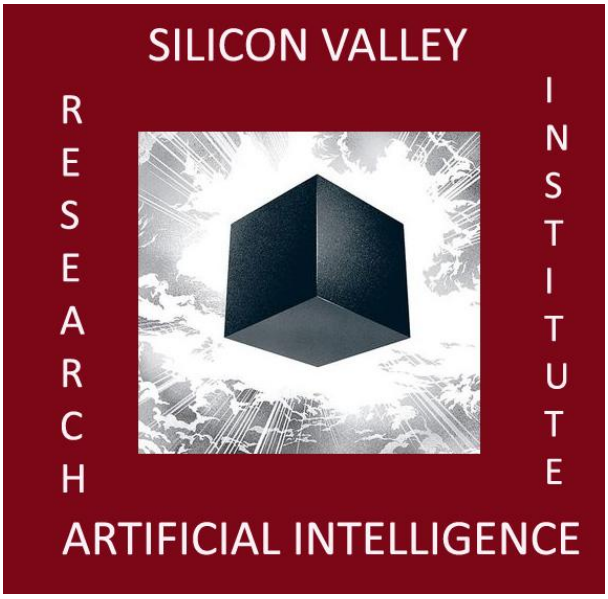

Bayesian Thinking: a brief introduction

The fundamental limitation of neural networks is that they need to be "trained" with thousands if not millions of examples before they can "learn". A child can usually learn a new concept from a single example, creating a generalization that she will be able to apply to similar objects or situations ("one-shot generalization").

Probabilistic induction was one of the very first proposals for artificial intelligence, notably Solomonoff's "Inductive Inference Machine" of 1956; and Judea Pearl's Bayesian reasoning ("Reverend Bayes on Inference Engines", 1982) provided the toolbox for probabilistic computation. Their motivation was the same that had originally motivated Fermi and Ulam: probabilistic reasoning is needed when the exact algorithm is too complicated and would result in unacceptable response time. In these cases it is preferrable to find an approximate (but quick) solution. In between, a generalization of subjective probability, the "theory of evidence", was developed by Arthur Dempster at Harvard University ("Upper and Lower Probabilities Induced by a Multivalued Mapping", 1967) and Glenn Shafer at Princeton University, who published the book "A Mathematical Theory of Evidence" (1976).

Morris DeGroot of Carnegie Mellon University published "Optimal Statistical Decisions" in 1970, but the book that revived probabilistic reasoning was the English translation of DeFinetti's book "Theory of Probability" (1970), many years after his 1928 talk.

Meanwhile, in 1970 Keith Hastings at the University of Toronto generalized the Metropolis algorithm ("Monte Carlo Sampling Methods using Markov Chains and their Applications", 1970).

Another popular MCMC algorithm, "Gibbs sampling", was developed in 1984 by the brothers Stuart and Donald Geman (respectively at Brown University and the University of Massachusetts) who were working on computer vision ("Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", 1984); Judea Pearl soon introduced Gibbs sampling in his Bayesian networks, also known as "belief networks" and "causal networks" ("Evidential Reasoning using Stochastic Simulation", 1987).

MCMC became especially popular after Adrian Smith at the University of Nottingham, in collaboration with Alan Gelfand, showed how these algorithms can be used in a wide variety of cases ("Sampling-based Approaches to Calculating Marginal Densities", 1990). Just then the statistical software BUGS (Bayesian inference Using Gibbs Sampling) was becoming available Bayesian inference using Markov chain Monte Carlo (MCMC): it had been developed since 1989 at the University of Cambridge by Andrew Thomas, working under David Spiegelhalter. In 1993 the International Society for Bayesian Analysis met in San Francisco for its first conference. Finally, Radford Neal of the University of Toronto published a report titled "Probabilistic Inference Using Markov Chain Monte Carlo Methods" (1993) that said it all, and in 1996 published the book "Bayesian Learning for Neural Networks".

Meanwhile, in 1983 Geoffrey Hinton at Carnegie Mellon University and Terry Sejnowski at John Hopkins University had introduced the Boltzmann machine that was still evolving. Because the computational cost of the undirected models is so high, the Boltzmann machine used an approximation method that was de facto already Gibbs sampling. In 1992 Radford Neal at the University of Toronto added "direction" to the connections (that were originally undirected, i.e. symmetric) of the Boltzmann machine in order to improve the training process. Pearl had described belief nets to represent expert knowledge (knowledge elicited from humans). Neal showed that belief nets can learn by themselves. In 1995 Hinton and Neal worked with the British neuroscientist Peter Dayan, who was keen on finding a similarity with Hermann von Helmholtz's theory of perception, and designed the Helmholtz Machine ("The Helmholtz Machine", 1995) for which they invented the "wake-sleep" algorithm for unsupervised learning ("The Wake-sleep Algorithm for Unsupervised Neural Networks", 1995). The wake-sleep algorithm, conceived purely from neuroscience, trains top-down and bottom-up probabilistic models (i.e. generative and inference models) against each other in a multilayer network of stochastic neurons. It is a form of "variational Bayesian learning": it approximates Bayesian inference when it becomes intractable, as it is typically the case in multilayer networks.

Meanwhile, the Swedish statistician Ulf Grenander (who in 1972 had established the Brown University Pattern Theory Group) fostered a conceptual revolution in the way a computer should describe knowledge of the world: not as concepts but as patterns. His "general pattern theory" provided mathematical tools for identifying the hidden variables of a data set. Grenander's pupil David Mumford studied the visual cortex and came up with a hierarchy of modules in which inference is Bayesian, and it is propagated both up and down ("On The Computational Architecture Of The Neocortex II", 1992). The assumption was that feedforward/feedback loops in the visual region integrate top-down expectations and bottom-up observations via probabilistic inference. Basically, Mumford applied hierarchical Bayesian inference to model how the brain works.

Hinton's Helmholtz machine of 1995 was de facto an implementation of those ideas: an unsupervised learning algorithm to discover the hidden structure of a set of data based on Mumford's and Grenander's ideas.

Another important idea originated in non-equilibrium thermodynamics (thermodynamics was that studies irreversible processes). Christopher Jarzynski at the Los Alamos National Laboratory used a Markov chain to gradually convert one distribution into another via a sequence of intermediate distributions ("Nonequilibrium Equality for Free Energy Differences", 1997), an idea that Radford Neal at the University of Toronto turned into yet another annealing method ("Annealed Importance Sampling", 2001).