MENTAL MODELS FOR NEURAL MODELS

Owen Lewis
Google
lewiso@google.com

1 INTRODUCTION

Neural networks are statistical models that excel at capturing regularities that *tend* to be true, but are not so good at deducing facts that *have* to be true logically. An example from Go playing, one of the signature successes of neural networks, illustrates the difference. One of the first patterns taught to a novice human Go player is "two eyes" (Figure 1A), a configuration of stones that cannot be captured: even when surrounded by black stones, the white stones can still "breathe" through at least one of the two central empty spaces. It is clear that models like AlphaGo know about two eyes; they wouldn't be able to play as well as they do otherwise. But it is equally clear that the way the models learn about the pattern is very different from the way humans do. For models, eyes are simply a pattern observed, over many human lifetimes of play, to be correlated to some extent with game wins. A human player, though, requires essentially no play experience to grasp the special properties of two eyes: if she knows the rules of the game, she can articulate an informal proof showing how the rules entail the invulnerability of stones in eyes.

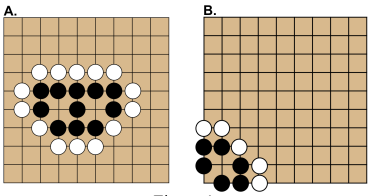

Figure 1.

Human-style deductive reasoning has several appealing computational properties. First, it avoids the notoriously large data requirements of standard neural models: humans can use knowledge and principles distilled in books, coaches' lessons and analyses of their own previous games to reach in a few years a level of play that would take AlphaGo centuries' worth of human-time game play to attain. Second, as ML models begin to make more and more safety-critical decisions, the interpretability of these decisions becomes increasingly important, and models that deduce their outputs from concretely expressible principles become more appealing. Finally, deductive models can benefit from explanation-based generalization (DeJong & Mooney (1986)): a model that understands *why* two eyes is important can immediately generalize to non-standard instantiations of the pattern like the one in Figure 1B.

Deduction is usually modeled as symbol manipulation: given a set of axioms, search for a sequence of inference rules that derives a target statement (see Ganesalingam & Gowers (2017) and many others). Deep learning research on deduction has largely followed this pattern, with neural networks being used to either guide the symbolic proof search process, as in Paliwal et al. (2019), or to parameterize a differentiable approximation to it, as in Rocktäschel & Riedel (2017).

The goal of this paper is to argue for an alternative approach which avoids search entirely and uses neural networks to learn "mental" models of a theory's semantics. Section 2 defines more precisely what we mean by "theory" and "model," and section 3 reviews evidence from cognitive science that humans reason in a model-based way. Finally, section 4 presents several approaches to implementing neural network model-based reasoners, some novel and some from the recent literature.

2 THEORIES AND MODELS

Here we introduce some terminology that frames the distinction between theories and models. A *logical language* consists of three sorts of symbols: function symbols ($+, \times, S(uccessor)$), predicate symbols ($\geq, =$), and constants (0). These symbols can be combined into *logical sentences* using connectives ($\wedge$, $\vee$, etc.).

An *axiom* is a sentence assumed to be true, and a *theory* is a set of sentences closed under entailment (i.e. a set of axioms, all sentences entailed by the axioms, all sentences entailed by those sentences,

and so on). A *model* of a theory is a choice of interpretation for each symbol in the theory: for some base set D, each constant is interpreted as an element of D, each k-ary function is a function $D^k \to D$, and each k-ary predicate is a function $D^k \to \{0, 1\}$. Loosely speaking, a theory is a syntactic object and a model gives it semantics. For example, a theory formalizing Euclidean geometry has no inherent geometric content (i.e. what we would picture as points, lines, etc.); any such content comes via an interpretation into a model.

This paper proposes to learn vector-space models of theories, using a set of axioms as training data to learn neural network interpretations of function and predicate symbols, and vector interpretations of constants. At test time, the learned model is used to predict truth values for a new set of logical sentences. The next section reviews evidence from cognitive science that people may reason in a similarly model-based way.

3 MODEL-BASED REASONING IN HUMANS

The idea of using models for deductive reasoning has been around since nearly the beginning of AI research (Gelernter & Rochester (1958)), and was formalized in cognitive science by Philip Johnson-Laird (Johnson-Laird (1983)), who showed that model-based reasoning can account for several peculiarities of human syllogistic reasoning. Johnson-Laird developed a detailed account of how model-based reasoning might work, particularly for syllogisms. His approach is a special case of model-based reasoning in general, and the vector-space models we propose in Section 4 are similar to it in spirit but different in setting and formulation.

If human reasoning was nothing other than symbol manipulation, its results should be uninfluenced by any meaning attached to the symbols. But if humans reason with models, then the assignment of meaning to symbols is an integral part of the reasoning process and a key determiner of the process's outputs. Several experiments seem to favor the model-based account. In one study (Simon (1996)), subjects were asked to play a game called "number scrabble," in which players take turns selecting digits from 1 to 9 from a central collection. The first player to collect three digits adding to 15 wins. Unsurprisingly, players struggled to discover a winning strategy. Observe, though, that this game is structurally identical to tic-tac-toe played on a 3×3 magic square. Despite this logical equivalence, though, number scrabble is considerably more difficult to learn. Later work (Kotovsky et al. (1985)) performed similar experiments on Towers of Hanoi.

Other experiments (Cosmides (1989)) showed similar interpretation-dependence for the Wason card selection task (Wason (1968)): subjects could solve the task if it was presented with a socio-political cover story, but not if it was presented abstractly. Finally, experiments summarized in Johnson-Laird (2010) show similar effects for syllogistic reasoning. In all, these results argue for a picture in which humans embue the symbols in a logical theory with meaning (i.e. interpret them in a model), and the meanings they choose determine, in part, the success of their reasoning.

Reaction time studies also support the idea that humans construct models of the domains they reason about. In (Moyer & Landauer (1967)), experimenters measured the time it took subjects to determine which of two textually-presented numbers was larger. Reaction times were approximately inverse-linear in the difference between the two numbers, and displayed similar patterns to RTs for analogue magnitude judgements. These results are consistent with a story on which humans model numbers as magnitudes (e.g. distances on a number line), and inconsistent with common symbolic models of arithmetic, Peano arithmetic for instance.

Two less quantitative arguments also support model-based reasoning. The first is the historical observation that areas of mathematics often emerge before axiomatizations of their objects of study. Calculus, for instance, was developed more than a century before the first axiomatization of the real numbers. This situation would seem impossible if the essence of mathematical reasoning was the symbolic manipulation of axioms, but it is entirely consistent with an account based instead on intuitively formulated mental models of how, e.g., the reals should behave.

Finally, it is suggestive, although by no means conclusive, that humans are not very good at formal symbolic reasoning. It takes years of training for most people to formalize and prove even simple statements like the fact that path connectivity in a graph is transitive (if I can get from a to b and from b to c, then I can get from a to c). This difficulty contrasts sharply with the facility of humans' everyday reasoning, and suggests that different mechanisms may underlie the two processes.

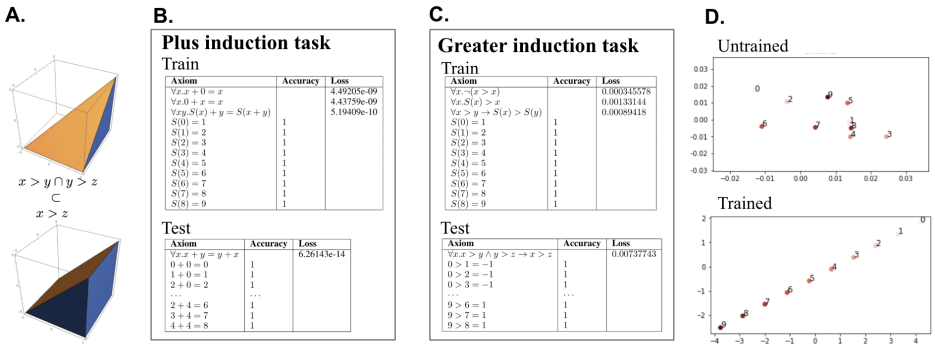

Figure 2.

4 MODEL-BASED REASONING IN NEURAL NETWORKS

Our goal is to build vector space models of logical theories in which neural networks are trained to represent a theory's functions and predicates, and embedding vectors to represent its constants. Once learned, such a model can be used to judge the truth or falsity of new statements.

The key source of difficulty in constructing such models is quantification. Suppose we have trained a model on the axioms of Peano arithmetic and learned a neural network representation of $+$. How are we to verify a statement like $\forall xy.x + y = y + x$? A null approach, which we explicitly wish to avoid, is enumeration: sample lots of inputs and verify that addition commutes for all of them. Not only is sampling an unreliable proxy for true quantification (lots of mathematical statements are only falsified by very rare counterexamples), it also seems to miss the essence of deduction, namely using some given statements to causally determine the truth value of others. So we seek an alternative method of evaluating quantified statements. Importantly, though, any such method has to agree locally with sampling: if we claim a model of addition commutes, it had better actually commute on real inputs; call this property *quantifier consistency*. This discussion frames the central challenge: how can we design neural networks whose behavior on all inputs can be characterized algorithmically?

This section presents several approaches to this challenge, two original and two from the literature. Given space constraints and the fact the two original models are very much works in progress, the goal of this section is not to present any method in detail, but rather to give the general flavor of some possible research directions.

4.1 QUANTIFICATION FROM PARAMETRIC CONSTRAINTS

Consider the statement $\forall xyz.x > y \wedge y > z \rightarrow x > z$, and suppose that $>$ is represented by a linear classifier. Then the implication reduces to a claim about set inclusion, as shown in Figure 2A: the halfspace in which $x > z$ must contain the intersection of the halfspaces where $x > y$ and $y > z$, and the parameterizations of the decision functions can be trained to make this inclusion true. Similar ideas have been used for taxonomy models in vision and NLP (Vilnis et al. (2018); Athiwaratkun & Wilson (2018); Mirzazadeh et al. (2015)). Overall, we can transform a statement about an infinite set of inputs into a statement about the finite set of decision function parameters.

A similar trick works for axioms with function symbols. In a statement like $\forall xy.S(x + y) = S(x) + y$, if $+$ and S are parameterized linearly, then the function compositions on the two sides of the equality are both products of matrices, and are therefore matrices themselves. Since the Frobenius norm F of the difference of these two matrices bounds the extent to which they can differ on an input, the equality can be enforced by minimizing F. Note that matrices are the simplest of various possible function and predicate parameterizations.

Figure 2 shows some preliminary results for a simple implementation along these lines. Given inductive definitions of $+$ (Figure 2B) and $>$ (Figure 2C) in terms of S (the successor function), the model can learn to apply the operators completely correctly, despite having seen no examples of such applications. Additionally, it can verify the truth of quantified statements ($\forall xy.x + y = y + x$ and $\forall xyz.x > y \wedge y > z \rightarrow x > z$ that are not entirely trivial to prove from the axioms. Finally,

the model learns sensible linear embeddings of the constants $0 \ldots 8$ from untrained initializations (Figure 2D).

4.2 QUANTIFICATION FROM SYMMETRY

Given a pair of points (p_1, p_2) in the plane, it is easy to verify that they are colinear. Moreover, the additional observation that (p_1, p_2) is related to any other pair of points by a colinearity-preserving Euclidean transformation licenses a universal generalization: $\forall q_1 q_2 . \texttt{colinear}(q_1, q_2)$.

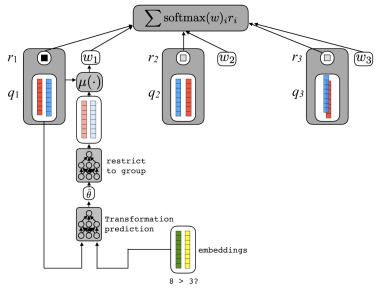

Figure 3.

If we squint a bit, we can see similar sorts of reasoning — forming abstractions by quotienting by a group of transformations — in other areas of math. For our path transitivity example above, for instance, a person might realize that changes to the numbers and locations of the nodes on a path is irrelevant to the path's end-to-end connectivity, an observation that would license reasoning in an abstract "quotient" domain in which paths are represented as smooth lines. Such processes may also account for some failures of human reasoning. The superficial convincingness, for instance, of inductive fallacies like the proof that all horses are the same color (after Pólya (1954)) may derive from our tendency to abstract sets of unknown cardinality into unenumerable "blobs." Within mathematics, the framing of logic via group invariances has been explored in, for example, Mautner (1946); Tarski & Corcoran (1986); Marquis (2008).

Work in computer vision and speech recognition has developed a range of tools for building invariance to nuisance transformations. Can we use these models to do quantification? More speculatively, could the invariances present in human's sensory systems underlie, in part, our ability to do deduction (Lakoff & Núñez (2000))?

Figure 3 shows the architecture of a neural model designed to capture quantification via symmetry. Briefly, the model parameterizes each k-ary predicate symbol P with a set of m "prototype inputs" $\{q_i\}_{i=1}^m$, each paired with a response r_i. Each prototype is a k-tuple of trainable vectors, and is interpreted as the representative of one orbit of P's input space under the action of a group G which leaves P invariant. In a sense, then, the entirety of P's behavior is characterized by its behavior on the prototypes, and statements quantified over the infinite input space can be replaced with finite conjunctions or disjunctions over the prototypes.

For example, in reasoning about order relations on the real line, G is the group of shifts and positive scales, and there are three prototypes: one in which the second element is larger than the first, one where the first is larger than the second, and one where the elements are equal.

To evaluate P on a set of constant-symbol inputs x, the model uses a self-supervised variant of a spatial transformer network (Jaderberg et al. (2015)) to estimate for each prototype q_i the transformation parameters θ_i^* that come closest to transforming the vector embeddings of x into q_i, recording the residuals in a weights vector $w_i = \|g_{\theta_i^*}(x) - q_i\|$. The model then returns a softmax-weighted average of the prototype responses r_i: $\sum_i \texttt{softmax}(w)_i r_i$.

Preliminary experiments using parts of this proposed model to reason about total orders are somewhat promising, showing correct orderings of approximately 80 out of 90 pairs of test inputs.

4.3 OTHER APPROACHES

Two other ideas proposed in the literature represent interesting alternative approaches to model-based quantification. The first of these (Minervini et al. (2017)) proposes to model quantification as adversarial search: a universal statement is true if a differentiable optimization process cannot discover a counterexample. So far, this approach has been used for link prediction in knowledge bases; it would be interesting to explore extending it to make judgements about quantified statements.

Second, (Evans et al. (2018)) proposes to evaluate statements by testing them on model-optimized input sets. Results for propositional logic are good; a similar approach may be possible for first- or higher-order logic.

REFERENCES

Ben Athiwaratkun and Andrew Gordon Wilson. Hierarchical density order embeddings. *arXiv preprint arXiv:1804.09843*, 2018.

Leda Cosmides. The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. *Cognition*, 31(3):187–276, 1989.

Gerald DeJong and Raymond Mooney. Explanation-based learning: An alternative view. *Machine learning*, 1(2):145–176, 1986.

Richard Evans, David Saxton, David Amos, Pushmeet Kohli, and Edward Grefenstette. Can neural networks understand logical entailment? *arXiv preprint arXiv:1802.08535*, 2018.

Mohan Ganesalingam and William Timothy Gowers. A fully automatic theorem prover with human-style output. *Journal of Automated Reasoning*, 58(2):253–291, 2017.

Herbert L Gelernter and Nathaniel Rochester. Intelligent behavior in problem-solving machines. *IBM Journal of Research and Development*, 2(4):336–345, 1958.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pp. 2017–2025, 2015.

Philip N Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.

Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press, 1983.

Kenneth Kotovsky, John R Hayes, and Herbert A Simon. Why are some problems hard? evidence from tower of hanoi. *Cognitive psychology*, 17(2):248–294, 1985.

George Lakoff and Rafael Núñez. *Where mathematics comes from*, volume 6. New York: Basic Books, 2000.

Jean-Pierre Marquis. *From a geometrical point of view: a study of the history and philosophy of category theory*, volume 14. Springer Science & Business Media, 2008.

FI Mautner. An extension of klein's erlanger program: Logic as invariant-theory. *American Journal of Mathematics*, 68(3):345–384, 1946.

Pasquale Minervini, Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. Adversarial sets for regularising neural link predictors. *arXiv preprint arXiv:1707.07596*, 2017.

Farzaneh Mirzazadeh, Siamak Ravanbakhsh, Nan Ding, and Dale Schuurmans. Embedding inference for structured multilabel prediction. In *Advances in Neural Information Processing Systems*, pp. 3555–3563, 2015.

Robert S Moyer and Thomas K Landauer. Time required for judgements of numerical inequality. *Nature*, 215(5109):1519–1520, 1967.

Aditya Paliwal, Sarah Loos, Markus Rabe, Kshitij Bansal, and Christian Szegedy. Graph representations for higher-order logic and theorem proving. *arXiv preprint arXiv:1905.10006*, 2019.

George Pólya. *Mathematics and plausible reasoning: Induction and analogy in mathematics*, volume 1. Princeton University Press, 1954.

Tim Rocktäschel and Sebastian Riedel. End-to-end differentiable proving. In *Advances in Neural Information Processing Systems*, pp. 3788–3800, 2017.

Herbert A. Simon. *The Sciences of the Artificial (3rd Ed.)*. MIT Press, Cambridge, MA, USA, 1996. ISBN 0262691914.

Alfred Tarski and John Corcoran. What are logical notions? *History and philosophy of logic*, 7(2): 143–154, 1986.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. *arXiv preprint arXiv:1805.06627*, 2018.

Peter C Wason. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3): 273–281, 1968.