Neural Network Semantics

Summary of Proposed Research

CALEB SCHULTZ KISBY

In the last 15 years, modern artificial intelligence (AI) systems have shown unprecedented success at learning from data with little human guidance. Consider for example large language models such as Llama and GPT [1; 8; 31], which have taken the world by storm with their ability to learn to converse in English merely from unstructured text data they scrape off the web. Or consider AlphaGo [26], which learned to play Go at a human expert level by repeatedly playing against itself. These breakthroughs in machine learning are in large part thanks to the widespread use of neural networks – brain-inspired computational models that are flexible and excel at learning from unstructured data.

But the danger of neural networks is that they come with no safety, reliability, or correctness guarantees. If you play with systems like GPT long enough, you eventually realize that they carry all sorts of misconceptions, make silly logical mistakes, and are quite happy to spew out disinformation [27]. Neural networks also lack transparency, which means diagnosing and correcting these errors is not feasible. (Imagine trying to determine which neurons and connections are responsible for believing that a sailfish is a mammal!) In practice, a computational learner is often a 'black-box' whose correct inferences, mistakes, and biases lack interpretation and explanation.

How can we better understand and control this seemingly black-box behavior of neural networks? The answer lies in symbolic (logic) systems, which were commonly used to model reasoning and intelligent behavior prior to the rapid growth of neural network systems. In contrast with neural networks, symbolic systems provide explicit rules for their reasoning in a human-interpretable language. However, this purely logic-based approach was largely abandoned due to logic's inability at the time to model flexible learning or update (known as the *frame problem* in AI [22; 25]).

There is still hope that we might be able to integrate neural networks and symbolic systems while retaining the advantages of both. The field of *neuro-symbolic AI* has emerged in response to this possibility [2; 5; 24]. As a result of this effort, there are now many different proposals for neuro-symbolic systems, including Logic Tensor Networks [3], Distributed Alignment Search [10], DeepProbLog [20], Logic Explained Networks [7], and neural network fibring [9]. But these systems form a scattered picture; some unifying perspective or theory is needed. In the preface to a recent neuro-symbolic survey book [5], Frank van Harmelen writes:

> What are the possible interactions between knowledge and learning? Can reasoning be used as a symbolic prior for learning … Can symbolic constraints be enforced on data-driven systems to make them safer? Or less biased? Or can, vice versa, learning be used to yield symbolic knowledge? … **neuro-symbolic systems currently lack a theory that even begins to ask these questions, let alone answer them.**

In this thesis, I will offer a new unifying perspective that sheds light on these questions. The basis for many neuro-symbolic systems is that they encode logical information into neural networks, or conversely, encode neural networks as models in logic [23]. Given these translations,

neural networks and logic models are able to represent the same information. This suggests that we can think of neural networks in the same way as a logician would think about a model.

The point here is to take this idea as far as it will go: I will develop logics with these *neural network semantics*, whose formulas are interpreted in terms of binary neural networks. First, I will consider *static* conditional and modal logics whose operators are given by the closure (or forward propagation) of signals in the neural network. Next, I will give a *dynamic* logic (inspired by Dynamic Epistemic Logic [28; 29; 30]) with an operator for Hebbian learning [14], a simple neural network update policy. Along the way, I will show how foundational questions about neural networks become natural and answerable questions in logic. Let's consider three questions that are natural to ask about for any logical system: *soundness*, *completeness*, and *expressivity*.

Soundness. What axioms are sound for the semantics? In neural network semantics, this question becomes: What properties can we formally verify for neural network inference? In the dynamic logic setting: What properties can we formally verify for neural network *learning policies*?

Completeness. What are the complete axioms for the semantics? This is equivalent to the question of whether we can build a model that obeys a set of logical constraints Γ. In neural network semantics, completness asks whether we can build a *neural network* that obeys Γ. And for dynamic logic, this asks whether we can build a neural network that obeys Γ before and after learning takes place. This is a key to the AI Alignment problem, which requires building neural networks with these kinds of guarantees.

Expressivity. What formulas can the semantics express or define? How does the expressive power of two different semantics compare? For neural networks, expressivity asks what kinds of formulas neural networks are capable of representing. Additionally, it provides a metric for comparing the power of neural networks against traditional models in logic. In the dynamic setting: What kinds of *learning policies* are neural networks able to support?

The History of Neural Network Semantics. I'll conclude this summary by situating my thesis in a broader history of development in neural network semantics, making sure to clarify which ideas are my own. The core idea behind neural network semantics – that neural networks can be treated as models for logic – actually dates back to the very first paper on neural networks. In McCulloch and Pitts [21], logical formulas are mapped directly to individual neurons in the net. This approach suffers from the well-known "grandmother cell" problem [13]: it is cognitively implausible that an individual neuron could represent a complex concept such as "grandmother". Instead, concepts in brain networks are distributed across multiple neurons at once.

Neural network semantics is based on a recent reimagining of this approach [4; 19], where logical formulas are mapped to sets of neurons rather than to individual neurons. Early work established the correspondence between forward propagation and nonmonotonic inference [4; 6; 17; 18]. More recently, [11; 12] proved soundness for forward propagation over fuzzy neural networks. My own work [15; 16] applies ideas from Dynamic Epistemic Logic to model Hebbian learning in neural network semantics. The key results of this work are the first ever soundness and completeness theorems for any learning policy on neural networks. This thesis will present these results, as well as ongoing work on the expressivity of neural network learning.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat et al. GPT-4 technical report. *ArXiv preprint arXiv:2303.08774*, 2023.

[2] Sebastian Bader and Pascal Hitzler. Dimensions of neural-symbolic integration – A structured survey. In *We Will Show Them! Essays in Honour of Dov Gabbay, Volume 1*, pages 167–194. College Publications, 2005.

[3] Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic Tensor Networks. *Artificial Intelligence*, 303:103649, 2022.

[4] Christian Balkenius and Peter Gärdenfors. Nonmonotonic inferences in neural networks. In *KR*, pages 32–39. Morgan Kaufmann, 1991.

[5] Tarek R Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Priscila Machado Vieira Lima, Leo de Penning et al. Neural-symbolic learning and reasoning: A survey and interpretation. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, pages 1–51. IOS press, 2021.

[6] Reinhard Blutner. Nonmonotonic inferences and neural networks. *Synthese*, 142:143–174, 2004.

[7] Gabriele Ciravegna, Pietro Barbiero, Francesco Giannini, Marco Gori, Pietro Lió, Marco Maggini, and Stefano Melacci. Logic Explained Networks. *Artificial Intelligence*, 314:103822, 2023.

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan et al. The Llama 3 herd of models. *ArXiv preprint arXiv:2407.21783*, 2024.

[9] Artur SD'Avila Garcez, Luis C Lamb, and Dov M Gabbay. *Neural-Symbolic Cognitive Reasoning*. Springer Science & Business Media, 2008.

[10] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR, 2024.

[11] Laura Giordano, Valentina Gliozzi, and Daniele Theseider Dupré. A conditional, a fuzzy and a probabilistic interpretation of self-organizing maps. *Journal of Logic and Computation*, 32(2):178–205, 2022.

[12] Laura Giordano and Daniele Theseider Dupré. Weighted defeasible knowledge bases and a multipreference semantics for a deep neural network model. In *Logics in Artificial Intelligence: 17th European Conference, JELIA 2021, Virtual Event, May 17–20, 2021, Proceedings 17*, pages 225–242. Springer, 2021.

[13] Charles G Gross. Genealogy of the "grandmother cell". *The Neuroscientist*, 8(5):512–518, 2002.

[14] Donald Hebb. *The Organization of Behavior*. Psychology Press, apr 1949.

[15] Caleb Kisby, Saúl Blanco, and Lawrence Moss. The logic of hebbian learning. In *The International FLAIRS Conference Proceedings*, volume 35. 2022.

[16] Caleb Schultz Kisby, Saúl A Blanco, and Lawrence S Moss. What do hebbian learners learn? Reduction axioms for iterated Hebbian learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14894–14901. 2024.

[17] Hannes Leitgeb. Nonmonotonic reasoning by inhibition nets. *Artificial Intelligence*, 128(1-2):161–201, 2001.

[18] Hannes Leitgeb. Nonmonotonic reasoning by inhibition nets II. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(supp02):105–135, 2003.

[19] Hannes Leitgeb. Neural network models of conditionals. In *Introduction to Formal Philosophy*, pages 147–176. Springer, 2018.

[20] Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Neural probabilistic logic programming in DeepProbLog. *Artificial Intelligence*, 298:103504, 2021.

[21] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, dec 1943.

[22] Drew McDermott. A critique of pure reason. *Computational intelligence*, 3(3):151–160, 1987.

[23] Simon Odense and Artur S. d'Avila Garcez. A semantic framework for neural-symbolic computing. *ArXiv*, abs/2212.12050, 2022.

[24] Md Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. Neuro-Symbolic Artificial Intelligence: Current Trends. *AI Communications*, 34, 2022 2022.

[25] Murray Shanahan. The frame problem. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016.

[26] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[27] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. Understanding the capabilities, limitations, and societal impact of large language models. *ArXiv preprint arXiv:2102.02503*, 2021.

[28] Johan Van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2011.

[29] Johan Van Benthem and Sonja Smets. Dynamic logics of belief change. In H. Van Ditmarsch, J. Halpern, W. van der Hoek, and B. Kooi, editors, *Handbook of Epistemic Logic*, pages 313–393. College Publications, London, UK, 2015.

[30] Hans Van Ditmarsch, Wiebe van Der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*, volume 337. Springer, 2007.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.