excels at low-level perception, probabilistic logics excel at high-level reasoning. As such, an integration of the two would have very promising properties. Recently, a number of researchers have revisited and modernized ideas originating from the field of neural-symbolic integration [7], searching for ways to combine the best of both worlds [8, 9, 10, 3], for example, by designing neural architectures representing differentiable counterparts of symbolic operations in classical reasoning tools. Yet, joining the full flexibility of high-level probabilistic and logical reasoning with the representational power of deep neural networks is still an open problem. Elsewhere [11], we have argued that neuro-symbolic integration should: 1) integrate neural networks with the two most prominent methods for reasoning, that is, logic and probability, and 2) that neuro-symbolic integrated methods should have the pure neural, logical and probabilistic methods as special cases.

With DeepProbLog, we tackle the neuro-symbolic challenge from this perspective. Furthermore, instead of integrating reasoning capabilities into a complex neural network architecture, we proceed the other way round. We start from an existing probabilistic logic programming language, ProbLog [12], and introduce the smallest extension that allows us to integrate neural networks: the neural predicate. The idea is simple: in a probabilistic logic, atomic expressions of the form  $q(t_1, \dots, t_n)$  (aka tuples in a relational database) have a probability  $p$ . We extend this idea by allowing atomic expressions to be labeled with neural networks whose outputs can be considered probability distributions. This simple idea is appealing as it allows us to retain all the essential components of the ProbLog language: the semantics, the inference mechanism, as well as the implementation.

Therefore, *one should not only integrate logic with neural networks in neuro-symbolic computation, but also probability.*

This effectively leads to an integration of probabilistic logics (hence statistical relational AI) with neural networks and opens up new abilities. Furthermore, although at first sight, this may appear as a complication, it actually can greatly simplify the integration of neural networks with logic. The reason for this is that the probabilistic framework provides a clear optimisation criterion, namely the probability of the training examples. Real-valued probabilistic quantities are

also well-suited for gradient-based training procedures, as opposed to discrete logic quantities.

### Example 1

Before going into further detail, the following example illustrates the possibilities of this approach. Consider the predicate `addition(X, Y, Z)`, where `X` and `Y` are images of digits and `Z` is the natural number corresponding to the sum of these digits. The goal is that after training, DeepProbLog allows us to make a probabilistic estimate on the validity of, for example, the example `addition(3, 5, 8)`. While such a predicate can be learned directly by a standard neural classifier, such an approach cannot incorporate background knowledge such as the definition of the addition of two *natural* numbers. In DeepProbLog such knowledge can easily be encoded in rules such as

$$\text{addition}(I_X, I_Y, N_Z) :- \text{digit}(I_X, N_X), \text{digit}(I_Y, N_Y), N_Z \text{ is } N_X + N_Y$$

with `is` the standard operator of logic programming to evaluate arithmetic expressions. All that needs to be learned in this case is the neural predicate *digit* which maps an image of a digit  $I_D$  to the corresponding natural number  $N_D$ . The trained network can then be reused for arbitrary tasks involving digits. Our experiments show that this leads not only to new capabilities but also to significant performance improvements. An important advantage of this approach compared to standard image classification settings is that it can be extended to multi-digit numbers without additional training. We note that the single digit classifier (i.e., the neural predicate) is not explicitly trained by itself: its output can be considered a latent representation, as we only use training data with pairwise sums of digits.

To summarize, we introduce DeepProbLog which has a unique set of features: (i) it is a programming language that supports neural networks and machine learning and has a well-defined semantics (ii) it integrates logical reasoning with neural networks; so both symbolic and subsymbolic representations and inference; (iii) it integrates probabilistic modeling, programming and reasoning with neural networks (as DeepProbLog extends the probabilistic programming language ProbLog, which can be regarded as a very expressive directed graphical modeling language [4]); (iv) it can be used to learn a wide range of probabilistic logical neural models from examples, including inductive programming.

This paper is a significantly extended and completed version of our previous work [1] (NeurIPS, spotlight presentation). This extended version now contains the necessary deep learning and probabilistic logic programming background and a more in depth theoretical explanation. It also contains additional experiments (see Section 6): the MNIST addition experiments from the short version are completed with the new experiments **T3** and **T4**, and we designed new experiments (**T8** and **T9**) to further investigate the use of DeepProbLog on combined probabilistic learning and deep learning. The code is available at <https://bitbucket.org/problog/deepproblog>.

## 2. Background

### 2.1. Logic programming concepts

In this section, we briefly summarize basic logic programming concepts; see e.g., Lloyd [13] for more details. Atoms are expressions of the form  $q(t_1, \dots, t_n)$  where  $q$  is a predicate (of arity  $n$ , or  $q/n$  in shorthand notation) and the  $t_i$  are terms. A literal is an atom or the negation  $\neg q(t_1, \dots, t_n)$  of an atom. A term  $t$  is either a constant  $c$ , a variable  $V$ , or a structured term of the form  $f(u_1, \dots, u_k)$  where  $f$  is a functor and the  $u_i$  are terms. We follow the Prolog convention and let constants, functors and predicates start with a lower case character and variables with an upper case. A rule is an expression of the form  $h :- b_1, \dots, b_n$  where  $h$  is an atom, the  $b_i$  are literals, and all variables are universally quantified. Informally, the meaning of such a rule is that  $h$  holds whenever the conjunction of the  $b_i$  holds. Thus  $:-$  represents logical implication ( $\leftarrow$ ), and the comma  $(,)$  represents conjunction ( $\wedge$ ). Rules with an empty body  $n = 0$  are called facts. A logic program is a finite set of rules.

A substitution  $\theta = \{V_1 = t_1, \dots, V_n = t_n\}$  is an assignment of terms  $t_i$  to variables  $V_i$ . When applying a substitution  $\theta$  to an expression  $e$  we simultaneously replace all occurrences of  $V_i$  by  $t_i$  and denote the resulting expression as  $e\theta$ . Expressions that do not contain any variables are called ground. The *Herbrand base* of a logic program is the set of ground atoms that can be constructed using the predicates, functors and constants occurring in the program.<sup>2</sup> Subsets of the Herbrand base are called *Herbrand interpretations*. A Herbrand interpretation is a *model* of a clause  $h :- b_1, \dots, b_n$  if for every substitution  $\theta$  such that the conjunction  $(b_1, \dots, b_n)\theta$  holds in the interpretation,  $h\theta$  is in the interpretation.

It is a model of a logic program if it is a model of all clauses in the program.

For negation-free programs, the semantics is given by the minimal such model, known as the least Herbrand model, which is unique. General logic programs use the notion of negation as failure, that is, the negation of an atom is true exactly if the atom cannot be derived from the program. These programs are not guaranteed to have a unique minimal Herbrand model, and several ways to define a canonical model have been studied. We follow the well-founded semantics here [14].

The main inference task in logic programming is to determine whether a given atom  $q$ , also called *query* (or *goal*), is true in the canonical model of a logic program  $P$ , denoted by  $P \models q$ . If the answer is yes (or no), we also say that the query *succeeds* (or *fails*). If such a query is not ground, inference asks for the existence of an *answer substitution*, that is, a substitution that grounds the query into an atom that is part of the canonical model.

### 2.2. Deep Learning

The following paragraphs provide a very brief introduction to deep learning, focusing on concepts needed for understanding our work. Extensive further

---

<sup>2</sup>If the program does not contain constants, one arbitrary constant is added.

details can be found, e.g., in [15]. This section is meant to provide readers with little or no knowledge of deep learning, with a conceptual understanding of the main ideas. In particular, we will focus on the setting of supervised learning, where the model learns to map an input item to a particular output, based on input-output examples.

An artificial neural network is a highly parameterized and therefore very flexible non-linear mathematical function that can be ‘trained’ towards a particular desired behavior, by suitably adjusting its parameters.

During training, the model learns to capture from the input data the most informative ‘features’ for the task at hand. The need for ‘feature engineering’ in classical (or rather, non-neural) machine learning methods has therefore been replaced by ‘architecture engineering’, since a wide variety of neural network components are available to be composed into a suitable model.

Deep neural networks are often designed and trained in an ‘end-to-end’ fashion, whereby only the raw input and the final target are known during training, and all components of the model are jointly trained. For example, for the task of hand-written digit recognition, an input instance consists of a pixel image of a hand-written digit, whereas its target denotes the actual digit.

Consider a supervised learning problem, with a training set  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$  containing  $N$  i.i.d. input instances  $\mathbf{x}_i$  and corresponding outputs  $\mathbf{y}_i$ . A model represented by a mapping function  $\mathcal{M}$  with parameters  $\Theta$ , maps an input item  $\mathbf{x}$  to the corresponding predicted output  $\hat{\mathbf{y}} = \mathcal{M}(\mathbf{x}|\Theta)$ .

To quantify how strongly the predicted output  $\hat{\mathbf{y}}$  deviates from the target output  $\mathbf{y}$ , a loss function  $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$  is defined. Training the model then comes down to minimizing the expected loss  $\bar{\mathcal{L}} = \frac{1}{N} \sum_i \mathcal{L}(\mathcal{M}(\mathbf{x}_i|\Theta), \mathbf{y}_i)$  over the training set. In the specific setting of multiclass classification, each input instance corresponds to one out of a fixed set of  $M$  output categories. The target vectors  $\mathbf{y}$  are typically represented as one-hot vectors: all components are zero, except at index  $m$  of the corresponding category. The predicted counterpart  $\hat{\mathbf{y}}$  at the model’s output is often obtained by applying a so-called softmax output function to intermediate real-valued scores  $\mathbf{s}$  obtained at the output of the neural network.

The  $i$ ’th component of the softmax is defined as

$$\text{softmax}(\mathbf{s})_i = \hat{y}_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

The softmax outputs are well-suited to model a probability distribution (i.e.,  $0 < \hat{y}_i < 1$  and  $\sum_i \hat{y}_i = 1$ ). The standard corresponding loss function is the cross-entropy loss, which quantifies the deviation between the empirical output distribution  $\hat{\mathbf{y}}$  (i.e., the softmax outputs) and the ground truth distribution (i.e., the one-hot target vector  $\mathbf{y}$ ) defined as

$$\mathcal{L} = - \sum_j y_j \log \hat{y}_j$$

The most widely used optimization approaches for neural networks are variations of the gradient descent algorithm, in which the parameters  $\Theta$  are iteratively

updated by taking small steps along the negative gradient of the loss. An estimate  $\Theta_n$  at iteration  $n$  is updated as  $\Theta_{n+1} = \Theta_n - \lambda \nabla_{\Theta} \bar{\mathcal{L}}$ , in which the step size is controlled by the learning rate  $\lambda$ . Typically, training is not performed over the entire dataset per iteration, but instead over a smaller ‘mini-batch’ of instances. This is computationally more efficient and allows for a better exploration of parameter space. Importantly, the loss gradient can only be calculated if all components of the neural network are differentiable.

A deep neural network typically has a layer-wise architecture: the different layers correspond to nested differentiable functions in the overall mapping function  $\mathcal{M}$ . The ‘forward pass’ through the network corresponds to consecutively applying these layer functions to a given input to the network. The intermediate representations obtained by evaluating these layer functions are called hidden states. After a forward pass, the gradient with respect to all parameters can then be calculated by applying the chain rule. This happens during the so-called ‘backward pass’: the gradients are calculated from the output back to the first layer. As an illustration of how the chain rule is applied, consider the network function  $\mathcal{M}(\mathbf{x}|\Theta) = \mathbf{g}(\mathbf{f}(\mathbf{x}, \theta_f), \theta_g)$ , which contains a first layer represented by the vector function  $\mathbf{f}$ , and a second layer  $\mathbf{g}$ . For simplicity, say each layer has one trainable parameter, respectively written as  $\theta_f$  and  $\theta_g$ . The derivative with respect to these parameters of a scalar loss function applied to the network output, becomes

$$\nabla_{\Theta} \mathcal{L}(\mathcal{M}(\mathbf{x}|\Theta)) = \left[ \frac{d\mathcal{L}}{d\theta_f}, \frac{d\mathcal{L}}{d\theta_g} \right] = \left[ \sum_i \frac{\partial \mathcal{L}}{\partial g_i} \sum_j \frac{\partial g_i}{\partial f_j} \frac{\partial f_j}{\partial \theta_f}, \sum_i \frac{\partial \mathcal{L}}{\partial g_i} \frac{\partial g_i}{\partial \theta_g} \right]$$

in which the individual derivatives are evaluated based on the considered input  $\mathbf{x}$  and current value of the parameters. The entire procedure to calculate the gradients is called the backpropagation algorithm. It requires a forward pass to calculate all intermediate representations up to the value of the loss. After that, in the backward pass, the gradients corresponding to all operations applied during the forward pass, are iteratively calculated, starting at the loss (i.e., with  $\partial \mathcal{L} / \partial g_i$  in the example). As such, the gradients with respect to parameters at a given layer can be calculated as soon as the gradients due to all operations further in the network are known, as governed by the chain rule.

To summarize, a single iteration in the optimization happens as follows: 1) A minibatch is sampled from the training data. 2) The output of the neural network is calculated during the forward pass. 3) The loss is calculated based on that output and the target. 4) The gradients for the parameters in the neural network are calculated using backpropagation. 5) The parameters are updated using a gradient-based optimizer.

The most basic neural networks building block is the so-called fully-connected layer. It consists of a linear transformation with weight matrix  $\mathbf{W}$  and bias vector  $\mathbf{b}$ , followed by applying a component-wise non-linear function, called the activation function. The input into such a layer can be the vector representation  $\mathbf{x}$  of the actual input to the model, or the output  $\mathbf{h}^{<i>}$  from a previous layer  $i$ , which is called a hidden representation. Its output is calculated as

$\mathbf{h}^{<i+1>} = a(\mathbf{W}\mathbf{h}^{<i>} + \mathbf{b})$ , in which typical choices for the activation function  $a$  are the Rectified Linear Unit (ReLU) defined as  $a(x) = \max(0, x)$  or the hyperbolic tangent  $a(x) = \tanh(x)$ . In other cases, a sigmoid activation can be used, given by  $\sigma(x) = (1 + e^{-x})^{-1}$ . Another important type of neural network layer is the convolutional layer, which convolves the input to pass it to the next layer, by means of a kernel with trainable weights, typically much smaller than the input size. Convolutional layers, followed by a similar activation function, are often used in image recognition models, whereby subsequent layers learn to extract useful features for the given task, from local patterns up to more global and often interpretable patterns. An architecture well-suited for modeling sequences are the so-called recurrent neural networks (RNN). In short, these define a mapping from an input element in the considered sequence into a hidden representation. Every input element is encoded with the same neural network, called the RNN ‘cell’, such that the model can be applied to variable-length sequences. In order to explicitly model the sequential behavior, when encoding a given item in the sequence, the cell takes as input that item, as well as the hidden state obtained while encoding the previous input item. When training with this recurrent setup, the gradient propagates back through the entire sequence. When encoding long sequences, this may lead to very small gradients. An important type of RNN, well-equipped to deal with this so-called vanishing gradient problem, is the Long Short-term Memory (LSTM). The same problem is solved by a similar architecture called the GRU.

More technical details, as well as several other popular types of neural network components, are provided in [15].

Deep neural networks can become very expressive, especially when deeper, or with large hidden representations. To avoid overfitting, various regularization approaches have been developed. A widespread technique, also used in some of the presented experiments in this work, is called dropout. For those layers on which dropout is applied, during training a random sample of the layer outputs are set to zero in each iteration, while accordingly compensating the amplitude of the remaining activations. At inference time, i.e., when applying the trained model to held-out data, all activations are kept.

As mentioned, many choices are possible in terms of architecture and training: dimensions, types of layers, learning rates, regularization strength, etc. These are so-called hyper-parameters, and are typically ‘tuned’ by evaluating on a validation set, not used explicitly for gradient-based training of the network parameters, and still separate from the final test data.

### 3. Introducing DeepProbLog

We now recall the basics of probabilistic logic programming using ProbLog (see De Raedt and Kimmig [6] for more details), and then introduce our new language DeepProbLog.

### 3.1. ProbLog

#### Definition 1 (ProbLog program)

A ProbLog program consists of a set of ground probabilistic facts  $\mathcal{F}$  of the form  $p :: f$  where  $p$  is a probability and  $f$  a ground atom and a set of rules  $\mathcal{R}$ .  $\triangleleft$

For instance, the following ProbLog program models a variant of the well-known alarm Bayesian network [16]:

```

0.1 :: burglary.
0.5 :: hears_alarm(mary).
0.2 :: earthquake.
0.4 :: hears_alarm(john).

alarm :- earthquake.
alarm :- burglary.
calls(X) :- alarm, hears_alarm(X).

```

Each probabilistic fact corresponds to an *independent Boolean random variable* that is true with probability  $p$  and false with probability  $1-p$ . Every subset  $F \subseteq \mathcal{F}$  defines a possible world  $w_F = F \cup \{h\theta \mid \mathcal{R} \cup F \models h\theta \text{ and } h\theta \text{ is ground}\}$ , that is, the world  $w_F$  is the canonical model of the logic program obtained by adding  $F$  to the set of rules  $\mathcal{R}$ , e.g.,

$$w_{\{\text{burglary}, \text{hears\_alarm}(\text{mary})\}} = \{\text{burglary}, \text{hears\_alarm}(\text{mary})\} \cup \{\text{alarm}, \text{calls}(\text{mary})\}$$

To keep the presentation simple, we focus on the case of finitely many ground probabilistic facts, but note that the semantics is also well-defined for the countably infinite case. The probability  $P(w_F)$  of such a possible world  $w_F$  is given by the product of the probabilities of the truth values of the probabilistic facts:

$$P(w_F) = \prod_{f_i \in F} p_i \prod_{f_i \in \mathcal{F} \setminus F} (1 - p_i) \quad (1)$$

For instance,

$$P(w_{\{\text{burglary}, \text{hears\_alarm}(\text{mary})\}}) = 0.1 \times 0.5 \times (1 - 0.2) \times (1 - 0.4) = 0.024$$

The probability of a ground fact  $q$ , also called *success probability of  $q$* , is then defined as the sum of the probabilities of all worlds containing  $q$ , i.e.,

$$P(q) = \sum_{F \subseteq \mathcal{F}: q \in w_F} P(w_F) \quad (2)$$

The probability of a query is also equal to the weighted model count (WMC) of the worlds where this query is true.



For ease of modeling, ProbLog supports non-ground probabilistic facts as a shortcut for introducing a set of ground probabilistic facts, as well as annotated disjunctions (ADs), which are expressions of the form

$$p_1 :: h_1 ; \dots ; p_n :: h_n :- b_1, \dots, b_m.$$

where the  $p_i$  are probabilities that sum to at most one, the  $h_i$  are atoms, and the  $b_j$  are literals. The meaning of an AD is that whenever all  $b_i$  hold, the AD causes one of the  $h_j$  to be true, or none of them with probability  $1 - \sum p_i$ . Note that several of the  $h_i$  may be true at the same time if they also appear as heads of other rules or ADs. This is convenient to model choices between different categorical variables, e.g. different severities of the earthquake:

`0.4::earthquake(none) ; 0.4::earthquake(mild) ; 0.2::earthquake(severe).`

or without explicitly representing the event of no earthquake:

`0.4::earthquake(mild) ; 0.2::earthquake(severe).`

In which neither `earthquake(mild)` nor `earthquake(severe)` will be true with probability 0.4. Annotated disjunctions do not change the expressivity of ProbLog, as they can alternatively be modeled through independent facts and logical rules; we refer to De Raedt and Kimmig [6] for technical details.

### 3.2. DeepProbLog

In ProbLog, the probabilities of all random choices are explicitly specified as part of probabilistic facts or annotated disjunctions. DeepProbLog extends ProbLog to basic random choices whose probabilities are specified through external functions implemented as neural networks.

#### Definition 2 (Neural annotated disjunction)

A *neural AD* is an expression of the form

$$nn(m_r, \vec{I}, O, \vec{d}) :: r(\vec{I}, O).$$

where  $nn$  is a reserved functor,  $m_r$  uniquely identifies a neural network model with  $k$  inputs and  $n$  outputs (i.e., its architecture as well as its trainable parameters) defining a probability distribution over  $n$  classes,  $\vec{I} = I_1, \dots, I_k$  is a sequence of input variables,  $O$  is the output variable,  $\vec{d} = d_1, \dots, d_n$  is a sequence of ground terms (the output domain of this neural network) and  $r$  is a predicate.


A *ground neural AD* is an expression of the form

$$nn(m_r, \vec{i}, d_1) :: r(\vec{i}, d_1) ; \dots ; nn(m_r, \vec{i}, d_n) :: r(\vec{i}, d_n).$$

where  $\vec{i} = i_1, \dots, i_k$  is a sequence of ground terms (the input to the neural network) and  $d_1, \dots, d_n$  are ground terms (the output domain of this neural network).  $\triangleleft$

The  $nn(m_r, \vec{i}, d_j)$  term in the definition can be considered a function that returns the probability of class  $d_j$  when evaluating the network  $m_r$  on input  $\vec{i}$ . As such, a ground nAD can be instantiated into a normal AD by evaluating the neural network and replacing the functor with the calculated probability. For instance, in the MNIST addition example, we would specify the nAD


$$nn(m\_digit, [X], Y, [0, \dots, 9]) :: digit(X, Y).$$

where `m_digit` is a network that classifies MNIST digits. Grounding this on an input image  would result in a ground nAD:

$$nn(m\_digit, [\text{3}], 0) :: digit(\text{3}, 0) ; \dots ; nn(m\_digit, [\text{3}], 9) :: digit(\text{3}, 9).$$

Evaluating this would result in a ground AD:

$$p_0 :: digit(\text{3}, 0) ; \dots ; p_9 :: digit(\text{3}, 9).$$

Where  $[p_0, \dots, p_9]$  is the output vector of the `m_digit` network when evaluated on .

The neural network could take any shape, e.g., a convolutional network for image encoding, a recurrent network for sequence encoding, etc. However, its output layer, which feeds the corresponding neural predicate, needs to be normalized.

We consider an output domain size of two as a special case. Instead of the neural network having two probabilities at the output that sum to one, we can simplify this to a single probability, with the second one the complement of that probability. This difference coincides with the difference between a softmax and single-neuron sigmoid layer in a neural network. We call such an expression a neural fact.

### Definition 3 (Neural fact)

A neural fact is an expression of the form

$$nn(m_r, \vec{I}) :: r(\vec{I}).$$

where  $nn$  is a reserved functor,  $m_r$  uniquely identifies a neural network model defining a probability distribution over  $n$  classes,  $\vec{I} = I_1, \dots, I_k$  is a sequence of input variables and  $r$  is a predicate.

A ground neural fact is an expression of the form

$$nn(m_r, \vec{i}) :: r(\vec{i}).$$

where  $\vec{i} = i_1, \dots, i_k$  is a sequence of ground terms (the input to the neural network). ◁

To exemplify, we use a neural network that gives a measure of the similarity between two input images. We can encode this with the following neural fact:

$$nn(m, [X, Y]) :: similar(X, Y).$$

Grounding this on the input  $\mathbf{3}$  and  $\mathbf{3}$  would result in the follow ground neural fact:

$$\text{nn}(\mathbf{m}, [\mathbf{3}, \mathbf{3}]) :: \text{similar}(\mathbf{3}, \mathbf{3}).$$

Evaluating this would result in a ground probabilistic fact:

$$p :: \text{similar}(\mathbf{3}, \mathbf{3}).$$

Where  $p$  is the output of the  $m$  network when evaluated on  $\mathbf{3}$  and  $\mathbf{3}$ .

#### Definition 4 (DeepProbLog Program)

A DeepProbLog program consists of a set of ground probabilistic facts  $\mathcal{F}$ , a set of ground neural ADs and ground neural facts  $\mathcal{N}$ , and a set of rules  $\mathcal{R}$ .  $\triangleleft$

The semantics of a DeepProbLog program is given by the semantics of the ProbLog program obtained by replacing each nAD with the AD obtained by instantiating the probabilities as mentioned above. While the semantics is defined with respect to ground neural ADs and facts, as in ProbLog, we write non-ground such expressions if the intended grounding is clear from context.

### 4. Inference

This section explains how a DeepProbLog model is used for a given query at prediction time. First, we provide more detail on ProbLog inference [17]. Next, we describe how ProbLog inference is adapted in DeepProbLog.

#### 4.1. ProbLog Inference

ProbLog inference proceeds in four steps. The first step is the grounding step, in which the logic program is grounded with respect to the query. This step uses backward reasoning to determine which ground rules are relevant to derive the truth value of the query, and may perform additional logical simplifications that do not affect the query’s probability.

The second step rewrites the ground logic program into a formula in propositional logic that defines the truth value of the query in terms of the truth values of probabilistic facts. We can calculate the query success probability by performing *weighted model counting* (WMC) on this logic formula (cfr. Fierens et al. [17]). However, performing WMC on this logical formula directly is not efficient.

The third step is knowledge compilation [18]. During this step, the logic formula is transformed into a form that allows for efficient weighted model counting. The current ProbLog system uses Sentential Decision Diagrams (SDDs, Darwiche [19]), the most succinct suitable representation available today. SDDs, being a subset of d-DNNFs allow for polytime model counting ([18]). However, they also support polytime conjunction, disjunction and negation while being

more succinct than OBDDs (Darwiche [19]).

The fourth and final step transforms the SDD into an arithmetic circuit (AC). This is done by putting the probabilities of the probabilistic facts or their negations on the leaves, replacing the OR nodes with addition and the AND nodes by multiplication. The WMC is then calculated with an evaluation of the AC.

### Example 2

In Figure 1, we apply the four steps of ProbLog inference on the earthquake example with query `calls(mary)`.

In the first step, the non-ground program (Figure 1a) is grounded with respect to the query `calls(mary)`. The result is shown in Figure 1b: the irrelevant fact `hears_alarm(john)` is omitted and the variable `X` in the `calls` rule is substituted with the constant `mary`. The resulting formula in the second step is

$$\text{calls}(\text{mary}) \leftrightarrow \text{hears\_alarm}(\text{mary}) \wedge (\text{burglary} \vee \text{earthquake})$$

The WMC of this formula is shown in Figure 1c. However, it is not calculated by enumeration as shown here, but an AC is used instead. The AC derived in step four is shown in Figure 1d, where rounded grey rectangles depict variables corresponding to probabilistic facts, and the rounded red rectangle denotes the query atom defined by the formula. The white rectangles correspond to logical operators applied to their children. The intermediate results are shown in black next to the nodes in Figure 1d.

## 4.2. DeepProbLog Inference

The only change required for DeepProbLog inference is that we need to instantiate the ground nADs and neural facts into the corresponding ground ADs and ground facts. This is done in a separate step after grounding, where the parameters for the regular AD are determined by making a forward pass on the relevant neural network with the ground input.

### Example 3

We illustrate this by evaluating the MNIST addition example (Figure 2a). The DeepProbLog program requires two lines: the first line defining the neural predicate, and the second line defining the addition. We evaluate it on the query `addition(0, 1, 1)`. In the first step, the DeepProbLog program is grounded into a ground DeepProbLog Program (Figure 2b). Note that the nADs are now all ground. As ProbLog only grounds the relevant part of the program, i.e. the part that can be used to prove the query, only the digits 0 and 1 are retained as the larger digits cannot sum to 1. The next step is the only difference between ProbLog and DeepProbLog inference: instantiating the ground nADs into regular ground ADs, which could, for instance, produce an AD as shown in Figure 2c. The probabilities in the

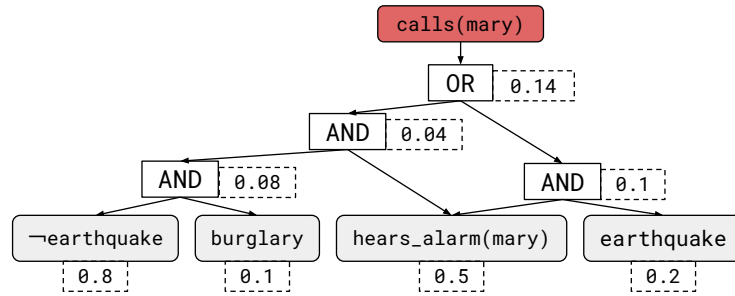
|                                 |                                       |
|---------------------------------|---------------------------------------|
| 0.2::earthquake.                | 0.2::earthquake.                      |
| 0.1::burglary.                  | 0.1::burglary.                        |
| 0.5::hears_alarm(mary).         | 0.5::hears_alarm(mary).               |
| 0.4::hears_alarm(john).         |                                       |
| alarm :- earthquake.            | alarm :- earthquake.                  |
| alarm :- burglary.              | alarm :- burglary.                    |
| calls(X):-alarm,hears_alarm(X). | calls(mary):-alarm,hears_alarm(mary). |

(a) The ProbLog program.

(b) The relevant ground program.

| Models of $\text{calls(mary)} \leftrightarrow \text{hears\_alarm(mary)} \wedge (\text{burglary} \vee \text{earthquake})$ | w           |
|--|-------------|
| $\{\}$   | 0.36        |
| $\{\text{hears\_alarm(mary)}\}$  | 0.36        |
| $\{\text{earthquake}\}$  | 0.09        |
| $\{\text{earthquake, hears\_alarm(mary), calls(mary)}\}$   | <b>0.09</b> |
| $\{\text{burglary}\}$  | 0.04        |
| $\{\text{burglary, hears\_alarm(mary), calls(mary)}\}$   | <b>0.04</b> |
| $\{\text{burglary, earthquake}\}$  | 0.01        |
| $\{\text{burglary, earthquake, hears\_alarm(mary), calls(mary)}\}$   | <b>0.01</b> |
| $\sum_{\text{calls(mary)} \in \text{model}}$   | <b>0.14</b> |

(c) The weighted count of the models where  $\text{calls(mary)}$  is true.



(d) The AC for query  $\text{calls(mary)}$ .

Figure 1: Inference in ProbLog using query  $\text{calls(mary)}$  and the program in (a). (Example 2)

---

```

nn(m_digit, [X], Y, [0...9]) :: digit(X,Y).
addition(X,Y,Z) :- digit(X,N1), digit(Y,N2), Z is N1+N2.

```

---

(a) The DeepProbLog program.

---

```

nn(m_digit, [0], 0) :: digit(0,0); nn(m_digit, [0], 1) :: digit(0,1).
nn(m_digit, [1], 0) :: digit(1,0); nn(m_digit, [1], 1) :: digit(1,1).
addition(0,1,1) :- digit(0,0), digit(1,1).
addition(0,1,1) :- digit(0,1), digit(1,0).

```

---

(b) The ground DeepProbLog program.

---

```

0.8 :: digit(0,0); 0.1 :: digit(0,1).
0.2 :: digit(1,0); 0.6 :: digit(1,1).
addition(0,1,1) :- digit(0,0), digit(1,1).
addition(0,1,1) :- digit(0,1), digit(1,0).

```

---

(c) The ground ProbLog program.

Figure 2: Inference in DeepProbLog (Example 3)

instantiated ADs do not sum to one, as the irrelevant terms ( $\text{digit}(0,2)$ , ...,  $\text{digit}(0,9)$  and  $\text{digit}(1,2)$ , ...,  $\text{digit}(1,9)$ ) have been dropped in the grounding process, although the neural network still assigns probability mass to them. Inference then proceeds identically to that of ProbLog: the ground program is rewritten into a logical formula, this formula is compiled and transformed into an AC. Finally, this AC is evaluated to calculate the query probability.

## 5. Learning in DeepProbLog

We now introduce our approach to learn the parameters in DeepProbLog programs. The parameters include the learnable parameters of the neural network (which we will call neural parameters from now on) and the learnable parameters in the logic program (which we will refer to as probabilistic parameters). We use the *learning from entailment* setting [20]

### Definition 5

*Learning from entailment* Given a DeepProbLog program with parameters  $\Theta$ , a set  $\mathcal{Q}$  of pairs  $(q, p)$  with  $q$  a query and  $p$  its desired success probability, and a loss function  $\mathcal{L}$ , compute:

$$\arg \min_{\Theta} \frac{1}{|\mathcal{Q}|} \sum_{(q,p) \in \mathcal{Q}} \mathcal{L}(P(q|\Theta), p)$$

In most of the experiments, unless mentioned otherwise, we only use positive examples for training (i.e., with desired success probability  $p = 1$ ). The model then needs to adjust the weights to maximize query probabilities  $P_{\Theta}(q)$  for all training examples. This can be expressed by minimizing the average negative log likelihood of the query, whereby Definition 5 reduces to:

$$\arg \min_{\Theta} \frac{1}{|\mathcal{Q}|} \sum_{(q,p) \in \mathcal{Q}} -\log P_{\Theta}(q)$$

The presented method however works for other choices in the loss function. For example, in experiment **T9** (Section 6.3) the mean squared error (MSE) is used.

### 5.1. Gradient descent in ProbLog

In contrast to the earlier approach for ProbLog parameter learning in this setting by Gutmann et al. [21], we use gradient descent rather than EM. This allows for seamless integration with neural network training. The key insight here is that we can use the same AC that ProbLog uses for inference for gradient computations as well. We rely on the automatic differentiation capabilities already available in ProbLog to derive these gradients. More specifically, to compute the gradient with respect to the probabilistic logic program part, we rely on Algebraic ProbLog (aProbLog [22]), a generalization of the ProbLog language and inference to arbitrary commutative semirings, including the gradient semiring [23]. In the following, we provide the necessary background on aProbLog, discuss how to use it to compute gradients with respect to ProbLog parameters and extend the approach to DeepProbLog.

*aProbLog and the gradient semiring.* ProbLog annotates each probabilistic fact  $f$  with the probability  $P$  that  $f$  is true, which implicitly also defines the probability  $1 - P$  that its negation  $\neg f$  is true. It then uses the probability semiring with regular addition and multiplication as operators to compute the probability of a query on the AC constructed for this query, cf. Figure 1d. The probability semiring is defined as follows:

$$a \oplus b = a + b \tag{3}$$

$$a \otimes b = ab \tag{4}$$

$$e^{\oplus} = 0 \tag{5}$$

$$e^{\otimes} = 1 \tag{6}$$

And the accompanying labeling function as:

$$L(f) = p \quad \text{for } p :: f \tag{7}$$

$$L(\neg f) = 1 - p \quad \text{with } L(f) = p \tag{8}$$

This idea is generalized in aProbLog to compute such values based on arbitrary commutative semirings. Instead of probability labels on facts, aProbLog uses a labeling function that explicitly associates values from the chosen semiring with

both facts and their negations, and combines these using semiring addition  $\oplus$  and multiplication  $\otimes$  on the AC. We use the gradient semiring, whose elements are tuples  $(p, \frac{\partial p}{\partial \theta})$ , where  $p$  is a probability (as in ProbLog), and  $\frac{\partial p}{\partial \theta}$  is the partial derivative of that probability with respect to a parameter  $\theta$ , that is, the probability  $p_i$  of a probabilistic fact with learnable probability, written as  $t(p_i) :: f_i$ . This is easily extended to a vector of parameters  $\vec{\theta} = [\theta_1, \dots, \theta_N]^T$ , the concatenation of all  $N$  probabilistic parameters in the ground program, as it is easier and faster to process all gradients in one vector. Semiring addition  $\oplus$ , multiplication  $\otimes$  and the neutral elements with respect to these operations are defined as follows:

$$(a_1, \vec{a}_2) \oplus (b_1, \vec{b}_2) = (a_1 + b_1, \vec{a}_2 + \vec{b}_2) \quad (9)$$

$$(a_1, \vec{a}_2) \otimes (b_1, \vec{b}_2) = (a_1 b_1, b_1 \vec{a}_2 + a_1 \vec{b}_2) \quad (10)$$

$$e^\oplus = (0, \vec{0}) \quad (11)$$

$$e^\otimes = (1, \vec{0}) \quad (12)$$

Note that the first element of the tuple mimics ProbLog’s probability computation, whereas the second simply computes gradients of these probabilities using derivative rules.

*Gradient descent with aProbLog.* To use the gradient semiring for gradient descent parameter learning in ProbLog, we first transform the ProbLog program into an aProbLog program by extending the label of each probabilistic fact  $p :: f$  to include the probability  $p$  as well as the gradient vector of  $p$  with respect to the probabilities of all probabilistic facts and ADs in the program, i.e.,

$$L(f) = (p, \vec{0}) \quad \text{for } p :: f \text{ with fixed } p \quad (13)$$

$$L(f_i) = (p_i, \mathbf{e}_i) \quad \text{for } t(p_i) :: f_i \text{ with learnable } p_i \quad (14)$$

$$L(\neg f) = (1 - p, -\nabla p) \quad \text{with } L(f) = (p, \nabla p) \quad (15)$$

where the vector  $\mathbf{e}_i$  has a 1 in the  $i$ th position and 0 in all others. For fixed probabilities, the gradient does not depend on any parameters and thus is 0. Note that after each update step, the probabilistic parameters are clipped to the  $[0, 1]$  range, and the parameters of an AD are re-normalized to ensure that they sum to one. For the other cases, we use the semiring labels as introduced above.

#### Example 4

Assume we want to learn the probabilities of **earthquake** and **burglary** in the example of Figure 1, while keeping those of the other facts fixed. Figure 3 shows the evaluation of the same AC as in Figure 1d, but with the gradient semiring. The nodes in the AC now also contain the gradient (the second element of the tuple). The result on the top node shows that



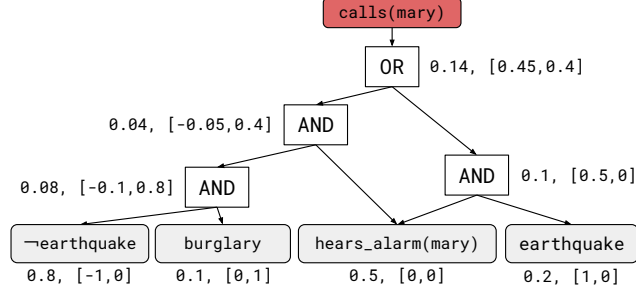


Figure 3: The AC evaluated using the gradient semiring. (Example 4)

the partial derivative of the query is 0.45 and 0.4 w.r.t. the earthquake and burglary parameters respectively.

## 5.2. Gradient descent for DeepProbLog

Just as the only difference between inference in ProbLog and DeepProbLog is the evaluation of the nADs, the only difference between gradient descent in ProbLog and DeepProbLog is optimizing the neural parameters alongside the probabilistic parameters. As mentioned in the previous section, the probabilistic parameters  $p_i$  in the logic program can be optimized by using the gradient semiring, which allows us to calculate  $\partial P(q)/\partial p_i$ . This gradient is then used to perform the update by using gradient descent. Note that since the outputs of the neural network are used as probabilities in the logic program and can be learned, we can view them as a kind of *abstract* parameters. However, although we can derive a gradient for these abstract parameters, we cannot optimize them directly, as the logic is unaware of the neural parameters that determine the value of these abstract parameters. Recall from Equation (1) that the gradient of the internal (neural) parameters in standard supervised learning can be derived using the chain rule in backpropagation. Below, we show how we can derive the gradient for these neural parameters of the loss applied to  $P(q)$  (Definition 5), rather than a loss function defined directly on the output of the neural network.

Specifically, consider the case of a single neural annotated disjunction, with probabilities  $\hat{p}_i$  (i.e., the aforementioned abstract parameters), calculated by evaluating a neural network with softmax output. The predicted probability that the query holds true, based on the current values of the neural and probabilistic parameters, is written  $P(q)$ . While training, true examples should yield a predicted query probability close to the expected query probability, which is expressed by means of a loss function  $\mathcal{L}$  as introduced in Definition 5.

Application of the chain rule leads to

$$\frac{d\mathcal{L}}{d\theta_k} = \frac{\partial \mathcal{L}}{\partial P(q)} \sum_i \frac{\partial P(q)}{\partial \hat{p}_i} \frac{\partial \hat{p}_i}{\partial \theta_k}$$

where the derivative of the loss with respect to any trainable parameter  $\theta_k$  in the neural network is decomposed into the partial derivative of the loss with

respect to the predicted output  $P(q)$ , the latter's derivative  $\partial P(q)/\partial \hat{p}_i$  with respect to each component of the annotated disjunction as obtained with the gradient semiring, and finally  $\partial \hat{p}_i/\partial \theta_k$ , the derivative of the neural network's output components with respect to the considered parameter. The latter is obtained by the standard application of the chain rule in the neural network. The backpropagation procedure in the neural network can thus be started by providing  $\partial P(q)/\partial \hat{p}_i$ , to systematically obtain the loss gradients for all neural parameters.

Extending this approach to the situation of multiple neural predicates is straightforward. If the same neural network is used for different neural predicates (e.g. in Example 3), the final derivative is obtained by summing over the contributions of each neural predicate.

Then, standard gradient-based optimizers (e.g. SGD, Adam, ...) are used to update the parameters of the network. During gradient computation with aProbLog, the probabilities of neural ADs are kept constant. Furthermore, updates on neural ADs come from the neural network part of the model, where the use of a softmax output layer ensures a normalized distribution, hence not requiring the additional normalization as for non-neural ADs.

To extend the gradient semiring to DeepProbLog programs, we define it for nADs and neural facts. The label for the nAD is defined as:

$$L(f_i) = (\hat{p}_j, \mathbf{e}_j) \quad \text{for } \dots ; nn(m, \vec{i}, d_j) :: r(\vec{i}, d_j) ; \dots \text{ a ground nAD} \quad (16)$$

Where  $d_j$  is the  $j$ -th domain element,  $\hat{p}_j$  is the  $j$ -th element of the output of the neural network  $m$  evaluated on input  $\vec{i}$ . The label for a neural fact is defined as:

$$L(f_i) = (\hat{p}, \mathbf{e}_j) \quad \text{for } nn(m, \vec{i}) :: r(\vec{i}) \text{ a ground neural fact} \quad (17)$$

where  $\hat{p}$  is the output of the neural network  $m$  evaluated on input  $\vec{i}$ . Since the first element of the tuple for nADs and neural facts is the evaluation of the neural networks as in Section 4.2, this change remains semantically equivalent.

### Example 5

To demonstrate the learning pipeline (Figure 5), we will apply it on the MNIST addition example show in Section 4.2 with a small extension: some of the labels have been corrupted and are picked randomly from a uniform distribution over  $[0, 18]$ . The goal is to also learn the fraction of noisy examples. The DeepProbLog program is given in Figure 4a. Grounding on the query `addition(a, b, 1)` results in the ground DeepProbLog program shown in Figure 4b. The arithmetic circuit corresponding to the ground program is shown in Figure 4c. As can be seen, the neural networks already have a confident prediction for both images (being 0 and 1 respectively). The top right shows how the different partial derivatives that are calculated: one w.r.t. to the noisy parameter, ten for the evaluation of the neural network on input a and ten for the evaluation on input b.

---

```

nn(classifier, [X], Y, [0 .. 9]) :: digit(X,Y).
t(0.2) :: noisy.

1/19 :: uniform(X,Y,0) ; ... ; 1/19 :: uniform(X,Y,18).

addition(X,Y,Z) :- noisy, uniform(X,Y,Z).
addition(X,Y,Z) :- \+noisy, digit(X,N1), digit(Y,N2), Z is N1+N2.

```

---

(a) The DeepProbLog program.

---

```

nn(classifier, [a], 0) :: digit(a,0); nn(classifier, [a], 1) :: digit(a,1).
nn(classifier, [b], 0) :: digit(b,0); nn(classifier, [b], 1) :: digit(b,1).
t(0.2)::noisy.

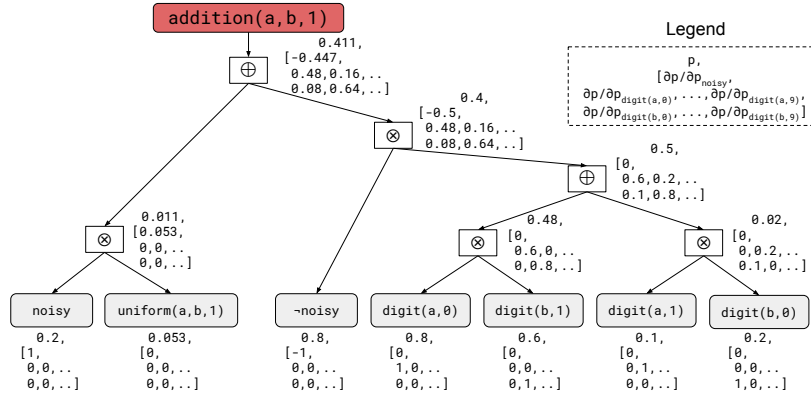
1/19::uniform(a,b,1).
addition(a,b,1) :- noisy, uniform(a,b,1).

addition(a,b,1) :- \+noisy, digit(a,0), digit(b,1).
addition(a,b,1) :- \+noisy, digit(a,1), digit(b,0).

```

---

(b) The ground DeepProbLog program.



(c) The AC for query `addition(a,b,1)`.

Figure 4: Parameter learning in DeepProbLog. (Example 5)

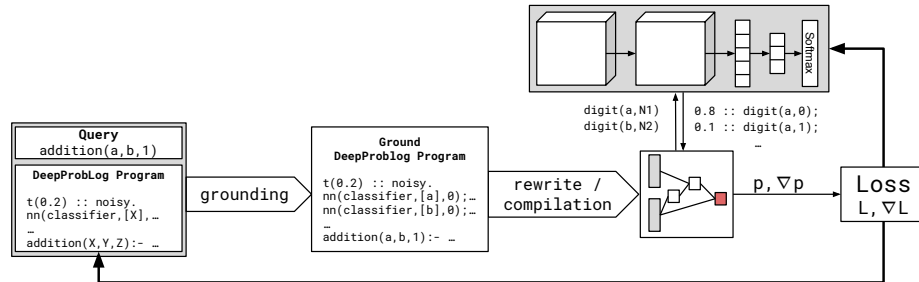


Figure 5: The learning pipeline.



## 6. Experimental Evaluation

We perform three sets of experiments to demonstrate that DeepProbLog supports (i) logical reasoning and deep learning; (ii) program induction; and (iii) probabilistic inference and combined probabilistic and deep learning.

We provide implementation details at the end of this section and list all programs in Appendix A.

### 6.1. Logical reasoning and deep learning

To show that DeepProbLog supports both logical reasoning and deep learning, we extend the classic learning task on the MNIST dataset [24] to four more complex problems that require reasoning:

**T1:** `addition(, 8)`

Instead of using labeled single digits, we train on pairs of images, labeled with the sum of the individual labels. This is the same as Example 3. The DeepProbLog program consists of the clause

$$\text{addition}(X, Y, Z) \text{:-} \text{digit}(X, X2), \text{digit}(Y, Y2), Z \text{ is } X2 + Y2$$

and a neural AD for the `digit/2` predicate, which classifies an MNIST image. We compare to a CNN baseline <sup>3</sup> classifying the two images into the 19 possible sums.

*Results.* Figure 6 shows the learning curves for the baseline (orange) and DeepProbLog (blue) on the single-digit addition. We evaluated on 3 levels of data availability: 30 000 examples, 3 000 and 300 examples. As can be seen in the figures, DeepProbLog converges faster and achieves a higher accuracy than the baseline. In the case for  $N = 30\,000$  (Figure 6a), the difference between the baseline and DeepProbLog is significant, but not immense. However, for  $N = 3000$  and especially  $N = 300$ , the difference becomes more apparent.

The reason behind this disparity is that the baseline needs to learn making a decision for the combined input digits (and there are a 100 different sums possible), whereas the DeepProbLog’s neural predicate only needs to recognize individual digits (with only 10 possibilities). Table 1 shows the average accuracy on the test set for the different models for different training set sizes.

**T2:** `addition(, , 63)`

The input consists of two lists of images, each element being a digit. Each list represents a multi-digit number. The label is the sum of the two numbers. The neural predicate remains the same. Learning the new predicate requires only a small change in the logic program. Because the CNN baseline cannot handle numbers of varying size, we fixed the size of the input to two-digit numbers.

---

<sup>3</sup>We’d like to thank Paolo Frasconi for the interesting discussion and idea for a new baseline.

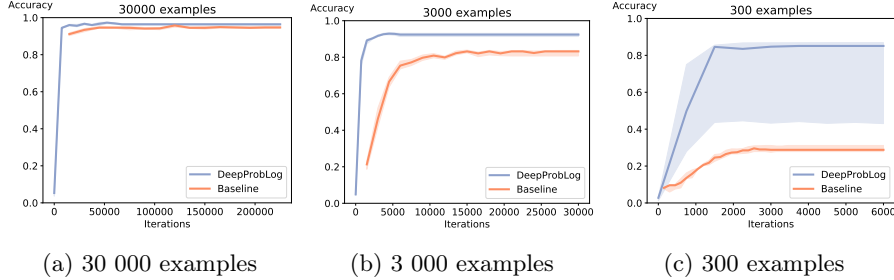


Figure 6: MNIST Single-Digit Addition (**T1**). The graphs show the accuracy on the validation set during training for different training set sizes.

| Model       | Number of training examples |                  |                   |
|-------------|-----------------------------|------------------|-------------------|
|             | 30 000                      | 3 000            | 300               |
| Baseline    | $93.46 \pm 0.49$            | $78.32 \pm 2.14$ | $23.64 \pm 1.75$  |
| DeepProbLog | $97.20 \pm 0.45$            | $92.18 \pm 1.57$ | $67.19 \pm 25.05$ |

Table 1: The accuracy on the test set for **T1**.

*Results.* First, we perform an experiment where we take the neural network trained in **T1** and use it in this model without any further training. Evaluating it on the same test set, we achieve an accuracy that is not significantly different from training on the full dataset of **T2**. This demonstrates that the approach used in DeepProbLog causes it to generalize well beyond training data. Figure 7 shows the learning curves for the baseline (orange) and DeepProbLog (blue) on the multi-digit addition. DeepProbLog achieves a somewhat lower accuracy compared to the single digit problem due to the compounding effect of the classification error on the individual digits, but the model generalizes well. The baseline fails to learn from few examples (150 and 1 500). It is able to learn with 15 000 examples, but converges very slowly. Table 2 shows the average accuracy on the test set for the different models for different training set sizes.

### **T3:** addition(**3**, **5**, **8**)

The input consists of 3 MNIST images such that the last is the sum of the first two. This task demonstrates potential pitfalls of only providing supervision on the logic level. Namely, without any regularization, the neural network quickly learns to predict 0 for all digits, i.e., the model collapses to always predicting  $0 + 0 = 0$ , as it is a valid logical solution. To avoid this, we add a regularisation term based on entropy maximization (Equation 18, Section 6.4). The intuition behind this regularisation term is that it penalizes mode collapse by requiring the entropy of the average output distribution per batch to be high. As such, this term encourages exploration, but is only necessary to start the training of the neural networks.

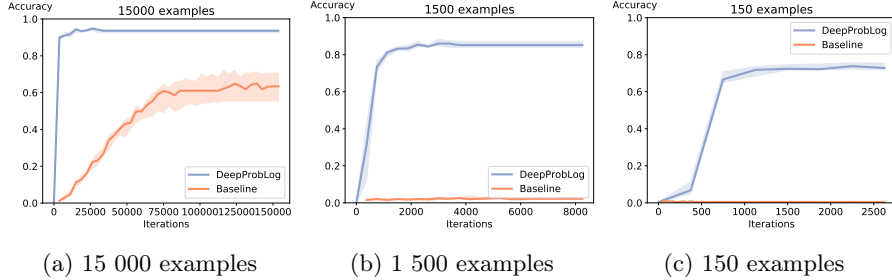


Figure 7: MNIST Multi-Digit Addition (**T2**). The graphs show the accuracy on the validation set during training for different training set sizes.

| Model       | Number of training examples |                  |                  |                    |
|-------------|-----------------------------|------------------|------------------|--------------------|
|             | 15 000                      | 1 500            | 150              | <b>T1</b> (30 000) |
| Baseline    | $60.85 \pm 9.77$            | $1.34 \pm 0.53$  | $0.80 \pm 0.14$  | –                  |
| DeepProbLog | $95.16 \pm 1.70$            | $87.21 \pm 1.92$ | $72.73 \pm 3.03$ | $93.36 \pm 1.18$   |

Table 2: The accuracy on the test set for **T2**.

If they are sufficiently trained, this term can be dropped. We call this additional loss term *infoloss*. This additional regularization loss is multiplied by a factor  $\lambda$  and added to the cross-entropy loss. We run the experiment for different values of  $\lambda$ .

*Results.* Figure 8 shows the accuracy of the neural predicate on classifying single digits for different levels of the regularization parameter. As can be seen, for  $\lambda = 2$ , the neural predicate converges on the trivial solution. For  $\lambda = 4$ , the neural predicate sometimes converges on the correct solution, but can also converge on the wrong solution. For  $\lambda = 8$ , the neural network consistently converges on the correct solution.

#### **T4:** addition(**3**, **5**, **14**)

This experiment is the example shown in Figure 4. It’s the same as **T1**, but with noise introduced in the labels. Namely, a fraction of the labels is replaced by uniformly selected number between 0 and 18. We compare three models: the CNN baseline from **T1**, the DeepProbLog model from **T1**, and a DeepProbLog model where the noise is explicitly modeled as in Figure 4.

*Results.* Table 3 shows the accuracy on the test set which has no noise. The baseline is not tolerant to noisy labels, quickly dropping in accuracy as the fraction of noisy labels increases. The DeepProbLog model from **T1** is more tolerant, but also drops noticeably in accuracy as the fraction of noise goes over 0.5. Explicitly modeling the noise makes the model very

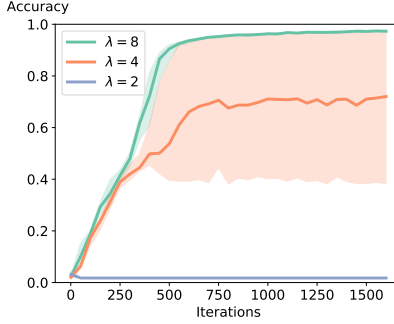


Figure 8: The accuracy on the MNIST test set for individual digits while training on (**T3**).

|                               | Fraction of noise |       |       |       |       |       |
|-------------------------------|-------------------|-------|-------|-------|-------|-------|
|                               | 0.0               | 0.2   | 0.4   | 0.6   | 0.8   | 1.0   |
| Baseline                      | 93.46             | 87.85 | 82.49 | 52.67 | 8.79  | 5.87  |
| DeepProbLog                   | 97.20             | 95.78 | 94.50 | 92.90 | 46.42 | 0.88  |
| DeepProbLog w/ explicit noise | 96.64             | 95.96 | 95.58 | 94.12 | 73.22 | 2.92  |
| Learned fraction of noise     | 0.000             | 0.212 | 0.415 | 0.618 | 0.803 | 0.985 |

Table 3: The accuracy on the test set for **T4**.

noise tolerant, even retaining an accuracy of 73.2% with 80% noisy labels. As shown in the last row, it is also able to learn the fraction of noisy labels in the data. This shows that the model is able to recognize which examples have noisy labels.

## 6.2. Program Induction

The second set of problems demonstrates that DeepProbLog can perform program induction. We follow the program sketching [25] setting of differentiable Forth ( $\partial 4$ ) [8], where holes in given programs need to be filled by neural networks trained on input-output examples for the entire program. As in their work, we consider three tasks: addition, sorting [26] and word algebra problems (WAPs) [27].

**T5:** `forth_addition([4], [8], 1, [1, 3])`

The input consists of two numbers, represented as lists of digits, and a carry. The output is the sum of the numbers and the carry. The program specifies the basic addition algorithm in which we go from right to left over all digits, calculating the sum of two digits and taking the carry over to the next pair. The hole in this program corresponds to calculating the resulting digit (`result/4`) and carry (`carry/4`), given two digits and the previous carry.

|                  | Test length | Training length |       |       |
|------------------|-------------|-----------------|-------|-------|
|                  |             | 2               | 4     | 8     |
| $\partial 4$ [8] | 8           | 100.0           | 100.0 | 100.0 |
|                  | 64          | 100.0           | 100.0 | 100.0 |
| DeepProbLog      | 8           | 100.0           | 100.0 | 100.0 |
|                  | 64          | 100.0           | 100.0 | 100.0 |

Table 4: Accuracy on the addition (**T5**) problem (results for  $\partial 4$  reported by Bošnjak et al. [8]).

*Results.* The results are shown in Table 4. Similarly to  $\partial 4$ , DeepProbLog achieves 100% on all training sizes.

**T6:** `forth_sort([8, 2, 4], [2, 4, 8])`

The input consists of a list of numbers, and the output is the sorted list. The program implements bubble sort, but leaves open what to do on each step in a bubble (i.e. whether to swap or not: `swap/2`).

*Results.* The results are shown in Table 5. Similarly to  $\partial 4$ , DeepProbLog achieves 100% on training sizes 2 and 3. However, whereas  $\partial 4$  fails to converge on training sizes larger than 3, DeepProbLog stills achieves 100% accuracy. As Bošnjak et al. [8] mention, the failure of  $\partial 4$  is due to computational issues arising from the long program trace resulting from sorting long lists. DeepProbLog does not suffer from these issues. As shown in Table 6, DeepProbLog runs faster and scales better with increasing training length.

**T7:** `wap('Robert has 12 books . ... How many does he have now ?', 12, 3, 1, 10)`

The input to the WAPs consists of a natural language sentence describing a simple mathematical problem. These WAPs always contain three numbers, which are extracted from the string and are given as part of the input. The output is the solution to the question. Every WAP can be solved by chaining the following 4 steps: permuting the three numbers (`permute/2`), applying an operation on the first two numbers (addition, subtraction or product `operation_1/2`), potentially swapping the intermediate result and the last digit (`swap/2`), and performing a last operation (`operation_2/2`). The hole in the program is in deciding which of the alternatives should happen on each step.

*Results.* DeepProbLog reaches an accuracy of up to 96.5%, similar to the results for  $\partial 4$  reported by Bošnjak et al. [8] (96%).



|                  |    | Training length |       |       |       |       |
|------------------|----|-----------------|-------|-------|-------|-------|
|                  |    | 2               | 3     | 4     | 5     | 6     |
| $\partial 4$ [8] | 8  | 100.0           | 100.0 | 49.22 | –     | –     |
|                  | 64 | 100.0           | 100.0 | 20.65 | –     | –     |
| DeepProbLog      | 8  | 100.0           | 100.0 | 100.0 | 100.0 | 100.0 |
|                  | 64 | 100.0           | 100.0 | 100.0 | 100.0 | 100.0 |

Table 5: Accuracy on the sorting (**T6**) problem (results for  $\partial 4$  reported by Bošnjak et al. [8]).

|                     |      | Training length |      |       |       |   |
|---------------------|------|-----------------|------|-------|-------|---|
|                     |      | 2               | 3    | 4     | 5     | 6 |
| $\partial 4$ on GPU | 42 s | 160 s           | –    | –     | –     | – |
| $\partial 4$ on CPU | 61 s | 390 s           | –    | –     | –     | – |
| DeepProbLog         | 11 s | 14 s            | 32 s | 114 s | 245 s | – |

Table 6: Time until 100% accurate on test length 8 for the sorting (**T6**) problem.

### 6.3. Probabilistic programming and deep learning

In this section we introduce two final experiments that show the intricacies involved in combining probabilistic logic programming and deep learning.

#### **T8:** *Coin classification and comparison*

In this experiment we train two neural networks using distant supervision. The input consists of a synthetic image containing two coins (an example is shown in Figure 9). They are either heads or tails. The image is labeled either with *same* or *different*. We train a neural network for each coin to predict either *heads* or *tails*. Solving this task requires solving two problems. On the one hand, the neural networks have to learn to recognize and separate the two different coins; on the other hand, they also have to each classify a different coin as heads or tails. The first question we ask is whether the neural networks can recover the latent structure imposed by the logic program. We expect the two neural networks to agree on which side of the coin is heads and which is tails, however, this might be the inverse of what is generally considered heads and tails. Furthermore, we expect the two neural networks to each pick one coin to label, but which network classifies which coin will vary between runs. As such, there are four possible solutions that the neural networks can converge on. The second question we ask is how many additionally labeled examples (with both the label for same/different and heads/tails of one of the coins given) we need for the neural network to recover the desired latent representation.

*Results.* We ran each experiment 100 times. The fraction of runs that converged on either no solution, the expected solution or a logically equivalent

| Labeled examples | Not solved | Expected solution | Other solution |
|------------------|------------|-------------------|----------------|
| 0                | 56%        | 11%               | 33%            |
| 5                | 39%        | 40%               | 21%            |
| 10               | 7%         | 92%               | 1%             |
| 20               | 4%         | 96%               | 0%             |
| 50               | 3%         | 97%               | 0%             |
| 100              | 4%         | 96%               | 0%             |

Table 7: The fraction of runs that converged to different outcomes for the Coins experiment (**T8**).

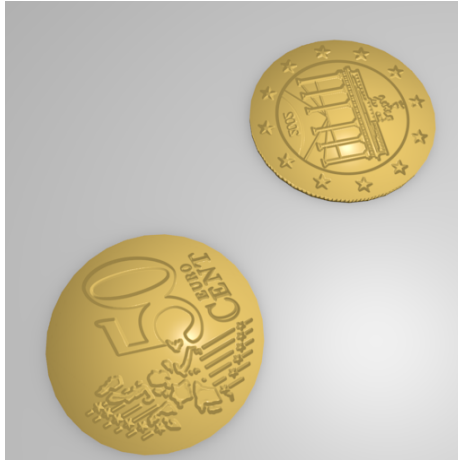


Figure 9: An example input image for the Coins (**T8**) experiment.

solution is shown in Table 7. We see that with no additionally labeled examples, DeepProbLog doesn’t converge on a satisfactory solution in about half of all runs. When it does converge on a solution, it converges on the *expected* solution 25% of the time, and on different solutions 75% of the time, which is conform with our expectations. We can also see that as the number of additionally labeled examples increases, DeepProbLog converges more reliably, and more on the expected solution. Starting with 10 additionally labeled examples, DeepProbLog reliably converges on the desired solution. Beyond 20 additionally labeled examples, we don’t see any further improvements.

**T9:** `0.8::poker([Q♥, Q♦, A♦, K♣],loss).`

In this last experiment we demonstrate that DeepProbLog can perform combined probabilistic reasoning, probabilistic learning and deep learning. We do this by playing a simplified poker game: there are two players that are dealt two cards from several decks. There is also one community card. Each player then makes a poker hand (e.g. pair, straight, ...) with their

| Distribution | Jack              | Queen             | King              | Ace               |
|--------------|-------------------|-------------------|-------------------|-------------------|
| Actual       | 0.2               | 0.4               | 0.15              | 0.25              |
| Learned      | $0.203 \pm 0.002$ | $0.396 \pm 0.002$ | $0.155 \pm 0.003$ | $0.246 \pm 0.002$ |

Table 8: The results for the Poker experiment (**T9**).

two cards and the community card.

For simplicity, we only use the jack, queen, king and ace. We also do not consider the suits of the cards.

The input consists of 4 images that show the cards dealt to the two players. Additionally, every example is labeled with the chance that the game is won, lost or ended in a draw, e.g.:

$$0.8 :: \text{poker}([\text{Q}\heartsuit, \text{Q}\diamond, \text{A}\diamond, \text{K}\clubsuit], \text{loss})$$

We expect DeepProbLog to:

- train the neural network to recognize the four cards
- reason probabilistically about the non-observed card
- learn the distribution of the unlabeled community card

To make DeepProbLog converge more reliably, we add some examples with additional supervision. Namely, in 10% of the examples we additionally specify the community card, i.e.

$$\text{poker}([\text{Q}\heartsuit, \text{Q}\diamond, \text{A}\diamond, \text{K}\clubsuit], \text{A}\diamond, \text{loss}).$$

This also showcases one of the strengths of DeepProbLog, namely, it can make use of examples that have different levels of observability. The loss function used in this experiment is the MSE between the predicted and target probabilities.

*Results.* We ran the experiment 10 times. Out of these 10 runs, 4 didn’t converge on the correct solution. The average values of the learned parameters for the remaining 6 runs are shown Table 8. As can be seen, DeepProbLog is able to correctly learn the probabilistic parameters. In these 6 runs, the neural network also correctly learned to classify all card types, achieving a 100% accuracy. The other runs did not converge because some of the classes were permuted (i.e., queens predicted as aces and vice versa) or multiple classes mapped onto the same one (queens and kings were both predicted as kings).

#### 6.4. Implementation details

For the implementation we integrate ProbLog2 [28] with PyTorch [29]. All programs are listed in Appendix A. In experiments **T1-T8** we optimize the cross-entropy loss between the predicted and target query probabilities, as we found that this works better to learn the probabilistic parameters. In experiment **T9** we optimize the MSE between the predicted and target query probabilities. We use Adam [30] optimization for the neural networks, and SGD for the logic parameters. For **T9**, we add random rotations (max 10 degrees) and shift the colours in the HSV by up to 5% to the images of the cards.

The neural network architectures are summarized in Table 11.  $Conv(o,k)$  denotes a convolutional layer with  $o$  output channels and kernel size  $k$ .  $Lin(n)$  denotes a fully connected layer of size  $n$ .  $BiGRU(h)$  denotes a single-layer bi-directional GRU with a hidden size  $h$ .  $(layer) \times 2$  means that there are two identical layers in parallel that are concatenated. A layer in bold means it is followed by a ReLU activation function. All neural networks end with a Softmax layer, unless otherwise specified. The hyperparameters used in the experiments are shown in Table 9. The sizes of the datasets used are specified in Table 10.

The regularisation term used in **T3** is calculated per network and per batch on the average of the neural network output. It is calculated as

$$1.0 - H_n \left( \frac{1}{N} \sum_{i=1}^n P_i \right) \quad (18)$$

where  $P_i$  is the  $i$ -th output of the neural network and  $H_n$  is the  $n$ -ary entropy (i.e. entropy using the base- $n$  logarithm).

#### 6.5. Computation time

Due to the nature of the exact inference used in DeepProbLog, which it inherits from aProbLog [22], the grounding and compilation steps can become expensive as the problem size grows. For example, in **T1**, which is the smallest problem we consider, grounding and compilation takes on average 0.01 seconds, while evaluating the NN and AC takes on average 0.002 seconds per example. In **T9**, which has the largest logic program out of all experiments, grounding and compilation takes on average 1.3 seconds, while NN and AC evaluation takes on average 0.05 seconds per example.

It is important to note that when we evaluate an example a second time, the structure of the AC, which is determined by the grounding and compilation, remains the same. Only the learned probabilities in the nAD change. We make use of this to improve the performance by caching the arithmetic circuits so that we only have to perform the (potentially expensive) grounding and compilation steps once. During evaluation, we only re-evaluate the neural networks and evaluate the AC with the updated probabilities.

Note that this optimization can also be applied to, for example, the queries  $\text{addition}(\text{3}, \text{5}, 8)$  and  $\text{addition}(\text{7}, \text{0}, 8)$ , as both have the same structure.

| Task         | Batch size | Learning rate | Parameter learning rate | Infloss |
|--------------|------------|---------------|-------------------------|---------|
| <b>T1-T3</b> | 2          | 1e-3          |                         |         |
| <b>T4</b>    | 2          | 1e-3          | 1e-3                    | 2, 4, 8 |
| <b>T5</b>    | 50         | 0.02          |                         |         |
| <b>T6</b>    | 16         | 1.0           |                         |         |
| <b>T7</b>    | 100        | 0.005         |                         |         |
| <b>T8</b>    | 5          | 1e-4          |                         | 0.25    |
| <b>T9</b>    | 50         | 1e-4          | 1e-3                    | 0.5     |

Table 9: Overview of the hyperparameters used in the experiments.

| Task      | Training set       | Validation Set | Test set |
|-----------|--------------------|----------------|----------|
| <b>T1</b> | 29 500, 3 000, 300 | 500            | 5 000    |
| <b>T2</b> | 14750, 1 500, 150  | 250            | 2 500    |
| <b>T3</b> | 16 000             | 2 000          | 3 000    |
| <b>T4</b> | 29 500             | 500            | 5 000    |
| <b>T5</b> | 512                | 256            | 1 024    |
| <b>T6</b> | 256                | 32             | 32       |
| <b>T7</b> | 300                | 100            | 200      |
| <b>T8</b> | 100                | –              | 20       |
| <b>T9</b> | 500                | –              | 25       |

Table 10: Overview of the sizes of the datasets used in the experiments.

To do this, we introduce placeholder constants and change both queries to the single query `addition(a,b,8)`, which reduces all queries in **T1** to 19 unique queries, one for each different label. During the evaluation of the neural networks, we replace the constants with the correct input and use the resulting probabilities in the cached ACs. We apply these optimizations to all experiments.

| Task   | Network  | Architecture  |
|--|----------|---|
| <b>T1-T4</b>   | digit/2  | MNISTConv, <b>Lin(120)</b> , <b>Lin(84)</b> , Lin(10)                             |
| <b>T1,T3</b>   | baseline | MNISTConv $\times 2$ , <b>Lin(120)</b> , <b>Lin(84)</b> , Lin(19)                 |
| <b>T2</b>  | baseline | (MNISTConv $\times 2$ , <b>Lin(100)</b> ) $\times 2$ , <b>Lin(128)</b> , Lin(199) |
| <b>T5</b>  | result/4 | Lin(50), TanH, Lin(10)  |
|  | carry/4  | Lin(10), TanH, Lin(2)   |
| <b>T6</b>  | swap/3   | <b>Lin(20)</b> , Lin(10)  |
| <b>T7</b>  | RNN      | Embedding(256), BiGRU(512), Dropout(0.5)*   |
|  | perm/2   | Lin(6)  |
|  | op1/2    | Lin(4)  |
|  | swap/2   | Lin(2)  |
|  | op2/2    | Lin(4)  |
| <b>T8</b>  | coin1/2  | AlexNetConv, <b>Lin(100)</b> , Lin(2)   |
|  | coin2/2  | AlexNetConv, <b>Lin(100)</b> , Lin(2)   |
| <b>T9</b>  | rank/2   | AlexNetConv, <b>Lin(100)</b> , Lin(4)   |
| MNISTConv: Conv(6,5), <b>MP(2,2)</b> , Conv(16,5), <b>MP(2,2)</b> *  |          |   |
| AlexNetConv: <b>Conv(64, 11, 2,2)</b> , MP(3,2), <b>Conv(192, 5, 2)</b> , MP(3,2), <b>Conv(384, 3, 1)</b> , <b>Conv(256, 3, 1)</b> , <b>Conv(256, 3, 1)</b> , MP(3,2)* |          |   |
| * Does not end with a Softmax layer.   |          |   |

Table 11: Overview of the neural network architectures used in the experiments.

## 7. Related Work

Most of the work on combining neural networks and logical reasoning comes from the *neuro-symbolic reasoning* literature [7, 31]. These approaches typically focus on approximating logical reasoning with neural networks by encoding logical terms in Euclidean space. However, they neither support probabilistic reasoning nor perception, and are often limited to non-recursive and acyclic logic programs [32]. DeepProbLog takes a different approach and integrates neural networks into a probabilistic logic framework, retaining the full power of both logical and probabilistic reasoning and deep learning.

At the same time, DeepProbLog also integrates probability in neuro-symbolic computation. Although this may appear as a complication, our work actually shows that it can greatly simplify the integration of neural networks with logic. The reason for this is that the probabilistic framework provides a clear optimisation criterion, namely the probability of the training examples. Real-valued probabilistic quantities are also well-suited for gradient-based training procedures, as opposed to discrete logic quantities.

The prominent recent lines of related work focus on three main branches: pushing the logic as regularisation, templating neural networks, and neural program induction.

### 7.1. Logic as regularisation

The main idea behind this line of research is that logic is included as a regularizer during the optimization of the neural network, or the learning of the embeddings. The goal is to encode the logic into the weights so that after training, when the logic is no longer explicitly present, the evaluation still shows the characteristics of the logic. [33, 34, 35, 36, 37, 38] all center around including logical background knowledge as a regularizer during training. Rocktäschel et al. [33] inject background knowledge into a matrix factorization model for relation extraction, by adding differentiable loss terms for propositionalized first-order rules. Demeester et al. [34] propose a more efficient alternative by inducing order relations in embedding space, effectively leading to a lifted application of the rules. This is further generalized by Minervini et al. [35], who investigate injecting rules by minimizing an inconsistency loss on adversarially-generated examples. Diligenti et al. [36] use FOL to specify constraints on the output of the neural network. They use fuzzy logic to create a differentiable way of measuring how much the output of the neural networks violates these constraints. This is then added as an additional loss term that acts as a regularizer. More recent work by Xu et al. [38] introduces a similar method that uses probabilistic logic instead of fuzzy logic, and is thus more similar to DeepProbLog. They also compile the formulas to an SDD for efficiency.

However, whereas DeepProbLog is based on (probabilistic) logic programming, these methods use first order logic instead. This is reminiscent to the difference between ProbLog and Markov Logic [39] or PSL [40]. Donadello et al. [37], though at first sight related to Diligenti et al. [36], work slightly differently. They learn functions that map numerical properties of logical constants onto truth values, which are then combined using fuzzy logic.

### 7.2. Templating neural networks

This line of work uses the logic as a template for constructing the architecture of neural networks. This is reminiscent of the knowledge base construction approach of statistical relational artificial intelligence [4].

Rocktäschel and Riedel [9] introduce a differentiable framework for theorem proving. They re-implemented Prolog’s theorem proving procedure in a differentiable manner and enhanced it with learning subsymbolic representation of the existing symbols, which are used to handle noise in data. Whereas Rocktäschel and Riedel use logic only to construct a neural network and focus on learning subsymbolic representations, DeepProbLog focuses on tight interactions between the two and parameter learning for both the neural and the logic components. In this way, DeepProbLog retains the best abilities of both worlds. Recently, Weber et al. [41] extend the notion of soft unification towards structured textual knowledge, i.e., unification can be performed between sentences, not only symbols. In contrast to Rocktäschel and Riedel [9], Weber et al. [41] retain the full ability of logical reasoning, and as such is closer to DeepProbLog, but it is specialised for NLP tasks.

Cohen et al. [10] introduce a framework to compile a tractable subset of logic programs into differentiable functions and to execute it with neural networks. It provides an alternative probabilistic logic but it has a different and less developed semantics. Furthermore, to the best of our knowledge it has not been applied to the kind of tasks tackled in the present paper. An idea similar in spirit to ours is that of Andreas et al. [42], who introduce a neural network for visual question answering composed out of smaller modules responsible for individual tasks, such as object detection. Whereas the composition of modules is determined by the linguistic structure of the questions, DeepProbLog uses probabilistic logic programs to connect the neural networks.

### 7.3. Neural program induction.

The third line of work has focused on learning programs from data by combining neural and symbolic approaches.

*Neural execution.* The first category captures a program behaviour with neural networks and therefore focuses on program execution. The approach most similar to ours is that of Bošnjak et al. [8], where neural networks are used to fill in *holes* in a partially defined Forth program. DeepProbLog differs in that it uses ProbLog as the host language which results in native support for both logical and probabilistic reasoning, while differentiable Forth uses a procedural language. Differentiable Forth has been applied to tasks T5-7, but it is unclear whether it could be applied to the remaining ones. Finally, Evans and Grefenstette [43] introduce a differentiable framework for rule induction, that does not focus on the integration of the two approaches like DeepProbLog.

*Neurally guided search.* The second line of research enhances the search procedures of the symbolic program induction techniques by incorporating neural components in the search itself. The key principle these techniques employ is to perform the search over programs in a systematic symbolic way, but guide the search with a heuristic learned by a deep neural network. For instance, Kalyan et al. [44] train a neural network to predict the scores of branches during the branch-and-bound search procedure, Zhang et al. [45] train a neural network to choose which candidate program to expand next while exploiting the constraints on the input-output examples, while Ellis et al. [46] use a neural network to efficiently search over a well-designed DSL.

*Neural program construction.* The final category involves techniques that decompose a problem into independent parts that can be individually solved by either neural or symbolic components and synchronize the individual components to solve the main problem. For instance, Yi et al. [47], Mao et al. [48] develop a neuro-symbolic approach towards visual question answering by using a neural network to generate a program computing the answer to the question and executing the program symbolically. Ellis et al. [49] generate a  $\text{\LaTeX}$  code from a hand-drawn sketch by using a neural network to recognise basic shapes within



a sketch and symbolically inducing the program describing the sketch. In contrast, Dong et al. [50] induce programs in a purely neural way and demonstrate favourable generalization; however, they do not induce symbolic programs, but rather express a program through a neural network.

#### 7.4. Symbolic deep learning

The success of neural deep learning has inspired several works introducing symbolic deep learning methods which, instead of representing the logical aspects in a vector space, retain the logical data representation in the latent representation. These include the symbolic versions of deep neural networks: Šourek et al. [51] treat symbolic rules expressed in first-order logic as a template for constructing a neural network, while Kazemi and Poole [52] compose a relational neural network by adding hidden layers to relational logistic regression [53]. Another research direction focuses on task-agnostic discovery of relational (symbolic) latent representations by exploiting approximate symmetries [54], a symbolic extension of the auto-encoding principle [55], or self-play [56].

## 8. Conclusion

We introduced DeepProbLog, a framework where neural networks and probabilistic logic programming are integrated in a way that exploits the full expressiveness and strengths of both worlds and can be trained end-to-end based on examples. This was accomplished by extending an existing probabilistic logic programming language, ProbLog, with neural predicates. Learning is performed by using aProbLog to calculate the gradient of the loss which is then used in standard gradient-descent based methods to optimize parameters in both the probabilistic logic program and the neural networks. We evaluated our framework on experiments that demonstrate its capabilities in combined symbolic and subsymbolic reasoning, program induction, and probabilistic logic programming.

Although we have shown promising results, DeepProbLog is currently using only exact inference. As exact inference does not always scale well and can be prohibitively expensive for large problems, the DeepProbLog implementation cannot yet be applied to problems such as KB completion. Future work will be concerned with incorporating approximate inference algorithms to speed-up the grounding and compilation process.

## Acknowledgements

RM is a SB PhD fellow at FWO (1S61718N). SD is supported by the Research Fund KU Leuven (GOA/13/010) and Research Foundation - Flanders (G079416N). This work has been partially supported by the European Research Council Advanced Grant project SYNTH (ERCAAdG-694980).

## References

- [1] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, L. De Raedt, Deep-problog: Neural probabilistic logic programming, in: *Advances in Neural Information Processing Systems*, 2018, pp. 3749–3759.
- [2] D. Kahneman, *Thinking, fast and slow*, Farrar, Straus and Giroux New York, 2011.
- [3] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, in: *Advances in Neural Information Processing Systems*, volume 30, 2017, pp. 4974–4983.
- [4] L. De Raedt, K. Kersting, S. Natarajan, D. Poole, Statistical relational artificial intelligence: Logic, probability, and computation, *Synthesis Lectures on Artificial Intelligence and Machine Learning* 10 (2016) 1–189.
- [5] L. Getoor, B. Taskar, *Introduction to statistical relational learning*, MIT press, 2007.
- [6] L. De Raedt, A. Kimmig, Probabilistic (logic) programming concepts, *Machine Learning* 100 (2015) 5–47.
- [7] A. S. d. Garcez, K. B. Broda, D. M. Gabbay, *Neural-symbolic learning systems: foundations and applications*, Springer Science & Business Media, 2012.
- [8] M. Bošnjak, T. Rocktäschel, S. Riedel, Programming with a differentiable forth interpreter, in: *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 2017, pp. 547–556.
- [9] T. Rocktäschel, S. Riedel, End-to-end differentiable proving, in: *Advances in Neural Information Processing Systems*, volume 30, 2017, pp. 3788–3800.
- [10] W. W. Cohen, F. Yang, K. R. Mazaitis, Tensorlog: Deep learning meets probabilistic databases, *Journal of Artificial Intelligence Research* 1 (2018) 1–15.
- [11] L. De Raedt, R. Manhaeve, S. Dumancic, T. Demeester, A. Kimmig, Neuro-symbolic= neural+ logical+ probabilistic, in: *NeSy’19@ IJCAI, the 14th International Workshop on Neural-Symbolic Learning and Reasoning*, 2019, pp. 1–4.
- [12] L. De Raedt, A. Kimmig, H. Toivonen, ProbLog: A probabilistic Prolog and its application in link discovery, in: *IJCAI*, 2007, pp. 2462–2467.
- [13] J. W. Lloyd, *Foundations of Logic Programming*, 2. ed., Springer, 1989.
- [14] A. Van Gelder, K. A. Ross, J. S. Schlipf, The well-founded semantics for general logic programs, *Journal of the ACM* 38 (1991) 620–650.

- [15] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.  
<http://www.deeplearningbook.org>.
- [16] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann Publishers Inc., 1988.
- [17] D. Fierens, G. Van den Broeck, J. Renkens, D. Shterionov, B. Gutmann, I. Thon, G. Janssens, L. De Raedt, Inference and learning in probabilistic logic programs using weighted Boolean formulas, *Theory and Practice of Logic Programming* 15 (2015) 358–401.
- [18] A. Darwiche, P. Marquis, A knowledge compilation map, *Journal of Artificial Intelligence Research* 17 (2002) 229–264.
- [19] A. Darwiche, SDD: A new canonical representation of propositional knowledge bases, in: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, IJCAI-11*, 2011, pp. 819–826.
- [20] M. Frazier, L. Pitt, Learning from entailment: An application to propositional horn sentences, in: *Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June 27-29, 1993*, 1993, pp. 120–127.
- [21] B. Gutmann, A. Kimmig, K. Kersting, L. De Raedt, Parameter learning in probabilistic databases: A least squares approach, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2008, pp. 473–488.
- [22] A. Kimmig, G. Van den Broeck, L. De Raedt, An algebraic Prolog for reasoning about possible worlds., in: *AAAI*, 2011.
- [23] J. Eisner, Parameter estimation for probabilistic finite-state transducers, in: *Proceedings of the 40th annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2002, pp. 1–8.
- [24] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in: *Proceedings of the IEEE*, 1998, pp. 2278–2324.
- [25] A. Solar-Lezama, Program sketching, *International Journal on Software Tools for Technology Transfer* 15 (2013) 475–495.
- [26] S. Reed, N. de Freitas, Neural programmer-interpreters, in: *International Conference on Learning Representations (ICLR)*, 2016.
- [27] S. Roy, D. Roth, Solving general arithmetic word problems, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1743–1752.

- [28] A. Dries, A. Kimmig, W. Meert, J. Renkens, G. Van den Broeck, J. Vlasselaer, L. De Raedt, Problog2: Probabilistic logic programming, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2015, pp. 312–315.
- [29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, in: Proceedings of the Workshop on The future of gradient-based machine learning software and techniques, co-located with the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), 2017.
- [30] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR), 2015, pp. 1–13.
- [31] B. Hammer, P. Hitzler, Perspectives of neural-symbolic integration, volume 8, Springer Heidelberg:, 2007.
- [32] S. Hölldobler, Y. Kalinke, H.-P. Störr, Approximating the semantics of logic programs by recurrent neural networks, *Applied Intelligence* 11 (1999) 45–58.
- [33] T. Rocktäschel, S. Singh, S. Riedel, Injecting logical background knowledge into embeddings for relation extraction, in: NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1119–1129.
- [34] T. Demeester, T. Rocktäschel, S. Riedel, Lifted rule injection for relation embeddings, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1389–1399.
- [35] P. Minervini, T. Demeester, T. Rocktäschel, S. Riedel, Adversarial sets for regularised neural link predictors, in: Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI), 2017.
- [36] M. Diligenti, M. Gori, C. Sacca, Semantic-based regularization for learning and inference, *Artificial Intelligence* 244 (2017) 143–165.
- [37] I. Donadello, L. Serafini, A. S. d’Avila Garcez, Logic tensor networks for semantic image interpretation, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, 2017, pp. 1596–1602.
- [38] J. Xu, Z. Zhang, T. Friedman, Y. Liang, G. V. den Broeck, A semantic loss function for deep learning with symbolic knowledge, in: Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, 2018, pp. 5498–5507.

- [39] M. Richardson, P. Domingos, Markov logic networks, *Machine learning* 62 (2006) 107–136.
- [40] S. H. Bach, M. Broecheler, B. Huang, L. Getoor, Hinge-loss markov random fields and probabilistic soft logic, *arXiv preprint arXiv:1505.04406* (2015).
- [41] L. Weber, P. Minervini, J. Münchmeyer, U. Leser, T. Rocktäschel, Nl-prolog: Reasoning with weak unification for question answering in natural language, in: *Proceedings of ACL 2018, Tutorial Abstracts*, 2019.
- [42] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Neural module networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 39–48.
- [43] R. Evans, E. Grefenstette, Learning explanatory rules from noisy data, *Journal of Artificial Intelligence Research* 61 (2018) 1–64.
- [44] A. Kalyan, A. Mohta, O. Polozov, D. Batra, P. Jain, S. Gulwani, Neural-guided deductive search for real-time program synthesis from examples, in: *ICLR*, 2018.
- [45] L. Zhang, G. Rosenblatt, E. Fetaya, R. Liao, W. E. Byrd, M. Might, R. Urtasun, R. Zemel, Neural guided constraint logic programming for program synthesis, in: *NeurIPS*, 2018.
- [46] K. Ellis, L. Morales, M. Sablé-Meyer, A. Solar-Lezama, J. Tenenbaum, Learning libraries of subroutines for neurally-guided bayesian program induction, in: *NeurIPS*, 2018.
- [47] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, J. B. Tenenbaum, Neural-symbolic vqa: Disentangling reasoning from vision and language understanding, in: *NeurIPS*, 2018.
- [48] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, J. Wu, The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision, in: *ICLR*, 2019.
- [49] K. Ellis, D. Ritchie, A. Solar-Lezama, J. Tenenbaum, Learning to infer graphics programs from hand-drawn images, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31*, Curran Associates, Inc., 2018, pp. 6059–6068.
- [50] H. Dong, J. Mao, T. Lin, C. Wang, L. Li, D. Zhou, Neural logic machines, in: *ICLR*, 2019.
- [51] G. Šourek, V. Aschenbrenner, F. Železný, S. Schockaert, O. Kuželka, Lifted relational neural networks: Efficient learning of latent relational structures, *Journal of Artificial Intelligence Research* to appear (2018).

- [52] S. M. Kazemi, D. Poole, RelNN: A deep neural model for relational learning, in: AAAI, 2018.
- [53] S. M. Kazemi, D. Buchman, K. Kersting, S. Natarajan, D. Poole, Relational logistic regression: The directed analog of markov logic networks, in: Proceedings of the 13th AAAI Conference on Statistical Relational AI, AAAIWS'14-13, AAAI Press, 2014, pp. 41–43.
- [54] S. Dumančić, H. Blockeel, Clustering-based relational unsupervised representation learning with an explicit distributed representation, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 1631–1637.
- [55] S. Dumančić, T. Guns, W. Meert, H. Blockeel, Learning relational representations with auto-encoding logic programs, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, 2019, p. To appear.
- [56] A. Cropper, Playgol: Learning programs through play, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, 2019, p. To appear.

## Appendix A. DeepProbLog Programs

---

```
nn(m_digit,[X],Y,[0,...,9]) :: digit(X,Y).

addition(X,Y,Z) :- digit(X,X2), digit(Y,Y2), Z is X2+Y2.
```

---

Listing 1: Single-digit MNIST addition (**T1**)

In Listing 1, `digit/2` is the neural predicate that classifies an MNIST image into the integers 0 to 9. The `addition/3` predicate’s first two arguments are MNIST digits, and the last is the sum. It classifies both images using `digit/2` and calculates the sum of the two results.

---



```
nn(m_digit,[X],Y,[0,...,9]) :: digit(X,Y).

number([],Result,Result).
number([H|T],Acc,Result) :-
    digit(H,Nr),
    Acc2 is Nr+10*Acc,
    number(T,Acc2,Result).
number(X,Y) :- number(X,0,Y).

multi_addition(X,Y,Z) :- number(X,X2), number(Y,Y2), Z is X2+Y2.
```

---

Listing 2: Multi-digit MNIST addition (**T2**)

In Listing 2, the only difference with Listing 1 is that the `multi_addition/3` predicate now uses the `number/2` predicate instead of the `digit/2` predicate. The `number/3` predicate’s first argument is a list of MNIST images. It uses the `digit/2` neural predicate on each image in the list, summing and multiplying by ten to calculate the number represented by the list of images (e.g. `number([,38)`).

---

```
nn(m_digit,[X],Y,[0,...,9]) :: digit(X,Y).

addition(X,Y,Z) :- digit(X,X2), digit(Y,Y2), digit(Z,Z2), Z2 is X2+Y2.
```

---

Listing 3: All-digit MNIST addition (**T3**)

In Listing 3, the only difference with Listing 1 is that all 3 inputs `X,Y,Z` are images. As such, the `digit/2` predicate is also used on the third input. The sum is also redefined as `Z2 is X2+Y2`.

In Listing 4, an additional probabilistic fact (`noisy/2`) is added that encodes the chance of an example being noisy. The `addition/3` predicate is split into two cases: when the noisy is true or when noisy is false. The latter is the same as in Listing 1. If `noisy` is true, `Z` is considered to be drawn from the uniform distribution (`uniform/3`).

---

```

nn(classifier, [X], Y, [0 .. 9]) :: digit(X,Y).
t(0.2) :: noisy.

1/19 :: uniform(X,Y,0) ; ... ; 1/19 :: uniform(X,Y,18).

addition(X,Y,Z) :- noisy, uniform(X,Y,Z).
addition(X,Y,Z) :- \+noisy, digit(X,N1), digit(Y,N2), Z is N1+N2.

```

---

Listing 4: Noisy MNIST addition (**T4**)

---

```

nn(m_result, [D1,D2,Carry], Y, [0,...,9]) :: result(D1,D2,Carry,Y).

nn(m_carry, [D1,D2,Carry], Y, [0,1]) :: carry(D1,D2,Carry,Y).

slot(I1,I2,Carry,NewCarry,Result) :-
    result(I1,I2,Carry,Result),
    carry(I1,I2,Carry,NewCarry).

add([], [], [C], C, []).

add([H1|T1], [H2|T2], C, Carry, [Digit|Res]) :-
    add(T1,T2,C,NewCarry,Res),
    slot(H1,H2,NewCarry,Carry,Digit).

forth_addition(L1,L2,C, [Carry|Res]) :- add(L1,L2,C,Carry,Res).

```

---

Listing 5: Forth addition sketch (**T5**)

In Listing 5, there are two neural predicates: **result/4** and **carry/4**. These are used in the **slot/4** predicate that corresponds to the slot in the Forth program. The first three arguments are the two digits and the previous carry to be summed. The next two arguments are the new carry and the new resulting digit. The **add/5** predicate's arguments are: the two list of input digits, the input carry, the resulting carry and the resulting sum. It recursively calls itself to loop over both lists, calling the **slot/5** predicate on each position, using the carry from the previous step.

In Listing 6, there's a single neural predicate: **swap/3**. Its first two arguments are the numbers that are compared, the last argument is an indicator whether to swap or not. The **bubble/3** predicate performs a single step of bubble sort on its first argument using the **hole/4** predicate. The second argument is the resulting list after the bubble step, but without its last element, which is the third argument. The **bubblesort/3** predicate uses the **bubble/3** predicate, and recursively calls itself on the remaining list, adding the last element on each step to the front of the sorted list.

In Listing 7, there are four neural predicates: **net1/2** to **net4/2**. Their first argument is the input question, and the second argument are indicator variables for the choice of respectively: one of six permutations, one of 4 operations, swapping and one of 4 operations. These are implemented in the **permute/7**, **swap/5** and **operator/4** predicates. The **wap/5** predicate then sequences these steps to



---

```

nn(m_swap, [X]) :: swap(X,Y).

hole(X,Y,X,Y) :- \+swap(X,Y).

hole(X,Y,Y,X) :- swap(X,Y).

bubble([X], [], X).
bubble([H1,H2|T], [X1|T1], X) :-
    hole(H1,H2,X1,X2),
    bubble([X2|T], T1, X).

bubblesort([], L, L).

bubblesort(L, L3, Sorted) :-
    bubble(L, L2, X),
    bubblesort(L2, [X|L3], Sorted).

forth_sort(L, L2) :- bubblesort(L, [], L2).

```

---

Listing 6: Forth sorting sketch (T6)

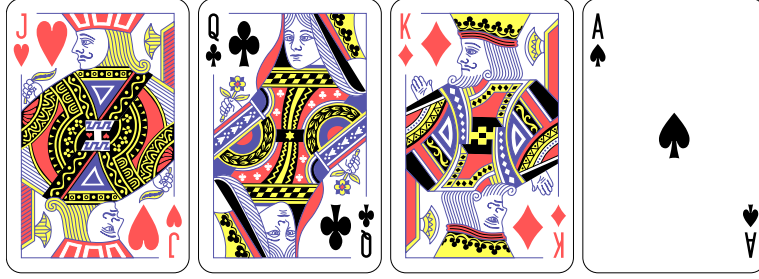


Figure A.10: Examples of cards used as input for the Poker without perturbations(T9) experiment.

calculate the result.

In Listing 8, there are two neural predicates: `coin1/2` and `coin2/2`. Their input is the image of the two coins (e.g. Figure 9). The output is heads or tails. The `coins/2` classifies both coins using these two predicates and then performs the comparison of the classes with the `compare/3` predicate.

In Listing 9, there's a single neural predicate `rank/2` that takes as input the image of a card and classifies it as either a jack, queen, king or ace. There's also an AD with learnable parameters that represents the distribution of the unseen community card (`house_rank/1`). The `hand/2` predicate's first argument is a list of 3 cards. It unifies the output with any of the valid hands that these cards contain. The valid hands are: high card, pair (two cards have the same rank), three of a kind (three cards have the same rank), low straight (jack, queen king) and high straight(queen, king, ace). Each hand is assigned a rank

---

```

permute(0,A,B,C,A,B,C).
permute(1,A,B,C,A,C,B).
permute(2,A,B,C,B,A,C).
permute(3,A,B,C,B,C,A).
permute(4,A,B,C,C,A,B).
permute(5,A,B,C,C,B,A).

swap(0,X,Y,X,Y).
swap(1,X,Y,Y,X).

operator(0,X,Y,Z) :- Z is X+Y.
operator(1,X,Y,Z) :- Z is X-Y.
operator(2,X,Y,Z) :- Z is X*Y.
operator(3,X,Y,Z) :- Y > 0, 0 := X mod Y,Z is X//Y.

nn(m_net1,[Repr],Y,[0,...,5])::net1(Repr,Y).
nn(m_net2,[Repr],Y,[0,...,3])::net2(Repr,Y).
nn(m_net3,[Repr],Y,[0,1])::net3(Repr,Y).
nn(m_net4,[Repr],Y,[0,...,3])::net4(Repr,Y).

wap(Text,X1,X2,X3,Out) :-
    net1(Text,Perm),
    permute(Perm,X1,X2,X3,N1,N2,N3),
    net2(Text,Op1),
    operator(Op1,N1,N2,Res1),
    net3(Text,Swap),
    swap(Swap,Res1,N3,X,Y),
    net4(Text,Op2),
    operator(Op2,X,Y,Out).

```

---

Listing 7: Forth WAP sketch (**T7**)

with the `hand_rank/2` predicate. The `best_hand_rank/2` predicate takes as input a list of cards, and unifies the second argument with the highest hand rank that is possible with the three given cards. The `outcome/3` predicate determines the outcome by comparing the two ranks of the best hand. The `game/3` predicate's first argument is a list of the 4 input images. It's second input is the labeled community card. It classifies the cards using the neural predicates, determines the best rank, and then unifies the last argument with the outcome. The `game/2` determines the community card from the learned distribution `house_rank/1`, and then determines the outcome using the `game/3` predicate. The `member/2` and `select/3` predicates are predicates from the *lists* library. `member/2` is true if it's second argument is a list and the first argument appears in that list. `select/3` is true if it's second argument is a list and the first argument appears in that list. It also unifies the last argument with the list that is the same as it's second argument, but with the first argument removed.

---

```
nn(net1, [X], Y, [heads, tails]) :: coin1(X,Y).
nn(net2, [X], Y, [heads, tails]) :: coin2(X,Y).

compare(X,X,same).
compare(X,Y,different) :- \+compare(X,Y,same).

coins(X,Comparison) :-
    coin1(X,C1),
    coin2(X,C2),
    compare(C1,C2,Comparison).
```

---

Listing 8: The coins experiment (**T8**)

---

```

t(1/4)::house_rank(jack);t(1/4)::house_rank(queen);
    t(1/4)::house_rank(king);t(1/4)::house_rank(ace).
nn(net1,[X],Y,[jack,queen,king,ace]):- rank(X,Y).

hand(Cards,straight(low)) :-
    member(card(jack),Cards),
    member(card(queen),Cards),
    member(card(king),Cards).
hand(Cards,straight(high)) :-
    member(card(queen),Cards),
    member(card(king),Cards),
    member(card(ace),Cards).
hand([card(R), card(R), card(R)],threeofakind(R)).
hand(Cards,pair(R)) :-
    select(card(R),Cards,Cards2),
    member(card(R),Cards2).
hand(Cards,high(R)) :-
    member(card(R),Cards).

hand_rank(high(jack),0).
...
hand_rank(straight(high),13).

best_hand_rank(Cards,R) :-
    hand(Cards,H),
    hand_rank(H,R),
    \+(hand(Cards,H2),hand_rank(H2,R2),R2>R).

outcome(R1,R2,win) :- R1 > R2.
outcome(R1,R2,loss) :- R1 < R2.
outcome(R,R,draw).

cards(C1,C2,House,[card(R1), card(R2), House]) :-
    rank(C1,R1),
    rank(C2,R2).

game([C1,C2,C3,C4],House,Outcome) :-
    cards(C1,C2,House,Hand1),
    cards(C3,C4,House,Hand2),
    best_hand_rank(C1,R1),
    best_hand_rank(C2,R2),
    outcome(R1,R2,Outcome).

game(Cards,Outcome) :-
    house_rank(House),
    game(Cards,House,Outcome).

```

---

Listing 9: The Poker experiment (**T9**)