Towards Deep Cognitive Probabilistic Programming

Research Background

I believe computation is a vital tool for comprehending things at all **levels of organization** in the world: from fundamental particles, to biomolecules that make life possible, to the wonders of biological intelligence and social structures. I dedicated my undergraduate research towards understanding complex biological systems using a spectrum of computational methods: from Boolean network dynamics for comprehending cellular apoptosis, to more abstractly, optimization algorithms inquiring how biological networks evolve interesting properties unseen in random networks. My bachelor's thesis focussed on quantifying local pairwise interactions that compose global real-world networks. This research path helped me appreciate the power of **abstraction** and **compositional modeling**. My work at the Wyss Institute at Harvard as a systems biologist allowed me to explore unsupervised deep learning methods for characterizing protein sequences, and hierarchical Bayesian models to develop probabilistic models of host tolerance that can encode conditional dependencies of biomolecules. My **concomitant study** of Bayesian modelling, deep learning and composable biological systems has strongly motivated me towards engineering deep probabilistic programs. I believe that cognitively informed probabilistic models are a principled way to fathom biological intelligence and expand horizons of artificial intelligence.

Motivation for Bayesian Deep Learning

In the past decade, there has been a captivating interest in using deep neural networks within the field of artificial intelligence and machine learning for a better understanding of the world [LeCun 2015]. Originally inspired by how neurons in the neocortex integrate-and-fire to transform input to output representations and hence “compute”, artificial feedforward networks gained traction in the mid-20th century. Its evolved descendants like the convolutional and recurrent neural nets have become state-of-the-art in machine understanding of images [Krizhevsky 2012] and language [Sutskever 2014]. Besides advances in these model architectures and learning algorithms, the success of deep learning also had a lot to do with the explosion of big data. These deep neural networks have many parameters to be tuned, which needs a large amount of data so that these networks do not overfit. To that end, there has been an increasing focus on creating strategies that allow these networks to **generalize to unseen data**, especially when there’s not a lot of it. Some of these techniques appear in the flavour of constraints on model parameters to better define the “ill-posed” problem of learning, like via L1 or L2 norm regularization. A seemingly ad hoc method while training very deep networks is of dropout, wherein subject to some probability, every neuron is dropped off from the forward and backward passes. These are very pragmatic methods used in deep learning today, with some mathematical backing, but most of their interpretation is rather lacking.

On the other hand, the less pragmatic but mathematically nuanced field of Bayesian statistics has allowed us to interpret these ideas more principally as imposition of uncertainty estimates over the space of inputs and model parameters. Many **equivalences** have been established between popular **regularization techniques and Bayesian probability**, treating model parameters as random variables following certain distributions. Assuming a Gaussian prior distribution and estimating the parameter posterior gives us the L2 norm, while placing a Laplacian prior renders the L1 norm, and something as ad hoc as dropout too can be seen as a proxy for placement of a spike-and-slab prior on the weights of the net [Gal 2016]. Choosing an appropriate prior can encourage simpler models to be preferred over complex ones, thus supporting a built-in **Occam's razor**.

Beyond these theoretical benefits, there are many pragmatic benefits of pursuing "Bayesian" deep learning. We can now talk about model predictions with estimates of **uncertainty** over them, and similarly of models themselves with uncertainties in the model space, which can proxy for a confidence in the model [Ghahramani 2015]. Given the surging emergence of autonomous driving, and use of AI for medicine and governance, ML algorithms are already affecting human lives in a very tangible way. Having model uncertainties will ensure that, especially in data-poor regimes, we diagnose the pitfalls of our models, issue relevant prognosis and plan procurement of future data and updates to the model. Additionally, the **interpretability** of probabilistic models can expose a suitable interface to "read-off" the black-box of these deep learning models in a more human-understandable manner.

Present Art in Bayesian Models of Cognition

Probabilistic modeling has been used extensively in the last couple of decades to directly uncover theories of biological intelligence, especially in the avatar of Bayesian models of cognition [Griffiths 2008]. Human learning of intuitive forms of structures in data was shown to be a hierarchical Bayesian model [Kemp 2008], which I studied and extended as an undergraduate project, hypothesizing the observed space to lie on a low-dimensional manifold. Human-level concept learning has been modeled as **probabilistic program induction**, wherein concepts are represented as simple programs that best explain the observed data under appropriate Bayesian criterion [Lake 2015]. This model has an ability to learn even with few examples of the order of tens vis-à-vis thousands needed for deep networks, beating the latter at one-shot learning in some visual domains including character recognition. The reason they work well is because, intuitively, the lack of data is compensated by imposition of prior knowledge that constrains the model space. Bayesian programs exploit the **compositional hierarchy** of visual concepts –that are composed by stochastic recombination of parts, which are themselves combination of further low-level subparts– encoded into appropriate symbolic representations following relevant distributions.

This hierarchy of representation is not an entirely new idea, and is often observed as an emerging property in deep learning. For instance, architecturally inspired by the ventral stream LGN-V1-V2-V4-IT, lower layers of a CNN appear to capture low-level features like corners or edges, while higher layers capture visual categories and concepts like "faces" vs. "animals" [LeCun 2015]. The notion remains implicit here, unlike in probabilistic programming. One of my undergraduate research projects tried explicating such hierarchical representations. For classifying plankton images into 120 categories, a hierarchical model stacked according to the plankton's evolutionary tree performed much better than a deep learning model agnostic to this structure information about the output classes. Growing research on the theoretical underpinnings of deep learning shows deep nets can guarantee avoiding the curse of dimensionality when a problem's input-output mappings are compositional [Poggio 2017]. A subset of such problems is those consisting of functions composed of a hierarchy of constituent functions that are local. It is this **compositional locality**, and not mere weight sharing, that gives ConvNets an exponential advantage. Recent work shows how the weights of initial layers matter more than those of the top ones [Raghu 2017], and that deep nets could, rather unnervingly, learn randomly assigned class labels in the said top layers [Zhang 2017]. Perhaps placing a structured hierarchical prior on the top-level features is a form of regularization that won't allow an overfit onto random mappings. To that end, Bayesian and deep learning hybrids, **hierarchical-deep models**, have been suggested and shown to learn well with even few examples [Salakhutdinov 2013]. Using fully probabilistic models by themselves is not entirely pragmatic in high-dimensional spaces, owing to difficulties of sampling, but advances in deep Gaussian processes and methods of approximate variational inference even allow a fully Bayesian treatment of deep networks [Damianou 2013, Lee 2017, Ghahramani 2015]. Acting as priors on functional mappings, GP kernels can be stacked or composed, and have been shown to be a good characterization of inductive biases observed in intuitive functions in human cognition [Schulz 2016].

Bayesian techniques have been used to model higher-level cognition and improve AI, but deep learning techniques have long been inspired by lower-level neuroscience. As mentioned earlier, the perceptron was modeled around a biological integrate-and-fire neuron. The idea of dropout, as is implicit in its Bayesian interpretation, can be motivated by the inherent stochasticity of neuronal firing according to Poisson-like statistics [Hinton 2012]. The idea of **distributed representation** of words and other real-world concepts, as high-dimensional vectors, also exploits the neurocomputing models of sentence processing [Hassabis 2017]. Amongst the contemporary challenges of AI lie the hallmarks of human intelligence –the ability to generalize, to transfer information between known and unknown contexts, to simulate and imagine new scenarios, and to learn *to learn*– some of which are being pursued by hierarchical Bayesian programs [Lake 2015], some by deep generative neural network models [Socher 2013, Kaiser 2017, Janner 2017], others by both [Kulkarni 2015]. There's a sense of **complementarity** in these two: high-level versus low-level cognition, symbolic low-dimensional versus distributed high-dimensional representation, pragmatic but uninterpretable versus unpragmatic but interpretable. To develop a truly intelligent system, and for reasons delineated above, we need to invest in a principled integration of the two. This is precisely what the deep probabilistic program attempts to do [Tran 2017].

Future of Deep Cognitive Probabilistic Programming

We've been able to exploit some of the quirks of neural network training and generate **adversarial** examples that can easily fool these models [Szegedy 2013]. This can be a threat to safety in AI systems, but more importantly, it provides motivation towards creating more robust model architectures. The human brain is a clear example of one such architecture. Ongoing advances in incorporating more cortically compliant architectural elements, such as the Capsule Network, have shown to withstand such white-box attacks much better [Sabour 2017]. The idea of distributed representations itself holds a clear analogy to neocortical spikings that encode a representation of real-world objects. Neuronal representation in the IT/V4 cortex of the visual stream is purported to encode very abstract visual concepts that are invariant to local changes in the input visual stimuli, such as the pose or shading of the object [Cadieu 2013]. There's a need to explore further how the brain generates and stores such **robust representations**. Using multi-electrode neuronal recordings, this data can be analyzed not just to understand neocortical representations better, but to ground and regularize representations of objects in AI systems. Algorithms inspired by how the brain embeds large amounts of real-world knowledge in a **compositional vector space**, such as in holographic embeddings inspired by associative memory [Nickel 2016], are not only faster to learn but may also produce more robust distributed representations. Alongside cognitively motivated model architectures, we must pursue research into **cognitively structured compositional priors**.

Inspiration from neuro and cognitive science to AI has been largely at the computational and algorithmic levels of David Marr's "levels of analysis" [Marr 1976], but plenty of avenues reside at the implementational level as well. Spatial topology of the cortex could enlighten us about the existence and relevance of feedback connections between neurons, and the synaptic density could be closely related to the amount of information content of the representations they harbour. The biophysical mechanisms by which neurons works are, even today, not completely well-understood. Electrical signals are tightly regulated phenomena sprinkled with opportunities for stochasticity, in which many metabolites, ion channels, proteins, their receptors, and even the microtubular vasculature take part [Hameroff 2014]. This intraneuronal stochastic orchestration can itself be modeled probabilistically, evident in some of my own work in systems biology, whose posterior estimate could be the observed integrate-and-fire behaviour that motivated the humble perceptron. It is in that sense that we come back to the need for deep hierarchical probabilistic models. At a **physical level**, the compositional structure of biological intelligence –biomolecular interactions composing to form neurons composing to form neural networks that encode highly distributed representations– can in principle be mirrored by a very deep probabilistic program. At an **algorithmic level**, instead of “error backpropagation”, methods of training that are biologically “more” plausible have come to closely depict learning stacked local layers of denoising auto-encoders with probabilistic interpretations [Bengio 2015]. And it is perhaps no surprise then that right at the highest **computational level** of our analysis of cognition, a Markov chain can be setup within human subjects to directly sample from their posterior distributions [Sanborn 2008]. This ubiquity of deep probabilistic programming is possibly a **mathematical artifact of compositionality**. But given that intelligence continues to lie in dealing with real-world uncertainty, it serves as a crucial narrative that encourages us to explore this burgeoning intersection further, if we are to concordantly demystify biological and refine artificial intelligence.

References

Bengio, Yoshua, et al. "Towards biologically plausible deep learning." *arXiv preprint arXiv:1502.04156* (2015).

Cadieu, Charles F., et al. "The neural representation benchmark and its evaluation on brain and machine." *arXiv preprint arXiv:1301.3530* (2013).

Damianou, Andreas, and Neil Lawrence. "Deep gaussian processes." *Artificial Intelligence and Statistics.* 2013.

Gal, Yarin, and Zoubin Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. 2016.

Ghahramani, Zoubin. "Probabilistic machine learning and artificial intelligence." *Nature* 521.7553 (2015): 452-459.

Griffiths, Thomas L., Charles Kemp, and Joshua B. Tenenbaum. "Bayesian models of cognition." (2008).

Hameroff, Stuart, and Roger Penrose. "Consciousness in the universe: A review of the ‘Orch OR’ theory." *Physics of life reviews* 11.1 (2014): 39-78.

Hassabis, Demis, et al. "Neuroscience-inspired artificial intelligence." *Neuron* 95.2 (2017): 245-258.

Hinton, Geoffrey E., et al. "Improving neural networks by preventing co-adaptation of feature detectors." *arXiv preprint arXiv:1207.0580* (2012).

Janner, Michael, et al. "Self-supervised intrinsic image decomposition." *Advances in Neural Information Processing Systems.* 2017.

Kaiser, Lukasz, et al. "One model to learn them all." *arXiv preprint arXiv:1706.05137* (2017).

Kemp, Charles, and Joshua B. Tenenbaum. "The discovery of structural form." *Proceedings of the National Academy of Sciences* 105.31 (2008): 10687-10692.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems.* 2012.

Kulkarni, Tejas D., et al. "Deep convolutional inverse graphics network." *Advances in Neural Information Processing Systems*. 2015.

Lake, Brenden M., Ruslan Salakhutdinov, and Joshua B. Tenenbaum. "Human-level concept learning through probabilistic program induction." *Science* 350.6266 (2015): 1332-1338.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *Nature* 521.7553 (2015): 436-444.

Lee, Jaehoon, et al. "Deep Neural Networks as Gaussian Processes." *arXiv preprint arXiv:1711.00165* (2017).

Marr, David, and Tomaso Poggio. "From understanding computation to understanding neural circuitry." (1976).

Nickel, Maximilian, Lorenzo Rosasco, and Tomaso A. Poggio. "Holographic Embeddings of Knowledge Graphs." *AAAI*. 2016.

Poggio, Tomaso, et al. "Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review." *International Journal of Automation and Computing* (2017): 1-17.

Raghu, Maithra, et al. "On the expressive power of deep neural networks." *arXiv preprint arXiv:1606.05336* (2016).

Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. "Dynamic Routing Between Capsules." *Advances in Neural Information Processing Systems*. 2017.

Sanborn, Adam, and Thomas L. Griffiths. "Markov chain Monte Carlo with people." *Advances in neural information processing systems*. 2008.

Salakhutdinov, Ruslan, Joshua B. Tenenbaum, and Antonio Torralba. "Learning with hierarchical-deep models." *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1958-1971.

Socher, Richard, et al. "Zero-shot learning through cross-modal transfer." *Advances in neural information processing systems*. 2013.

Schulz, Eric, et al. "Probing the compositionality of intuitive functions." *Advances in neural information processing systems*. 2016.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.

Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199* (2013).

Tran, Dustin, et al. "Deep probabilistic programming." *arXiv preprint arXiv:1701.03757* (2017).

Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization." *arXiv preprint arXiv:1611.03530*(2016).