Learning Reasoning Strategies in End-to-End Differentiable Proving

Pasquale Minervini 1  Sebastian Riedel 1 2  Pontus Stenetorp 1  Edward Grefenstette 1 2  Tim Rocktäschel 1 2

Abstract
Attempts to render deep learning models interpretable, data-efficient, and robust have seen some success through hybridisation with rule-based systems, for example, in Neural Theorem Provers (NTPs). These neuro-symbolic models can induce interpretable rules and learn representations from data via back-propagation, while providing logical explanations for their predictions. However, they are restricted by their computational complexity, as they need to consider all possible proof paths for explaining a goal, thus rendering them unfit for large-scale applications. We present Conditional Theorem Provers (CTPs), an extension to NTPs that learns an optimal rule selection strategy via gradient-based optimisation. We show that CTPs are scalable and yield state-of-the-art results on the CLUTRR dataset, which tests systematic generalisation of neural models by learning to reason over smaller graphs and evaluating on larger ones. Finally, CTPs show better link prediction results on standard benchmarks in comparison with other neural-symbolic models, while being explainable. All source code and datasets are available online.

1. Introduction
Neural Natural Language Understanding (NLU) systems—wherein a deep neural network is used as a function approximator (LeCun et al., 2015; Goodfellow et al., 2016)—have been extremely successful at various natural language tasks, such as Question Answering (QA) and Natural Language Inference (NLI) (Goldberg, 2017), achieving strong generalisation results on datasets available for these tasks (Seo et al., 2017; Hu et al., 2018; Shen et al., 2016; Huang et al., 2018). Even strong performance on NLU problems have been recently achieved with advent of large models pre-trained via self-supervision, such as BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), and RoBERTa (Liu et al., 2019).

Generalisation in Neural Models However, there are growing concerns about the ability of NLU systems, and neural networks more generally, to generalise in a systematic and robust way (Bahdanau et al., 2019; Lake & Baroni, 2018; Johnson et al., 2017; Sinha et al., 2019). For instance, Jia & Liang (2017) highlight the brittleness of NLU systems to adversarial examples, while Gururangan et al. (2018); Kaushik & Lipton (2018) show that neural NLU models tend to exploit annotation artefacts and spurious correlations in the data. Furthermore, analysing and supervising the inner workings of such models is not trivial, due to their inherent black-box nature (Doshi-Velez & Kim, 2017; Lipton, 2018).

More generally, Garnelo & Shanahan (2019) emphasise several limitations of neural models, in terms of:

1. Data inefficiency and high sample complexity—the need of high volumes of training data in order to be effective,
2. Poor generalisation—modern neural models may not produce the correct predictions when exposed to data outside the training distribution, and
3. Lack of interpretability—such models are black boxes where internal representations and computations are hardly interpretable by humans.

In this vein, Sinha et al. (2019) measured and compared the systematic generalisation abilities of several neural models (including very strong baselines such as BERT (Devlin et al., 2019) and Graph Attention Networks (GATs) (Velickovic et al., 2018)) on the task of answering questions about family relationship graphs, by evaluating on held-out combinations of reasoning patterns and by adding curated distracting noisy facts. Interestingly, they found that performance degrades monotonically for every model in their pool as they increase the complexity of the relational graph, highlighting the challenge of systematic generalisation (Lake & Baroni, 2018; Sodhani et al., 2018).

Neuro-Symbolic Reasoning A promising direction for overcoming these issues consists in combining neural models and symbolic reasoning given their complementary strengths and weaknesses (d’Avila Garcez et al., 2013).
We focus on NTPs (Rocktäschel & Riedel, 2017), a family of neuro-symbolic reasoning models. NTPs are continuous relaxations of the backward-chaining reasoning algorithm that replace discrete symbols with their continuous embedding representations. NTPs have interesting properties: they can jointly learn representations and interpretable rules from data via backpropagation, and can potentially combine such rules in ways that may have not been observed during training. However, a major limitation in NTPs is that, during training, they need to consider all rules for explaining a given goal or sub-goal. This quickly renders them ineffective in settings requiring a large number of rules or reasoning steps.

Conditional Theorem Provers For addressing limitations of NTPs, we propose CTPs, an extension that is able to learn an adaptive strategy for selecting subsets of rules to consider at each step of the reasoning process. This is achieved by a select module that, given a goal, produce the rules needed for proving it. Predicates and constants in the produced rules lie in a continuous embedding space. Hence, the select module is end-to-end differentiable, and can be trained jointly with the other modules via gradient-based optimisation.

2. End-to-End Differentiable Proving

NTPs (Rocktäschel & Riedel 2017) are a continuous relaxation of the backward chaining algorithm (Russell & Norvig 2010): this algorithm works backward from the goal, chaining through rules to find known facts supporting the proof.

Given a query (or goal) \( G \), backward chaining first attempts to unify it with the facts available in a given Knowledge Base (KB). If no matching fact is available, it considers all rules \( H \Rightarrow B \), where \( H \) denotes the head (or consequence) and \( B \) the body (or premise), and \( H \) can be unified with the query \( G \) resulting in a substitution for the variables contained in \( H \). Then, the backward chaining algorithm applies the substitution to the body \( B \), and recursively attempts to prove the atoms contained therein.

Backward chaining can be seen as a type of and/or search: or because the goal can be proven by any rule in the KB and and because all the conjuncts in the premise of a rule must be proven.

Example 2.1 (Backward Chaining). Consider a KB composed by the facts \( p(\text{RICK, BETH}) \) and \( p(\text{BETH, MORTY}) \), and by the rule \( g(X, Y) \Rightarrow p(X, Z), p(Z, Y) \), where \( p \) and \( g \) denote the relationships parent and grandparent, respectively. The goal \( G = g(\text{RICK, MORTY}) \) can be proven by unifying \( G \) with the head of the rule \( g(X, Y) \), with the substitution \( \{X/\text{RICK}, Y/\text{MORTY}\} \), and then by recursively proving the subgoals \( p(\text{RICK}, Z), p(Z, \text{MORTY}) \), which hold true for the substitution \( \{Z/\text{BETH}\} \).

NTPs make this reasoning process more flexible and end-to-end differentiable by replacing the comparison between symbols with a soft matching of their respective trainable dense vector representations. NTPs recursively build a neural network enumerating all possible proof paths for proving a query given \( \text{KB} \) and aggregate all proof scores via max pooling. They do so by relying on three modules: a unification module, which compares sub-symbolic representations of logic atoms, and mutually recursive or and and modules, which jointly enumerate all possible proof paths, before the final aggregation selects the highest scoring one. The whole process is outlined in Algorithm 1.

Example 2.2 (Neural Theorem Provers). Consider a variant of Example 2.1, where each predicate and constant lives in a continuous embedding space, i.e., \( \theta_p, \theta_x, \theta_R, \theta_B, \theta_M \in \mathbb{R}^k \). Rocktäschel & Riedel (2017) propose replacing comparisons between symbols with a differentiable similarity measure \( K : \mathbb{R}^k \times \mathbb{R}^k \rightarrow [0, 1] \), such as a Gaussian kernel, between their embeddings. Their model enumerates all possible proofs for a goal \( G \), and generates a proof score for each of them, given by the minimum of all embedding similarities. For instance, if \( G = \text{grandPa}(\text{RICK, MORTY}) = \{\text{grandPa}, \text{RICK, MORTY}\} \), one candidate proof consists in using the facts \( F = p(\text{RICK, BETH}) \) and \( F' = p(\text{BETH, MORTY}) \) and the rule \( g(X, Y) \Rightarrow p(X, Z), p(Z, Y) \) from the \( \text{KB} \), yielding the score \( K(\theta_{\text{grandPa}}, \theta_g) \). It is important to mention that NTPs allow unifying symbols like \( \text{grandPa} \) and \( g \) (which, in this case, share the same semantics), even though they are lexically different. The score for \( G \) is given by the maximum of all proof scores.

3. Conditional Proving Strategies

The NTPs proposed by Rocktäschel & Riedel (2017) use a fixed set of rules, either specified by the user or learned from data via provided rule templates. In this model, given a goal, there is no hard decision mechanism for deciding which rules can be used for reformulating a given goal into subgoals: all rules in the \( \text{KB} \) need to be considered when proving each goal. For this reason, NTPs were shown not to scale to large datasets (Rocktäschel & Riedel 2017).

Differentiable Goal Reformulation In order to explicitly learn which rule to consider at each step, we propose the following solution. Rather than relying on a fixed, potentially very large set of rules, we propose to dynamically generate a minimal set of rules via a neural network architecture conditioned on the goal to prove.

Example 3.1 (Conditional Theorem Proving). Assume that
Algorithm 1 Overview of the neural backward chaining algorithm proposed by Rocktäschel & Riedel (2017) – intuitively, it recursively proves each goal with all rules in the KB (OR module) and, for each rule, it proves its premise (AND module), up to $d$ recursion steps.

1: function or($G$, $d$, $S$)  
2: for $H := B \in K$ do  
3: for $S \in$ and($B$, $d$, unify($H$, $G$, $S$)) do  
4: yield $S$  
5: function and($B$, $d$, $S$)  
6: if $B = []$ or $d = 0$ then yield $S$ else  
7: for $S' \in$ or($\text{sub}(B_0, \rho, d - 1, S)$) do  
8: for $S'' \in$ and($B_1, d, S'$) do  
9: yield $S''$

where $\Psi$ is a set of variables, and \[ A \triangleq \mathbb{R}^k \times (\mathbb{R}^k \cup \Psi) \times (\mathbb{R}^k \cup \Psi) \] denotes the embedding representation of a goal, such as $g(\text{RICK}, \text{MORTY})$.

For instance, the select module in Eq. (1) can be implemented by a neural network that, given a goal such as $G = [\theta_{G_1}, \theta_{MORTY}]$, generates $H := B$ with $H = [\theta_{Y}; X, Y]$ and $B = [[\theta_{p}; X, Z], [\theta_{p}; Z, Y]]$, corresponding to the symbolic rule $g(X, Y) := p(X, Z) \cup p(Z, Y)$. If the positions of the variables in the rule are fixed, the whole module is end-to-end differentiable with respect to its parameters $\theta$.

Neural Goal Reformulation Here, we define select as a linear function of the goal predicate:

\[ \text{select}_\theta(G) \triangleq [F_H(G) := F_{B_1}(G), F_{B_2}(G)], \tag{2} \]

where the head and body of the resulting rule are given by $F_H(G) = [f_H(\theta_{G_1}), X, Y]$, $F_{B_1}(G) = [f_{B_1}(\theta_{G_1}), X, Z]$, and $F_{B_2}(G) = [f_{B_2}(\theta_{G_1}), Z, Y]$. Every $f_i : \mathbb{R}^k \rightarrow \mathbb{R}^k$ is a differentiable function, such as the linear projection $f_i(x) = W_i x + b$, with $W_i \in \mathbb{R}^{k \times k}$ and $b \in \mathbb{R}^k$. Thus, instead of iterating through a possibly very large set of rules in the KB, we can generate a significantly smaller set of rules, whose generation is conditioned on the goal $G$ and can be trained end-to-end on downstream reasoning tasks.

Attentive Goal Reformulation We can incorporate a useful prior in the select module architecture—namely that predicate symbols in the rule already exist in the KB among the available relations $R$. A method for incorporating this prior consists in using the given goal $G$ for generating a distribution over the set of relations $R$:

\[ f_i(x) = \alpha \mathbf{E}_R, \quad \alpha = \text{softmax} (\mathbf{W}_i x) \in \Delta^{|R|-1}, \tag{3} \]

where $\mathbf{E}_R \in \mathbb{R}^{|R| \times k}$ denotes the predicate embedding matrix, $\mathbf{W}_i \in \mathbb{R}^{k \times |R|}$, and $\alpha$ is an attention distribution $\alpha \in \Delta^{|R|-1}$ over the predicates in $R$, where $\Delta^n$ denotes
the standard $n$-simplex $\Delta^n$. This is especially helpful if the embedding size is larger than the number of relationships, i.e. $k \gg |R|$.

Memory-Based Goal Reformulation A problem with using black-box neural networks for reformulating goals into sub-goals is that it can be difficult to inspect the rules by analysing the model parameters, as in Rocktäschel & Riedel (2017). For this reason, we propose an alternative goal reformulation module, where rules are stored in a differentiable memory.

More precisely, $n$ rules are stored as memory matrices $[M_1, \ldots, M_m]$, where each $M_i \in \mathbb{R}^{n \times k}$ denotes the $k$-dimensional embeddings of the $i$-th predicates in the $n$ rules. Then, the goal $G$ is used to compute an attention distribution over rules $\alpha \in \Delta^{n-1}$, where each $\alpha_i$ denotes the attention weight on the $i$-th rule. The attention distribution can be formalised as follows:

$$f_i(x) = \alpha M_i, \quad \alpha = \text{softmax}(Wx) \in \Delta^{n-1}, \quad (4)$$

where each $f_i : \mathbb{R}^k \rightarrow \mathbb{R}^k$ is a differentiable function that, given the goal, produces an attention distribution $\alpha \in \Delta^{n-1}$ over the rules and for indexing a memory $M_i$, analogous to a key-value memory network (Miller et al. 2016).

4. Related Work

Memory-Augmented Networks Memory-augmented neural architectures aim at improving the generalisation and reasoning abilities in neural networks by disentangling representations from computations. By enriching neural networks with a differentiable external memory, these models were able to multi-hop reason over texts (Sukhbaatar et al. 2015), induce algorithmic behaviours (Graves et al. 2014; Joulin & Mikolov 2015; Grefenstette et al. 2015; Kaiser & Sutskever 2016), and rapidly assimilate new data (Santoro et al. 2016).

Neuro-Symbolic Models Differentiable interpreters enable translating declarative or procedural knowledge into neural networks exhibiting strong inductive biases of that knowledge (Bosnjak et al. 2017; Rocktäschel & Riedel 2017; Evans & Grefenstette 2018). Bosnjak et al. (2017) propose $\partial A$, a differentiable abstract machine for the Forth programming language.

Rocktäschel & Riedel (2017) propose a differentiable implementation for the backward chaining algorithm, effectively implementing a differentiable Datalog interpreter. Evans & Grefenstette (2018) propose a differentiable forward-chaining reasoning process, while Donadello et al. (2017) propose a continuous generalisation of the semantics of first-order logic. Yang et al. (2017) and Sadeghian et al. (2019) propose an approach for learning function-free Datalog clauses from KBs by means of a differentiable graph traversal operator, while Das et al. (2018) propose learning policies for navigating a KB via reinforcement learning.

A major problem with these approaches is their computational complexity, which renders them unusable for large-scale learning problems. In order to address this issue, Minervini et al. (2020) propose Greedy NTPs (GNTPs), an extension to NTPs where, for each goal, only the top-$k$ facts and rules are considered during the differentiable reasoning process.

Neural Module Networks Andreas et al. (2016) introduce Neural Module Networks (NMNs), an end-to-end differentiable composition of jointly trained neural modules. Analogously, NTPs can be seen as a recursive differentiable composition of or and and modules, jointly trained on downstream reasoning tasks. NMNs allow defining and
training end-to-end differentiable composable models, and interpret and execute their compositions as simple programs. This is especially useful when dealing with reasoning tasks from visual and natural language inputs, such as reasoning over text with arithmetic modules (Gupta et al., 2019) or visual question answering (Andreas et al., 2016). Interestingly, in this work the structure of the composition is statically drawn from the data, while Jiang & Bansal (2019) propose a way of learning the model composition via a coordination model (Jiang & Bansal, 2019).

Incorporating Knowledge via Regularisation Another branch of works uses symbolic background knowledge to learn better representations for entities and relationships in a KB. An early work in this space is Rocktäschel et al. (2015), which regularise a relation extraction model by penalising inconsistency with respect to a set of logical constraints and sampled entities. Minervini et al. (2017a) regularise relation representations to incorporate equivalence and inversion axioms for a set of neural link prediction models, while Demeester et al. (2016) focus on simple implication axioms. Minervini et al. (2017b) propose adversarially regularising neural models by identifying, during training, inputs that violate a given set of constraints, and regularising the model to decrease the degree of such violations. Xu et al. (2018) propose a similar idea, by using a semantic loss measuring to which extent a model matches a set of given constraints.

5. Experiments

We evaluate CTPs on two datasets: systematic generalisation on the CLUTRR dataset, and link prediction in Knowledge Graphs (KGs). Datasets are introduced in Section 5.1 while baselines are described in Section 5.2.

5.1. Datasets and Tasks

Systematic Generalisation CLUTRR—Compositional Language Understanding and Text-based Relational Reasoning (Sinha et al., 2019)—contains a large set of graphs modelling hypothetical family relationships. Given a set of family relations, encoded as a graph with a variable number of nodes and edges, the goal is to infer the relationship between two family members, whose relationship is not explicitly mentioned. To solve this task, a learning agent should be able to induce the logical rules governing the kinship relationships, such as the parent of a parent is a grandparent, and use these rules to infer the relationship between a given pair of entities.

CLUTRR allows testing a learning agent’s ability for systematic generalisation, by testing on graphs containing combinations of logical rules that were not seen during training. Each edge in the graph is labelled with one out of nine family relation type from \( \mathcal{R} = \{ \text{child, grand, in-law, inv-child, inv-grand, inv-in-law, inv-un, sibling, un} \} \), and the task consists in inferring the relationship between two of the nodes in the graph. During training, a model is trained to infer such relationship by traversing a limited number of edges (such as two, three, and four edges), and during evaluation the model has to traverse up to ten edges.

Link Prediction Furthermore, we evaluate CTPs on neural link prediction tasks, following the same evaluation protocols as Rocktäschel & Riedel (2017) on the Countries (Bouchard et al., 2015), Nations, UMLS, and Kinship (Kemp et al., 2006) datasets.

The Countries dataset contains countries, regions, and sub-regions as entities, and it is carefully designed to test the logical reasoning and learning capabilities of neural link prediction models: queries have the form \( \text{located in}(\cdot, \cdot) \), where the answers are regions. This dataset comes with three tasks (S1, S2, and S3) each requiring reasoning skills of increasing complexity; we refer to Rocktäschel & Riedel (2017) for more details about this dataset.

The Unified Medical Language System (UMLS) dataset is from bio-medicine: entities are biomedical concepts, and relations include treatments and diagnoses. The Kinship dataset contains kinship relationships among members of the Alyawar tribe from Central Australia.

5.2. Models and Baselines

Models We consider the following three CTP model variants: i) CTP_\(L\), where the mappings \( f_i \) from goal predicates to rule predicates are implemented via a linear projection, ii) CTP_\(A\), where it is implemented via attentive goal reformulation, as described in Section 3, and iii) CTP_\(M\), where it is implemented via memory-based goal reformulation, also described in Section 3.

Baselines We consider two classes of baselines: graph-based and sequence-based. Graph-based baselines consist in neural architectures specifically designed for graph data and. We consider GNTPs, a neuro-symbolic reasoning model, and two Graph Neural Network (GNN) architectures, namely GATs (Velickovic et al., 2018) and Graph Convolutional Networks (GCNs) (Kipf & Welling, 2017).

Sequence-based baselines are neural architectures originally proposed for encoding sequences: by linearising the relational graphs into sequences of subject-predicate-object triples, we can use such models for encoding graphs. We
Figure 3: Results on the CLUTRR dataset after training on stories of lengths $\{2, 3\}$ and evaluating on stories of length $\{4, 5, \ldots, 10\}$ – hyperparameters were fine-tuned on either short stories (left) and long stories (right). Significance testing was assessed via an unequal variances $t$-test in comparison with CTP$_L$: $\triangledown$ (resp. $\triangledown$) represents a $p$-value lower than $10^{-3}$ (resp. $10^{-5}$).

For model selection, we generate a CLUTRR-like dataset using the code published by Sinha et al. (2019) composed of training set graphs with $\{2, 3\}$ edges, and two validation sets, one with graphs with three edges, and another with graphs with nine edges. We then select two sets of hyperparameters for each of the baselines: one that maximises the validation accuracy on graphs with three edges, and another that maximises the test accuracy on graphs with nine edges. All details on the hyperparameter selection process can be found in Appendix A.

To assess the statistical significance of our results, we ran each of the experiments 10 times, each time with a different seed, and compared the resulting accuracy values using an unequal variances $t$-test, or Welch’s $t$-test$^3$.

$^3$We assume accuracy values to be Gaussian-distributed, as they approach a normal distribution for large numbers of re-runs, due to the Central Limit Theorem.
Learning Reasoning Strategies in End-to-End Differentiable Proving

---

Table 1: Comparison of different models on the CLUTRR dataset. The baseline model is CTPL, and the models are compared in terms of accuracy on the CLUTRR dataset. The models are CTP_{L}, CTP_{A}, CTP_{M}, GAT, GCN, RNN, LSTM, GRU, CNNH, CNN, MHA.

<table>
<thead>
<tr>
<th>Model</th>
<th>5 Hops</th>
<th>6 Hops</th>
<th>7 Hops</th>
<th>8 Hops</th>
<th>9 Hops</th>
<th>10 Hops</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTP_{L}</td>
<td>0.98 ± 0.04</td>
<td>0.97 ± 0.03</td>
<td>0.95 ± 0.06</td>
<td>0.93 ± 0.07</td>
<td>0.91 ± 0.05</td>
<td></td>
</tr>
<tr>
<td>CTP_{A}</td>
<td>0.98 ± 0.04</td>
<td>0.97 ± 0.03</td>
<td>0.95 ± 0.06</td>
<td>0.93 ± 0.07</td>
<td>0.91 ± 0.05</td>
<td></td>
</tr>
<tr>
<td>CTP_{M}</td>
<td>0.98 ± 0.04</td>
<td>0.97 ± 0.03</td>
<td>0.95 ± 0.06</td>
<td>0.93 ± 0.07</td>
<td>0.91 ± 0.05</td>
<td></td>
</tr>
<tr>
<td>GAT</td>
<td>0.86 ± 0.04</td>
<td>0.79 ± 0.02</td>
<td>0.73 ± 0.01</td>
<td>0.65 ± 0.01</td>
<td>0.54 ± 0.01</td>
<td></td>
</tr>
<tr>
<td>GCN</td>
<td>0.86 ± 0.04</td>
<td>0.79 ± 0.02</td>
<td>0.73 ± 0.01</td>
<td>0.65 ± 0.01</td>
<td>0.54 ± 0.01</td>
<td></td>
</tr>
<tr>
<td>LSTM</td>
<td>0.86 ± 0.04</td>
<td>0.79 ± 0.02</td>
<td>0.73 ± 0.01</td>
<td>0.65 ± 0.01</td>
<td>0.54 ± 0.01</td>
<td></td>
</tr>
<tr>
<td>GRU</td>
<td>0.86 ± 0.04</td>
<td>0.79 ± 0.02</td>
<td>0.73 ± 0.01</td>
<td>0.65 ± 0.01</td>
<td>0.54 ± 0.01</td>
<td></td>
</tr>
<tr>
<td>CNNH</td>
<td>0.86 ± 0.04</td>
<td>0.79 ± 0.02</td>
<td>0.73 ± 0.01</td>
<td>0.65 ± 0.01</td>
<td>0.54 ± 0.01</td>
<td></td>
</tr>
<tr>
<td>CNN</td>
<td>0.86 ± 0.04</td>
<td>0.79 ± 0.02</td>
<td>0.73 ± 0.01</td>
<td>0.65 ± 0.01</td>
<td>0.54 ± 0.01</td>
<td></td>
</tr>
<tr>
<td>MHA</td>
<td>0.86 ± 0.04</td>
<td>0.79 ± 0.02</td>
<td>0.73 ± 0.01</td>
<td>0.65 ± 0.01</td>
<td>0.54 ± 0.01</td>
<td></td>
</tr>
</tbody>
</table>

---

Figure 4: Results on the CLUTRR dataset after training on stories of lengths \{2, 3, 4\} and evaluating on stories of length \{5, \ldots, 10\} – hyperparameters were fine-tuned on either short stories (left) and long stories (right). Significance testing was assessed via an unequal variances \textit{t}-test in comparison with CTP_{L}. ▽ (resp. ▼) represents a \textit{p}-value lower than 10^{-3} (resp. 10^{-5}).

5.3. Results

CLUTRR We evaluated three CTP variants and all considered baselines on two datasets published by Sinha et al. (2019) under the identifiers 0899076f8 and db9b8f04. We refer to these datasets as CLUTRR_2\{k = 2, 3\} and CLUTRR_2\{k = 2, 3, 4\}, where \(k\) denotes the number of edges in the training graphs. Results for CLUTRR_2\{k = 2, 3\} are summarised in Fig. 3 while results for CLUTRR_2\{k = 2, 3, 4\} are summarised in Fig. 4.

In Fig. 5 we observe that, after training on graphs with two and three edges, baseline models tend to be able to generalise correctly to slightly longer stories (such as graphs with four and five edges), but that predictive accuracy quickly decreases with increasing graph sizes and this phenomenon happens when tuning hyper-parameters either on graphs with three edges or on graph with nine edges.

In our experiments, LSTMs had a strikingly different behaviour in comparison with other baselines: for graphs with nine edges, the accuracy decrease caused by using the LSTMs baseline is only significant with \(p \leq 10^{-2}\) (for all other baselines this change is significant with \(p \leq 10^{-4}\), with a drop in significance for smaller graphs.

The phenomenon that LSTMs yield surprisingly accurate results on the CLUTRR dataset can be seen across every experiment in our empirical evaluation, while other recurrent models such as RNNs and GRUs do not show this.

Model Analysis A great feature of CTPs is that we can analyse the goal reformulation process to understand the reasoning process underlying a given prediction, and extract explainable rules. In Fig. 2 we show a sample of the rules and common-sense reasoning patterns learned by CTPs on the CLUTRR dataset.

We can see that, for example, CTP successfully identifies e.g. \textit{the child of a child is a grandchild, the child of...}
## Learning Reasoning Strategies in End-to-End Differentiable Proving

### Table 1: Comparison of CTPs with GNTPs (Minervini et al., 2020), NeuralLP (Yang et al., 2017) and MINERVA (Das et al., 2018) (from Minervini et al., 2020) on benchmark datasets: hyperparameters were selected based on the validation MRR, and we report the mean and standard deviation over 10 random seeds.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Metrics</th>
<th>CTP</th>
<th>GNTP</th>
<th>NeuralLP</th>
<th>MINERVA</th>
<th>Learned Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CTP_L</td>
<td>CTP_A</td>
<td>CTP_M</td>
<td>GNTP (k = 5)</td>
</tr>
<tr>
<td>S1</td>
<td></td>
<td></td>
<td>100.0 ± 0.00</td>
<td>99.98 ± 0.05</td>
<td>100.0 ± 0.00</td>
<td>100.0 ± 0.00</td>
</tr>
<tr>
<td>Countries</td>
<td>AUC-PR</td>
<td>91.81 ± 1.07</td>
<td>90.82 ± 0.88</td>
<td>93.48 ± 3.29</td>
<td>75.1 ± 0.3</td>
<td>92.36 ± 2.41</td>
</tr>
<tr>
<td>S3</td>
<td></td>
<td></td>
<td>94.78 ± 0.00</td>
<td>87.70 ± 4.79</td>
<td>91.27 ± 4.02</td>
<td>92.20 ± 0.2</td>
</tr>
</tbody>
</table>

### Training Dynamics

![CLUTRR_G(k=2,3,4) -- Training Iteration × Accuracy](image)

Figure 5: Training dynamics of CTPs and GNTPs.

One's significant other is also one's child, and the parent of a significant other is an in-law.

### Training Dynamics

We analyse the training dynamics of CTP_L, CTP_A, CTP_M, and GNTPs (Minervini et al., 2020) on CLUTRR_G(k=2,3,4).

The CTP variants consisted of 5 goal reformulators, each implemented by an independent select module, while GNTP has a KB of 32 rules and $k \in \{5, 7, 9, 11\}$. For all models, the embedding size for entities and relations was set to 50.

Figure 5 demonstrates how the training accuracy of such models evolves during training. We can see that, while the three CTP variants get to a nearly-perfect training set accuracy in less than 300 iterations, GNTPs is unable to match this result, even after careful hyperparameter tuning.

A possible explanation is that, in order to scale to large rule sets, GNTPs only considers the top-$k$ rules, based on the similarity between the goal and the head of the rules. This is equivalent to a hard attention mask, which is known to be problematic to train via gradient-based optimisation (Luong et al., 2015).

### Link Prediction

In Table 1, we show link prediction results in comparison with three other neuro-symbolic reasoning methods, namely GNTPs (Minervini et al., 2020), NeuralLP (Yang et al., 2017) and MINERVA (Das et al., 2018). GNTPs are an extension of NTPs where rules are heuristically selected by search for the rules where the head predicate is closest to the sub-goal predicate in embedding space.

Our experiments show that CTPs produce significantly more accurate or very competitive link prediction results, while controlling the complexity of the reasoning process via the goal-conditioned rule selection. For instance, in the Nations dataset, only two rules were generated by CTPs for each goal, while in Rocktäschel & Riedel (2017) NTPs were required to iterate over sixty rules.

Furthermore, in this case, CTPs were able to produce explanations for each of their predictions. For instance, in the Nations dataset, CTPs successfully extracted logical patterns such as commonbloc1(X,Y) :- relngo(Y,X), timesincewar(X,Y) :- independence(X,Y), unweightedunvote(X,Y) :- relngo(X,Y), and ngo(X,Y) :- independence(Y,X).
6. Conclusions

We introduced CTPs, an extension to NTPs, for learning the optimal rule selection strategy via gradient-based optimisation. For each sub-goal, a select module produces a smaller set of rules, which is then used during the proving mechanism. Furthermore, we proposed three variants of the rule selection mechanism, where the sub-goal reformulations are obtained by linear projections of the sub-goal predicate, attention distributions over predicate embeddings, and a key-value memory lookup over a set of rules.

We showed that CTPs are scalable and yield state-of-the-art results on the CLUTRR dataset, which explicitly tests the systematic generalisation of neural models, in comparison with a wide set of neural baselines. Finally, we demonstrated that CTPs yield competitive results in standard link prediction benchmark in comparison with other neuro-symbolic approaches.

Future Work An open problem is how CTPs can be able to process CLUTRR instances where family relationships are not directly provided as a labelled graph, but rather as free-form text. A possible solution, proposed by Minervini et al. (2020), consists in having an end-to-end differentiable encoder for producing the fact embeddings conditioned on the text, and we are currently analysing several options in this space.

Acknowledgements This work was supported by the EU Horizon 2020 Research and Innovation Programme under the grant 875160. We thank Yihong Chen, Joe Stacey, and all the amazing folks in the UCL NLP group for the enlightening discussions and support. Finally, we thank NVIDIA for GPU donations.

References


A. Hyperparameter Selection

For model selection in baseline models, we generated a CLUTRR-like dataset using the code published by Sinha et al. (2019), composed by a training set of graphs with \{2, 3\} edges, and two test sets, one with graphs with three edges, and another with graphs with nine edges. We then selected two sets of hyperparameters for each of the baselines: one that maximises the validation accuracy on graphs with three edges, and another that maximises the test accuracy on graphs with nine edges. For each of the baselines, we considered a wide range of hyperparameters: the dimensionalities of node and edge embeddings were varied in \{10, 50, 100, 200, 500\}, the number of attention heads in attention-based architectures in \{1, 2, \ldots, 10\}, the number of filters in convolutional architectures in \{1, 2, \ldots, 10\}, and the number of hidden units in recurrent architectures in \{32, 64, 128, 256, 512\}.

To assess the statistical significance of our results, we ran each of the experiments 10 times, each time with a different seed, and compared the resulting accuracy values using an unequal variances t-test, or Welch’s t-test. This is motivated by the observation that accuracy values to be Gaussian-distributed, as they approach a normal distribution for large numbers of re-runs, due to the Central Limit Theorem.

A.1. Optimal Hyperparameters

Note that all recurrent architectures are bidirectional.

Graph Attention Networks: the hyperparameters that maximise the accuracy on validation graphs with 3 edges are \(h = 10\) for the number of attention heads, \(k = 50\) for the dimensionality of node embeddings, \(k_e = 200\) for the dimensionality of edge embeddings. For validation graphs with 9 edges, \(k = 50, k_e = 500, h = 10\).

Graph Convolutional Networks: the hyperparameters that maximise the accuracy on validation graphs with 3 edges are \(k = 50\) for the dimensionality of node embeddings, \(k_e = 500\) for the dimensionality of edge embeddings. For validation graphs with 9 edges, \(k = 50, k_e = 50\).

Convolutional Neural Networks: the hyperparameters that maximise the accuracy on validation graphs with 3 edges are \(k = 50\) for the dimensionality of node and edge embeddings, \(f = 8\) convolutional filters, and conditional encoding. For validation graphs with 9 edges, \(k = 200, f = 4\), and conditional encoding.

Recurrent Neural Networks: the hyperparameters that maximise the accuracy on validation graphs with 3 edges are \(k = 50\) for the dimensionality of node and edge embeddings, \(h = 64\) for the size of the hidden state representations, and conditional encoding. For validation graphs with 9 edges, \(k = 500, h = 512\), and conditional encoding.

Long Short-Memory Networks: the hyperparameters that maximise the accuracy on validation graphs with 3 edges are \(k = 50\) for the dimensionality of node and edge embeddings, \(h = 64\) for the size of the hidden state representations, and conditional encoding. For validation graphs with 9 edges, \(k = 100, h = 512\), and independent encoding.

Gated Recurrent Units: the hyperparameters that maximise the accuracy on validation graphs with 3 edges are \(k = 50\) for the dimensionality of node and edge embeddings, \(h = 64\) for the size of the hidden state representations, and conditional encoding. For validation graphs with 9 edges, \(k = 200, h = 512\), and conditional encoding.

CNN with Highway Encoder: the hyperparameters that maximise the accuracy on validation graphs with 3 edges are \(k = 200\) for the dimensionality of node and edge embeddings, \(h = 2\) highway layers, and conditional encoding. For validation graphs with 9 edges, \(k = 200, h = 1\), and conditional encoding.

Multi-Head Attention: the hyperparameters that maximise the accuracy on validation graphs with 3 edges are \(k = 500\) for the dimensionality of node and edge embeddings, \(h_{\text{top LSTM}} = 100\), \(h = 512\) for the size of the hidden state representation of the top LSTM layer, and conditional encoding. For validation graphs with 9 edges, \(k = 500, h = 10, h_{\text{top LSTM}} = 128\), and conditional encoding.