A Computational Perspective on Neural-Symbolic Integration

Gustav Šír
Czech Technical University in Prague, Czech Republic
E-mail: gustav.sir@cvut.cz

Abstract. Neural-Symbolic Integration (NSI) aims to marry the principles of symbolic AI techniques, such as logical reasoning, with the learning capabilities of neural networks. In recent years, many systems have been proposed to address this integration in a seemingly efficient manner. However, from the computational perspective, this is in principle impossible to do. Specifically, some of the core symbolic problems are provably hard, hence a general NSI system necessarily needs to adopt this computational complexity, too. Many NSI methods try to circumvent this downside by inconspicuously dropping parts of the symbolic capabilities while mapping the problems into static tensor representations in exchange for efficient deep learning acceleration.

In this paper, we argue that the aim for a general NSI system, properly covering both the neural and symbolic paradigms, has important computational implications on the learning representations, the structure of the resulting computation graphs, and the underlying hardware and software stacks. Particularly, we explain how the currently prominent, tensor-based deep learning with static computation graphs is conceptually insufficient as a foundation for such general NSI, which we discuss in a wider context of established (statistical) relational and structured deep learning methods. Finally, we delve into the underlying hardware acceleration aspects and outline some promising computational directions toward fully expressive and efficient NSI.

Keywords: Neural-Symbolic Integration, Computation Graphs, Learning Representations, Relational Learning, Hardware Acceleration, Logical Reasoning, Dynamic Deep Learning, Computational Complexity

1. Introduction

During the past decade, deep learning (DL) has taken over the wider domain of artificial intelligence (AI) so rapidly that many people now commonly confuse the fields as equivalent, expecting neural networks to eventually solve all AI problems at hand. This has only been accentuated with the recent rise of foundation (neural) models [1], which exhibit many AI capabilities that have previously been though as out of reach for DL. Indeed, the dramatic progress in difficult, symbolic reasoning benchmarks has left a lot of people wondering whether there are even any limits to the current strategy of scaling large Transformers [2].

However, from the computational perspective, one thing remains striking - these neural models arrive at the answer to every possible input after a *fixed number* of the same computation steps, which is clearly suboptimal. For simple inputs, this is a waste of computation, and for complex inputs, there will always be a limit where this is insufficient. Particularly, many (symbolic) AI capabilities, such as reasoning or planning, are provably NP-hard (or worse) and thus require a number of computation steps exponential in the size of the input. Hence, this static DL strategy is clearly not universal. However, such a fixed form of computation is the core characteristic of DL that perpetuates the whole domain, as it is crucially required to make use of the underlying hardware-specific, data-parallel acceleration, particularly with the Graphics Processing Units (GPUs).

Researchers often like to think of their models and algorithms independently of the underlying hardware (HW) and software (SW). However, the resurrection of modern DL has been a very clear demonstration that these worlds

are deeply intertwined. And while the recent computing evolution in the direction of specialized HW accelerators has been transformational for the success of DL, it unfortunately started to further deepen the gap between DL and the, more general, symbolic AI compute, making the exploration of novel neural-symbolic methods aimed at their integration increasingly more difficult. On the one hand, we have the symbolic, logic-based approaches to AI working in the regime of (hard) discrete optimization, with long historical roots reaching all the way back to the origins of AI and computing itself. These make heavy use of stateful computation with conditional branching, leading to very sparse and irregularly structured computation graphs that change *dynamically* from input to input. This form of compute builds on the universal capabilities of the Central Processing Unit (CPU) architecture, which dominated computing throughout history. On the other hand, we have the more recent, static, tensor-based neural approaches, working in the simpler regime of continuous optimization, now depending crucially on the GPU architecture that, however, lacks the universal support required for the (dynamic) symbolic compute. This underlying dichotomy is then heavily reflected on the SW layer, too, with largely incompatible learning representations and models.

In this paper, we argue that the current trend of shifting to the static tensor-based (neural) methods is principally limiting the expressiveness of the resulting NSI systems. To clarify the position, we overview evolution of the NSI approaches to AI, with a particular focus on *relational* logic representations that lie at the core of the dichotomy between the (static) neural and (dynamic) symbolic computation. We then discuss the limitations through insights into so-called "lifted models" [3] from the associated field of (statistical) relational learning [4]. Finally, we link these to the computing structures of the, increasingly popular, (deep) graph representation learning models [5], and conclude the paper with a vision of how the field of NSI might benefit from the associated HW evolution [6].

2. A Brief History of Neural-Symbolic Computation

Historically, the mainstream vision for AI has been fluctuating from the almost purely symbolic approaches in the early days to the almost purely neural approaches we see today. However, history teaches us that whenever there are two streams of very opposing schools of thought, the resolution is typically somewhere in the middle.

2.1. From Neurons to Symbols

Perhaps surprisingly, the correspondence between the neural and logical calculus has been well-established throughout history. Dating all the way back to the introduction of the first computational *neuron* by McCulloch and Pitts in their 1943's paper "A logical calculus of the ideas immanent in nervous activity" [7], we can see that it was actually thought to emulate *logic* gates over binary propositions. The idea was based on the, now commonly exemplified, fact that logical connectives of conjunction and disjunction can be easily encoded by binary threshold units with weights (Fig. 1) – i.e., the perceptron, a learning algorithm for which was introduced shortly after.

While stacking these on top of each other was a clear path towards representation of more complex, nested logical functions, it took until the 1980's to reintroduce the chain rule for differentiation of nested functions as the "backpropagation" method to calculate gradients in such "*neural networks*" which, in turn, could be efficiently trained with gradient descent. For that, however, researchers had to replace the original binary threshold units with *differentiable* activation functions, such as the logistic sigmoid, which started digging a gap between the neural networks and their crisp logical interpretations.

2.2. From Deep Learning to Logic

These historical parallels might seem outlandish in the context of modern DL. However, interestingly, even the modern idea of DL was not originally described as bound to the neural networks, but rather universally to methods modeling hierarchical composition of useful concepts that are reused in different inference paths of target variables from the input samples [8]. And while this concept has most commonly been instantiated through "hidden layers" in DL, such hierarchical abstractions are common in logical reasoning, too. In fact, *logic* is the very science of deduction of useful concepts from simpler premises in a hierarchical (nested) fashion. While constructing a proof in logic, auxiliary lemmas are often created to reduce the complexity of the theory in scope, in a fashion very similar

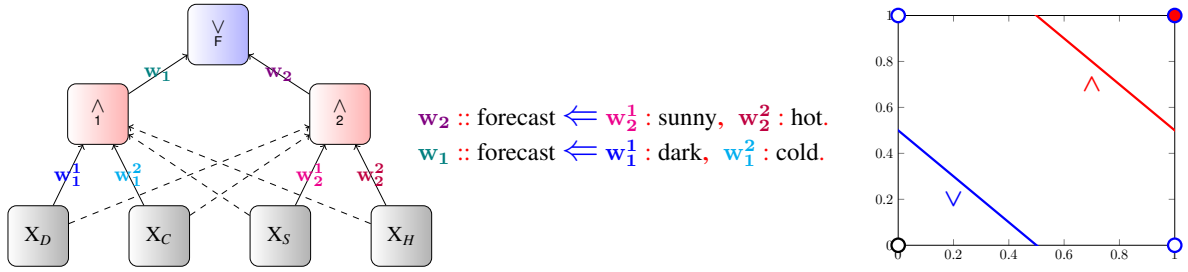

Fig. 1. An example correspondence between a neural network (left) and a propositional logic program (middle), based on conjunctive and disjunctive activations emulated by the respectively weighted neurons (right), in the spirit of KBANN [9].

to DL. Thus, instead of hidden neural layers, these hierarchical abstractions may also be thought of as "complicated propositional formulae re-using many sub-formulae".[1]

And from the NSI perspective, this is not just some metaphor, but actual systems reflecting this paradigm started to emerge throughout the '90s, such as the popular Knowledge-Based Artificial Neural Network (KBANN) [9] (Fig. 1). However, as imagined by Bengio, such a direct neural-symbolic correspondence was insurmountably limited to the aforementioned *propositional* setting, popularly referred to as "propositional fixation" of neural networks [10].

2.3. Problems with Relational Logic

While the aforementioned computational correspondence between the propositional logic programs and neural networks was very direct, transferring the same principle to the *relational*[2] setting has remained a major challenge [11]. The issue is that in the propositional setting, only the (binary) values of the existing input propositions are changing, while the structure of the logical program stays fixed. This is easy to think of as a boolean circuit (neural network) sitting on top of a propositional interpretation (feature vector).

However, the relational input interpretations can no longer be thought of as individual values over a fixed (bounded) number of propositions (interpretation vector), but an *unbound set* of related atoms that are true in a given world (a Herbrand model [12]). Consequently, also the structure of the logical inference unfolded upon this representation can no longer be represented by a fixed boolean circuit. Naturally, when proving a query in relational (first-order) logic, we do not know the number of objects that will form input into the proof, we do not know the number of auxiliary lemmas that will be created in the process, and we do not know the length of the proof nor the overall size of the computation. Hence, there is just no way of mapping that relational reasoning into a static neural computation with a fixed number of inputs, neurons in each layer, and overall depth, effectively corresponding to a finite state machine, principally incapable of capturing the unboundedness, characteristic to the relational logic.

2.4. Modern NSI Systems

In the field of NSI, the issue with relational expressiveness, not present in the propositional setting, has alternatively been viewed as the problem of *variable binding* [13]. Indeed, it is actually the relational logic variables that effectively represent the unbound sets of objects, enabling to make general statements about different individuals, and facilitating the unbound variability in symbolic AI computing structures. This is in contrast to the bounded processing of fixed-size vector (tensor)[3] representations characteristic to modern DL compute.

In recent years, a large number of new NSI systems building directly on top of popular deep learning frameworks emerged [14]. Many of these systems are designed to operate with some level of first-order expressiveness, seemingly resolving the old issue of propositional fixation [10] with the efficient, modern-day, static tensor-based deep

[1] quotation from the abstract of "Learning Deep Architectures for AI" by Y. Bengio [8]

[2] We use the common term "relational logic" to refer to first-order logic without function symbols, which would unnecessarily complicate the computational perspective taken in this paper by leading further to infinite interpretations.

[3] We do not make much differences between vectors and tensors since, from the representation expressiveness point of view, they are the same.

learning [15]. However, following the aforementioned computational perspective, this is in principle impossible to do. Consequently, in one way or another, these efficient, modern neural-symbolic systems are necessarily dropping parts of the symbolic (relational-level) capabilities, often somewhat inconspicuously.[4] This is because building upon the modern, GPU-accelerated, DL foundation always requires some mapping into the fixed-size tensor computation of the existing frameworks [16, 17], inherently insufficient to capture the unbound structures of relational inference.

3. Relational Machine Learning

Mapping problems to fixed-size numeric vectors (tensors) has a long history in machine learning. This representation is highly appealing since treating each sample as an independent point in an $n-$dimensional space allows to directly adopt classic results from multivariate statistics [18] and the efficient computing machinery of basic linear algebra subprograms [19]. However, real-world problems are not stored in numeric vectors, but in the interlinked structures of the internet pages, knowledge graphs, and databases, which are inherently *relational* representations.

3.1. Propositionalization

This representation dichotomy leads to a natural urge to turn these structures into the familiar form of the vectors, and continue with our standard ML pipelines [18], following the common cognitive bias of "If all you have is a hammer, everything looks like a nail". In relational ML, this approach is generally referred to as "propositionalization" [20] which, similarly to the "tensorization" [21, 22] in modern NSI systems (Sec. 2.4), is an effective strategy that often works well in practice. However, it is conceptually limited.

Firstly, it is clear that one cannot map from unbound relational structures into the fixed-size vectors or tensors without (unwanted) loss of information. This obvious limitation is commonly addressed by bounding the number of domain elements, seemingly restricting the maximum size of the resulting (neural) computation [23]. However, even if we restrict ourselves to such bounded structures, designing a suitable learning representation in the form of a numeric vector or tensor is still deeply problematic. For example, let us take even just graph structures – a restricted form of relational data. If there was a definite way to map a graph into the standard learning form of a (fixed-size) numeric vector (or tensor), it would trivially solve the graph isomorphism problem, since to test if two graphs are isomorphic or not, it would suffice to turn them into such vectors and compare these for equality instead. Of course, one can simply decide to ignore this problem and choose one of the plethora of the (NSI) approximations [24]. However, any such vector (tensor) mapping that ignores the isomorphism symmetry will naturally break the underlying logical semantics that is inherently invariant to it. Therefore, to address this problem properly, it is necessary to leave the space of fixed-size vector (tensor) transformations and revisit the original space of symbolic logic.

3.2. Inductive Logic Programming

Much out of the lights of the ML mainstream, there has been a community of Inductive Logic Programming (ILP) [25], concerned with *proper* ways of learning (interpretable) models from exactly such relational representations. For illustration, let us see how the graph-structured learning problems can be approached with the relational logic representations of ILP. To represent a graph, we simply define a binary "*edge*" relation, with a set of instantiations $edge(x, y)$ for all the adjacent nodes x, y. Additionally, we may also use other (unary) relations to assign attributes to the sets of nodes, such as $red(x)$, etc.

The models then take the form of *relational* logic rules, enabling to capture an unbound variety of structures, from characteristic motifs in molecules to paths in a network (Fig. 2). For instance, a (learned) relational model *correctly* capturing paths in a graph, can be easily learned with a basic ILP system from but a few examples. The simplicity of this problem within the relational representation formalism of ILP is then in direct contrast to the neural models operating on the propositional foundation of fixed-size vectors (tensors). There, in order to properly embed the same problem into the hypothesis space of a classic, fixed neural architecture, we would generally need to restrict

[4] or opting for merely compositional (unintegrated) approaches, such as those explained in Sec. 3.3 and Sec. 5.2

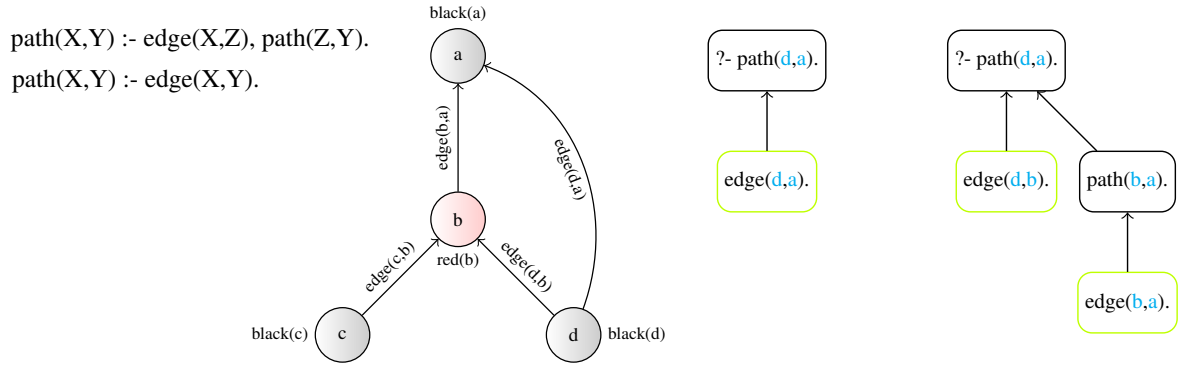

Fig. 2. An example learned ILP model of a path (left), being unfolded on a labeled graph sample encoded in relational logic (middle), resulting into *two different* computation graphs for the same target query "path(d, a)" (right).

ourselves to fixed-size graphs, design an exponentially large network to capture all the possible (inference) paths, in which it would be close to impossible for gradient descent to find the correct solution, and we would still not be able to properly generalize to different (larger) graphs. Moreover, the learning difficulties remain even if we move to theoretically expressive (Turing-complete) neural architectures, such as the (DeepMind's) "Differentiable Neural Computer" [26], which required an extreme number of examples and additional hacking (e.g., pruning out invalid path predictions) to approximate the very same path-finding task through the (propositional) tensor representation, emulating a "differentiable memory" [27].

While ILP is able to correctly capture the expressiveness of relational logic, it is not well suited for dealing with noise and uncertainty, commonly present in real-world learning tasks. To address that issue, a number of methods arose to merge ILP with statistical learning under the notion of Statistical Relational Learning (SRL) [28], such as the "graphicalization"[5] approach of [29], transforming the relational domains into graphs on top of which kernel methods were utilized. Generally, there have been two major streams of approaches in SRL — (i) probabilistic logic programs [30] and (ii) lifted graphical models [3]. While the former builds heavily on the classic, symbolic ILP formalism, the latter is much closer to standard statistical models, the expressiveness of which it "lifts" from the propositional into the relational setting, while retaining proper logical semantics. Naturally, this is highly relevant to the aforementioned problem of propositional fixation of neural networks and relational-level NSI (Sec. 2.3).

3.3. Lifted Graphical Models

As opposed to standard ("ground") machine learning models, such as neural networks, *lifted* models do not specify a particular computation structure, but rather a logical "template" from which the standard models are being unfolded as part of the inference (evaluation) process, given the varying context of the relational input data. For instance, the (arguably) most popular lifted model — a Markov Logic Network (MLN) [31] may be seen as such a template for the classic Markov networks. For example, such an MLN template may express a general prior that "friends of smokers tend to be smokers" and "smoking may lead to cancer". The learning data may then describe a social network of people with a subset of smokers labeled as such. The lifted modeling paradigm of MLNs then allows to induce the smoking probabilities of all the other people based on their social relationships, as if modeled by a regular Markov network, yet correctly generalizing over social networks of *diverse structures and sizes*, overcoming the propositional fixation of the base Markov model (Fig. 3).

The crucial ability to capture the unbound variability of the relational computation structures with a bounded computation template necessarily induces some form of repeated regularities, i.e. *symmetries*, within any such unfolded computation. In relational logic, these symmetries stem from the use of logical variables within the relations, i.e. subsets of the Cartesian products defined over the respective sets of objects. That is, unless the computation is

[5] which can be seen as a lossless alternative to the discussed propositionalization.

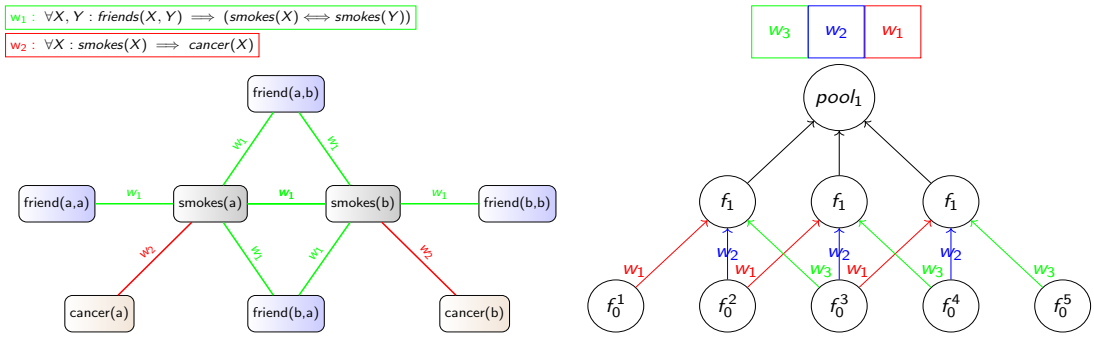

Fig. 3. Illustration of two forms of relational expressiveness, with CNNs (right) capturing a spatial pattern of a 3-neighborhood, and an MLN (left) capturing a social pattern of smoking habits amongst friends, both reflected by the symmetries in the respective ground model computations.

effectively propositional as a corner case, one should be able to observe these repeated computational substructures, including shared parameterization, in the unfolded computation of the (ground) model. And while the, relational logic-based, lifted graphical models make this observation very explicit, we can actually observe the same principle being effectively exploited in many successful DL architectures, too, such as the popular CNNs, exploiting the symmetry of translational equivariance.[6] From this computational view, the CNNs can thus be seen as a particular instance of an (implicit) relational modeling template (Fig. 3).

The proper logical semantics of the lifted graphical models is, however, redeemed by computational difficulties. Due to the explicit treatment of the logic variable binding, these models are commonly viewed as computationally expensive. Nevertheless, from the perspective of a general NSI system, this view is somewhat misleading since the computation is exactly as complex as necessary to properly capture the respective relational (logic) expressiveness.[7] However, the problematic factor is the *structure* of the resulting computation.

4. Structured Computation Graphs

For the sake of delving deeper into the (relational) computing structures in NSI, let us now define, a bit more formally, the common DL abstraction of a *computation graph*. A computation graph $\mathrm{G} = (\mathcal{N}, \mathcal{E}, \mathcal{F}, \mathcal{W})$, composed of nodes $\mathcal{N}$, edges $\mathcal{E}$, activation functions $\mathcal{F}$, and parameters $\mathcal{W}$, is a general way to represent nested parameterized functions using the language of graph theory. The neural networks are then commonly conveyed by the means of directed, differentiable, data-flow graphs where the data, commonly represented as the numeric vectors (or tensors) $\boldsymbol{x}^k$, "flow" through the directed edges $e \in \mathcal{E}$ while being successively transformed by the $\mathcal{W}$-parameterized, differentiable activation functions $f \in \mathcal{F}$ associated with the nodes $N \in \mathcal{N}$, commonly referred to as "neurons".

4.1. Basic Deep Learning

By far the most common computation paradigm in deep learning is to structure the computation graphs regularly into homogeneous "*layers*", which are used to refer to the sets of neurons $\{N \mid \mathsf{depth}_{\mathrm{G}}(\mathsf{N}) = \mathsf{k}\}$ residing at the same depth k in G. An input layer $k = 0$ is then commonly used to represent the feature values $\boldsymbol{x}$ of the input data samples $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ themselves. An output layer $k = \mathsf{depth}(\mathrm{G})$ then corresponds to the target values $\boldsymbol{y}$. The arguably most popular computation structure is then a "fully-connected" design pattern, where the interconnections between nodes in subsequent layers k and $k+1$ follow $N^k, N^{k+1} : (N^k, N^{k+1}) \in \mathcal{E}$, forming a complete bipartite graph. Moreover, each edge here is associated with a unique weight as $\mathcal{E} \xrightarrow{1:1} \mathcal{W}$. Consequently, assuming the common vector form of the input data samples $\boldsymbol{x}_i$, the computation can be efficiently reduced to a linear series of full (dense) matrix W_k^{k+1} multiplications, each followed by an element-wise application of some non-linear function f^{k+1}.

[6] or invariance with the use of the pooling operator

[7] For instance, there are many useful relational templates (patterns) the computational complexity of which is simply *linear*, such as the CNNs and almost all other weight-sharing neural models.

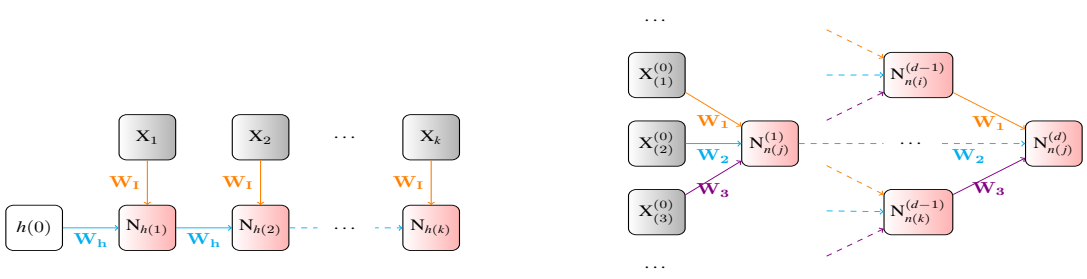

Fig. 4. The symmetric weight-sharing patterns in the structured computation graphs of recurrent (left) and *recursive* (right) neural models.

4.2. Relational Neural Models

The regular, dense, and homogeneous structure of the computation graph G, mapping to the basic matrix (tensor) operations, is a salient feature of the basic neural models. Note that from the aforementioned perspective of the lifted (graphical) models from SRL (Sec. 3.3) such computation, lacking any symmetries, is effectively *propositional*. However, there are also neural models designed to operate with some forms of relational data structures, such as sequences, trees, and graphs, which necessarily need to reflect some level of relational expressiveness. As a consequence, following the SRL reasoning from Sec. 3.3, their computations will, in contrast to the basic neural models, need to reflect some type of sparsity, heterogeneity, or irregularity, within the scope of the necessary symmetries reflected by some form of regular weight-sharing patterns within their computation graphs. Let us now inspect some common representatives of these neural models.

A *Recursive* Neural Network (RecNN) [32, 33] is a neural architecture the computation graphs G_i of which directly follow the structure of each input example x_i, which takes the form of a $k-$regular *tree*. This enables to learn neural networks *directly* from differently-structured regular tree examples x_i, as opposed to using some transformation into the fixed-size tensors $\boldsymbol{x}$ (Sec. 3.1).[8] Here, the (vector) representations of every c leaf nodes $x^j, \ldots, x^{j+c}$ with the same parent node N_j^1 in the respective computation tree G_i are consequently combined by a given $\boldsymbol{c}$-parameterized operation c, such as a $\boldsymbol{c}$-weighted dot-product, to compute the representation of the parent N_j^1. This combining operation c then continues recursively for all the interior nodes, until the representation for the root node $N^{k=d}$ is computed, which forms the output of the model for x_i. Similarly to the convolution in CNNs, this parameterized combining operation c over the children nodes is shared across the tree [34] however, differently from CNNs, the resulting computation graph can be highly and *irregularly sparse* (Fig. 4).

Note that the basic form of the more common *Recurrent* Neural Networks (RNNs) [35] can then be seen as a restriction of the RecNN computation to *sequential structures* x_i (Fig. 4). There, the computation graph G in the form of a linear binary tree is successively unfolded along the input sequence to compute the hidden representation for each node N_i based on the previous node's N_{i-1} representation, and the current input features $\boldsymbol{x_i}$.

The most popular recent addition in this category are then Graph Neural Networks (GNN) [36, 37] which, in contrast to the RNNs, can be seen as an extension of RecNNs to completely *irregular graph data* $x_i = \{\mathcal{N}_i, \mathcal{E}_i\}$. For that purpose, they unfold the computation graph G_i structure from each input structure x_i, similarly to the RecNNs. However, a GNN still follows the standard layered approach (Sec. 4.1), where the structure of *each layer* k exactly follows the structure of the *whole* input graph x_i . For computation of the next layer $k+1$ representations of the nodes in G_i, each node N calculates its own value $h(N)$ by *aggregating* A ("pooling") the (vector) values of the nodes $M : (N, M) \in \mathcal{E}_i$ adjacent in the input graph x_i ("message passing"), transformed by some parametric function C_{W_1} ("convolution"), which is again being reused with the same parameterization W_1^k within each layer k (Fig. 5). However, in contrast to RecNNs, a different parameterization is typically used at each layer.

[8] which can also be seen as graphs with completely regular grid topologies

4.3. Lifted Relational Neural Networks

Lifted Relational Neural Networks [38] can be seen as a further generalization of the GNNs to arbitrary relational structures, covering all the previous models as special cases [39]. For that, they transfer the strategy of *lifted modeling* (Sec. 3.3) into a deep learning setting, effectively extrapolating the original, direct neuro-symbolic correspondence between deep learning and propositional logic (Sec. 2.2) straight into fully expressive *relational* setting.

Similarly to the lifted graphical models (Sec. 3.3), they adopt the language of weighted relational logic as their foundation, endowing them with explicit variable binding and symbolic reasoning capabilities, as found in the ILP systems (Sec. 3.2). This covers the sought-after ability to manipulate *unbound* structures x_i, such as correctly resolving various path-finding (planning) problems (Fig. 2). However, differently from ILP and SRL, they introduce a *differentiable* semantics to the language, endowing them with ability to also directly model various deep learning architectures. Particularly, all the introduced CNN, RNN, RecNN, and GNN examples come directly from trivial differentiable logic programs in the LRNN framework [39]. This is in contrast to the modern (tensorization) NSI systems (Sec. 2.4) that build on top of the existing static DL compute stack, forcing them to map the logical semantics into fixed-size tensors, losing the relational expressiveness in the process (Sec. 3.1). Nevertheless, this also means that LRNNs necessarily share the computational difficulties associated with the relational expressiveness, particularly the *dynamically structured* computation graphs G_i, notoriously incompatible with the GPUs [40].

5. Hardware and Software Lottery

It is instructive to note again the RecNNs architecture, which can be seen as a strict generalization of the common RNNs, adding significant flexibility by exploiting also the sentence *structure* in NLP, while retaining comparable computation complexity.[9] Moreover, in contrast to the RNNs, the RecNNs can, in principle, be parallelized, thanks to the recursive decomposition. Nevertheless, they have never reached anywhere near the popularity of the simpler RNNs, as they were notoriously difficult to effectively implement in the mainstream, fixed-size tensor-based deep learning frameworks.[10] A similar story then followed with the subsequent replacement of the RNNs with Transformers [2], the core "attention" concept of which can be seen simply as a restriction of the GNN computation to fixed-size, fully connected graphs. This hardwired restriction removes the relational flexibility w.r.t. the GNNs and restricts the length of the input sequence w.r.t. the RNNs, while also substantially increasing the computation complexity from linear to quadratic. Nevertheless, it is exactly these apparent drawbacks that provide the architecture one crucial advantage - it is perfectly aligned with the current HW stack, particularly the GPUs.

The GPU architecture is inherently based on the "single instruction, multiple data" (SIMD)[11] concept [42], enabling for high-throughput data-parallel processing of dynamic values through static operations. With neural networks, this generally means static, regular, and dense structure of the computation graph, which maps perfectly to the homogeneous, fixed-size tensor processing underlying the mainstream deep learning architectures and frameworks. On the contrary, the dynamic, sparse, irregular, and heterogeneous nature of the discussed (lifted) neural models, reflecting some level of relational expressiveness, is highly problematic to this mainstream regime of compute. And the situation is further considerably worse with the symbolic (logic) computation (Sec. 3.2), commonly relying on conditional branching and looping (recursion).

5.1. Accelerating Symbolic Computation

The maturity of the existing HW&SW DL ecosystem naturally incentivizes NSI researchers away from the expressive symbolic computation towards further "tensorization" of their systems [21], even in cases where such a choice would otherwise be conceptually inappropriate, which has a self-reinforcing side-effect of regressing to the

[9] depending on the specific parse/dependency tree representation, but generally $O(n \cdot log(n))$ at worst

[10] A notable, yet largely unsuccessful, attempt to improve this situation was, e.g., Tensorflow Fold [41].

[11] A bit more precise term with modern GPUs is SIMT, combining SIMD with multi-threading which, however, shares the same limitation as all the threads effectively execute the same logic.

same models [43]. This is becoming noticeable with more and more NSI methods operating simply as modules on top of existing tensor-based frameworks [44]. However, the symbolic computation, as exploited in the classic (relational) AI tasks, is often inherently sequential, stateful, and conditional, leading to dynamically sized and structured computation graphs, the forms of which are just impossible to properly embed into the common fixed-size tensor processing on GPUs.

For instance, the computation of a relational (lifted) model (Sec. 3.3) may completely change from sample to sample, making it principally incompatible with the SIMD philosophy of the GPU architecture. Importantly, this does not mean that the relational (symbolic) compute strategy would itself be somehow inferior or impossible to accelerate. On the contrary, akin to batching in DL, one can always naively parallelize in the data dimension across multiple CPU threads, utilizing its "multiple instructions, multiple data" (MIMD) capabilities. Consequently, should the evolution of CPU parallelism have reflected its GPU counterpart, i.e. chips with (hundreds of) thousands of CPU threads be readily available, there would be no need to sacrifice the symbolic capabilities in NSI systems. Moreover, the symbolic model representation also offers a much broader range of interesting acceleration strategies, such as lifted inference [45], in contrast to the comparatively brute-force approach of the GPU-based DL.

5.2. Accelerating Graph Neural Networks

An interesting borderline example between these two, largely separate, worlds of symbolic and neural compute are the aforementioned GNN models, which can be seen as an instance of deep learning evolving in the direction of relational learning. Since the GNNs are able to (partially) capture graph-structured data, they necessarily reflect some level of (symbolic) relational logic expressiveness. Particularly, their expressive power corresponds to a guarded fragment of C2 logic [39, 46].[12] Similarly to the other relational models, this increased, albeit still relatively limited (e.g., compared to the LRNNs [39]), expressiveness already takes its toll in their compute structure. Particularly, the variability of the input graph data, reflected in the irregularly sparse interconnection patterns within the GNN layers requiring random access to the underlying tensor indices (through respective gather/scatter operations [47]), is highly detrimental to the classic GPU acceleration.

However, the significant rise in popularity of the GNN-like architectures, driven by the omnipresence of the graph data structures in real-world problems, has come with an interesting paradigm shift in which the SW and HW developers started to actively reflect back on these alternative computing concepts in AI, too. Consequently, there have been new SW frameworks emerging [47, 48], and most of the latest GPU cards support some levels of sparsity [49]. This is an interesting step in a direction beneficial for NSI [50], however, it is still far from a proper foundation for the full NSI compute, which can never fully fit the SIMD design of the GPUs. The corresponding low-level optimizations here are still largely incremental, building heavily on the legacy of the matrix-multiplication-based architectural design. Moreover, they tend to be increasingly specific, e.g., enforcing particular, fixed sparsity patterns [49], or even containing hardwired acceleration for particular models, such as the Transformers [51].

In contrast, the high-level, HW-independent, symbolic acceleration techniques are much more principled and reusable. For instance, in the spirit of NSI, the symbolic techniques of lifted inference (Sec. 3.3) can be transferred into deep learning to effectively accelerate relational neural models, such as the GNNs, by exploiting the inherent symmetries contained in their computation graphs. This results into explicitly smaller neural computation that effectively performs the exact same function (Fig. 5), which was unprecedented in deep learning [52]. Consequently, using this symbolic technique, even a naive CPU implementation of a GNN model can become faster than a specialized GPU-accelerated implementation [52].[13] However, should we completely fixate on the GPU-based deep learning as the foundation for NSI, our possibilities for such a beneficial transfer of symbolic techniques into deep learning (and vice-versa) might gradually become extinct.

6. Conclusion – A Call for Neuro-Symbolic HW

In the past, AI researchers used to design their algorithms independently of the underlying HW, relying on the universal capabilities of the CPU architecture, and its all-encompassing acceleration through the Moore's law [53].

[12] Relational logic with neighbor "guards", counting quantifiers, and at most 2 variables [46].

[13] Particularly in relational domains with highly symmetric graph structures, such as molecular chemistry.

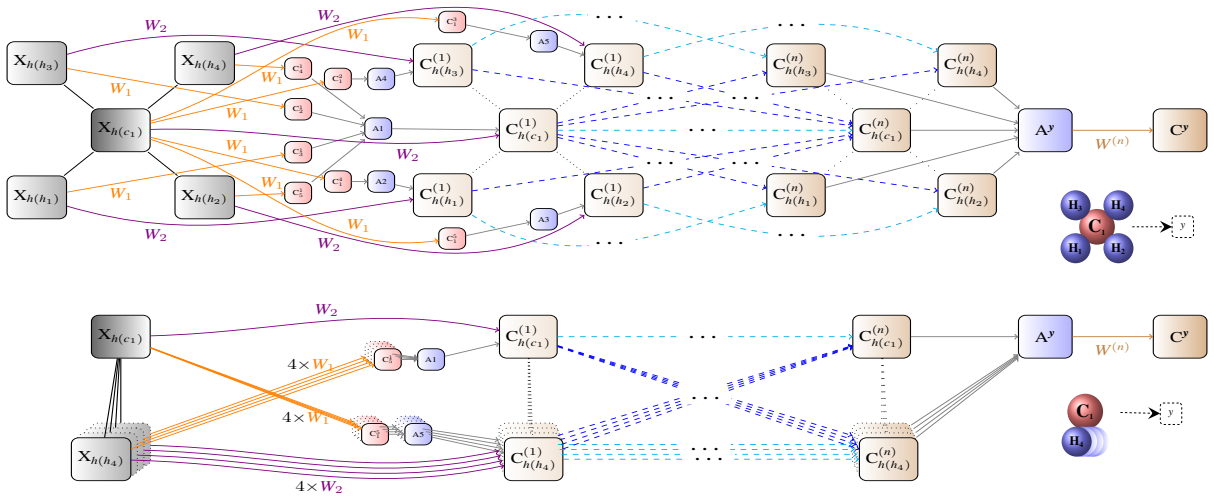

Fig. 5. Compressing a typical GNN model computation (up) based on the contained computation graph symmetries, resulting into a much smaller yet functionally equivalent computation (down), in the spirit of (symbolic) lifted inference [52], as implemented in LRNNs (Sec. 4.3).

However, in this new era of AI compute, an inadequate alignment with the existing specialized HW can easily turn conceptually strong ideas upside down. The prospect of the GPU acceleration, together with the mature DL SW ecosystem built upon that foundation, naturally incentivizes researchers to adjust their ideas to this mainstream regime of compute in order to reap its benefits. However, as argued throughout this paper, this choice has important consequences, especially for NSI.

While the SIMD architecture of the GPUs is a perfect fit for the current, tensor-based neural models, the principles of the symbolic methods are closely connected to the universal flexibility of the CPUs. The vision of a proper, fully expressive *and efficient* NSI computation thus calls for a much tighter integration between the CPU and GPU features than currently present. Instead of incremental improvements of the, principally limited, SIMD approach, for NSI it would be much more meaningful to transfer the MIMD principles from CPUs onto a significantly more parallel, high-throughput, low-latency foundation.

While the AI industry is currently dominated by the GPU-based applications of DL with largely predictable business potential, it is highly unlikely that we will see the same amount of interest in acceleration of the, more demanding, symbolic computing. The NSI community is just too small of a niche to attract the astronomical investments needed for development of such a new HW architecture. However, with the rapid growth in popularity of the GNN models, reflecting some of the computing demands of the symbolic methods, we start to see first signs of the needed HW convergence with newly emerging architectures, such as the Intelligence Processing Unit (IPU) [6], finally breaking beyond the SIMD legacy of the GPUs, and currently leading the GNN performance benchmarks [54]. While still far from meeting all the compute demands of a general NSI system, it might be very interesting for researchers to watch this emerging stream of novel computing architectures, bearing promise to enable exploring truly novel and expressive neural-symbolic systems.

Acknowledgments

The author of this work acknowledges support from the European Union's Horizon Europe Research and Innovation program under the grant agreement TUPLES No 101070149, and funding from the Czech Science Foundation grant. no. 24-11664S.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., Language models are few-shot learners, *Advances in neural information processing systems* **33** (2020), 1877–1901.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, Attention is All you Need, *Neural Information Processing Systems* (2017).

[3] A. Kimmig, L. Mihalkova and L. Getoor, Lifted graphical models: a survey, *Machine Learning* **99**(1) (2015), 1–45.

[4] G. Marra, S. Dumančić, R. Manhaeve and L. De Raedt, From statistical relational to neurosymbolic artificial intelligence: A survey, *Artificial Intelligence* (2024), 104062.

[5] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and S.Y. Philip, A comprehensive survey on graph neural networks, *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[6] Z. Jia, B. Tillman, M. Maggioni and D.P. Scarpazza, Dissecting the graphcore ipu architecture via microbenchmarking, *arXiv preprint arXiv:1912.03413* (2019).

[7] W.S. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity, *The Bulletin of Mathematical Biophysics* **5**(4) (1943), 115–133. ISBN 0007-4985.

[8] Y. Bengio, Learning Deep Architectures for AI, *Foundations and Trends in Machine Learning* **2**(1) (2009), 1–127. ISBN 2200000006.

[9] G.G. Towell, J.W. Shavlik and M.O. Noordewier, Refinement of approximate domain theories by knowledge-based neural networks, in: *Proceedings of the eighth National conference on Artificial intelligence*, Boston, MA, 1990, pp. 861–866.

[10] J. McCarthy, Epistemological challenges for connectionism, *Behavioral and Brain Sciences* **11**(1) (1988), 44–44.

[11] S. Bader and P. Hitzler, Dimensions of Neural-symbolic Integration - A Structured Survey, *arXiv preprint* (2005).

[12] J.H. Gallier, *Logic for computer science: foundations of automatic theorem proving*, Courier Dover Publications, 2015.

[13] K. Greff, S. Van Steenkiste and J. Schmidhuber, On the binding problem in artificial neural networks, *arXiv preprint arXiv:2012.05208* (2020).

[14] P. Hitzler and M.K. Sarker, Neuro-symbolic artificial intelligence: The state of the art (2022).

[15] I. Donadello, L. Serafini and A.D. Garcez, Logic tensor networks for semantic image interpretation, *arXiv preprint arXiv:1705.08968* (2017).

[16] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., Tensorflow: A system for large-scale machine learning, in: *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016.

[17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga and A. Lerer, Automatic differentiation in pytorch (2017).

[18] D. Haussler and M. Warmuth, *The probably approximately correct (PAC) and other learning models*, Springer, 1993.

[19] J.J. Dongarra, J. Du Croz, S. Hammarling and I.S. Duff, A set of level 3 basic linear algebra subprograms, *ACM Transactions on Mathematical Software (TOMS)* **16**(1) (1990), 1–17.

[20] M.-A. Krogel, S. Rawles, F. Železný, P.A. Flach, N. Lavrač and S. Wrobel, *Comparative evaluation of approaches to propositionalization*, Springer, 2003.

[21] A. Garcez, M. Gori, L. Lamb, L. Serafini, M. Spranger and S. Tran, Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning, *Journal of Applied Logics* **6**(4) (2019), 611–631.

[22] L. De Raedt, S. Dumančić, R. Manhaeve and G. Marra, From Statistical Relational to Neuro-Symbolic Artificial Intelligence, *arXiv preprint arXiv:2003.08316* (2020).

[23] M. Vitor, M.V.M. França and A.S. d'Avila Garcez, Neural Relational Learning Through Semi-Propositionalization of Bottom Clauses, *2015 AAAI Spring Symposium in Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches* (2015), 53–56. ISBN 1412920388.

[24] M.V. Franca, G. Zaverucha and A. Garcez, Fast relational learning using bottom clause propositionalization with artificial neural networks, *Machine learning* **94**(1) (2014), 81–104.

[25] A. Cropper, S. Dumančić and S.H. Muggleton, Turning 30: New ideas in inductive logic programming, *arXiv preprint arXiv:2002.11002* (2020).

[26] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S.G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou et al., Hybrid computing using a neural network with dynamic external memory, *Nature* **538**(7626) (2016), 471–476.

[27] A. Graves, G. Wayne and I. Danihelka, Neural turing machines, *arXiv preprint arXiv:1410.5401* (2014).

[28] L. Getoor and B. Taskar, *Introduction to Statistical Relational Learning*, 2007, p. 586. ISBN 0262072882.

[29] P. Frasconi, F. Costa, L. De Raedt and K. De Grave, klog: A language for logical and relational learning with kernels, *Artificial Intelligence* **217** (2014), 117–143.

[30] L. De Raedt and A. Kimmig, Probabilistic (logic) programming concepts, *Machine Learning* **100** (2015), 5–47.

[31] M. Richardson and P. Domingos, Markov logic networks, *Machine learning* (2006).

[32] R. Socher, A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C. Potts et al., Recursive deep models for semantic compositionality over a sentiment treebank, in: *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, Vol. 1631, Citeseer, 2013, p. 1642.

[33] J.B. Pollack, Recursive distributed representations, *Artificial Intelligence* **46**(1) (1990), 77–105.

[34] R. Socher, D. Chen, C.D. Manning and A. Ng, Reasoning with neural tensor networks for knowledge base completion, in: *Advances in neural information processing systems*, 2013.

[35] Z.C. Lipton, J. Berkowitz and C. Elkan, A critical review of recurrent neural networks for sequence learning, *arXiv preprint arXiv:1506.00019* (2015).
[36] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner and G. Monfardini, The graph neural network model., *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* **20**(1) (2009), 61–80.
[37] M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam and P. Vandergheynst, Geometric deep learning: going beyond euclidean data, *IEEE Signal Processing Magazine* **34**(4) (2017), 18–42.
[38] G. Šourek, V. Aschenbrenner, F. Železný, S. Schockaert and O. Kuželka, Lifted relational neural networks: Efficient learning of latent relational structures, *Journal of Artificial Intelligence Research* **62** (2018), 69–100.
[39] G. Šourek, F. Železný and O. Kuželka, Beyond graph neural networks with lifted relational neural networks, *Machine Learning* **110**(7) (2021), 1695–1738.
[40] G. Neubig, C. Dyer, Y. Goldberg, A. Matthews, W. Ammar, A. Anastasopoulos, M. Ballesteros, D. Chiang, D. Clothiaux, T. Cohn et al., Dynet: The dynamic neural network toolkit, *arXiv preprint arXiv:1701.03980* (2017).
[41] M. Looks, M. Herreshoff, D. Hutchins and P. Norvig, Deep learning with dynamic computation graphs, *arXiv preprint arXiv:1702.02181* (2017).
[42] M.J. Flynn, Very high-speed computing systems, *Proceedings of the IEEE* **54**(12) (1966), 1901–1909.
[43] S. Hooker, The hardware lottery, *Communications of the ACM* **64**(12) (2021), 58–65.
[44] K. Ahmed, T. Li, T. Ton, Q. Guo, K.-W. Chang, P. Kordjamshidi, V. Srikumar, G. Van den Broeck and S. Singh, PYLON: A PyTorch framework for learning with constraints, in: *NeurIPS 2021 Competitions and Demonstrations Track*, PMLR, 2022, pp. 319–324.
[45] G. Van den Broeck, N. Taghipour, W. Meert, J. Davis and L. De Raedt, Lifted probabilistic inference by first-order knowledge compilation, in: *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, AAAI Press/International Joint Conferences on Artificial Intelligence; Menlo . . . , 2011, pp. 2178–2185.
[46] M. Grohe, The logic of graph neural networks, in: *2021 36th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, IEEE, 2021, pp. 1–17.
[47] M. Fey and J.E. Lenssen, Fast graph representation learning with PyTorch Geometric, *arXiv preprint arXiv:1903.02428* (2019).
[48] M. Innes, Flux: Elegant Machine Learning with Julia, *Journal of Open Source Software* (2018). doi:10.21105/joss.00602.
[49] J. Choquette, W. Gandhi, O. Giroux, N. Stam and R. Krashinsky, NVIDIA A100 tensor core GPU: Performance and innovation, *IEEE Micro* **41**(2) (2021), 29–35.
[50] L.C. Lamb, A.S. d'Avila Garcez, M. Gori, M.O.R. Prates, P.H.C. Avelar and M.Y. Vardi, Graph Neural Networks Meet Neural-Symbolic Computing: A Survey and Perspective, in: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, ed., ijcai.org, 2020, pp. 4877–4884.
[51] J. Choquette, Nvidia hopper h100 gpu: Scaling performance, *IEEE Micro* (2023).
[52] G. Šourek, F. Železný and O. Kuželka, Lossless Compression of Structured Convolutional Models via Lifting, *International Conference on Learning Representations* (2021).
[53] R.R. Schaller, Moore's law: past, present and future, *IEEE spectrum* **34**(6) (1997), 52–59.
[54] W. Hu, M. Fey, H. Ren, M. Nakata, Y. Dong and J. Leskovec, Ogb-lsc: A large-scale challenge for machine learning on graphs, *arXiv preprint arXiv:2103.09430* (2021).