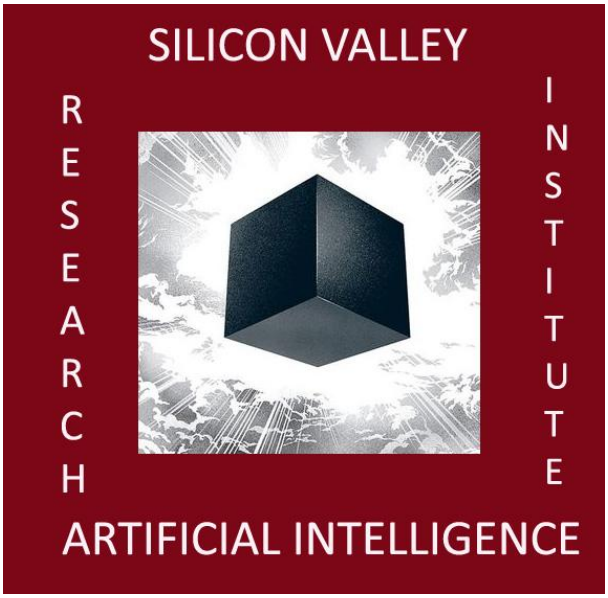

Variational Inference

Inference in probabilistic models in often intractable. The most common approximation algorithms were still based on Monte Carlo methods. An optimization-based alternative to the sampling-based Monte Carlo methods is variational inference, and the wake-sleep algorithm was a case of variational inference. In 1993 Geoffrey Hinton and Drew VanCamp at the University of Toronto had already proposed a kind of variational inference (that they called "ensemble learning") for neural networks ("Keeping Neural Networks Simple by Minimizing the Description Length of the Weights", 1993). Variational inference and MCMC do the same job (approximate inference of posterior probability) but in wildly different ways, each with pros and cons.

Variational methods convert a complex problem into a simpler problem. A practical example in physics is the study of many-body system, which is intractable until it is approximated as a one-body system. The mathematical tool is the calculus of variations that we owe to Swiss mathematicians: pioneered by the brothers Jacob and Johann Bernoulli in 1696, it was formalized as a distinct branch of calculus by Leonhard Euler in his book "Methodus Inveniendi Curvas Lineas" of 1744, and Euler also gave it a name in a 1756 lecture titled "Elementa Calculi Variationum". The calculus of variations (mainly developed by the German mathematician Karl Weierstrass in the 1860s) searches through a space of functions for the "best" function. Therefore the calculus of variations deals not with functions but with "functionals", functions of functions: the solution of the problem is not a number but a function. The calculus of variations, used for optimization purposes, is an alternative to Richard Bellman's dynamic programming. It turns out that this calculus also helps to approximate probabilistic inference, and that's what is referred to as variational inference.

After Hinton's attempts, variational inference for probabilistic models was pioneered by Michael Jordan at UC Berkeley ("Mean Field Theory for Sigmoid Belief Networks", 1996) and by Tommi Jaakkola at MIT ("Variational Methods for Inference and Estimation in Graphical Models", 1997), and was inspired by statistical physics, notably by the Italian physicist Giorgio Parisi ("Mean Field Theory for Spin Glasses", 1980).

Variational inference is about maximizing the "evidence lower bound" or ELBO. In other words, one can view problems of probabilistic inference (i.e. infer the value of something given the value of something else) as problems of optimization (find the values that minimize or maximize some function, in this case maximize the ELBO or, equivalently, minimize another quantity called the "Kullback-Leibler divergence"). Several methods have been proposed to maximize the ELBO.

Alex Graves (who had studied with Schmidhuber at IDSIA, but now at the University of Toronto) applied the "minimum description length principle" to variational inference ("Practical Variational Inference for Neural Networks", 2011). He employed the "minimum description length principle" that had already been used to train an autoencoder by Hinton in 1993 ("Keeping Neural Networks Simple by Minimizing the Description Length of the Weights", 1993). This principle dates back to Jorma Rissanen in Finland ("Modeling by Shortest Data Description", 1978) and is basically a computational version of Occam's razor in which the best model for a dataset is the one that yields to the best compression of the data. In 2012 David Blei of Princeton University and Michael Jordan of UC Berkeley (and Jordan's student John Paisley) presented an alternative to the mean-field method, based on Robbins' 1951 stochastic optimization ("Variational Bayesian Inference", 2012). Then Matt Hoffman at Adobe applied this method to analyze large libraries of texts ("Stochastic Variational Inference", 2013). In 2013 David Blei and his student Rajesh Ranganath at Princeton University developed a "black box" variational inference algorithm, one that can be quickly applied to many models with little changes ("Black Box Variational Inference", 2014), which then evolved into the "hierarchical variational models" of 2016. In general, the expression "black box" is used when one analyzes a system without looking at its internal workings, only by studying its inputs and outputs.

Bayesian models to classify data divide into "discriminative" and "generative". The former simply makes a prediction (for example, whether a sentence is German or English). The latter builds a model of the data (e.g. a model of the German and English languages) and therefore "understands" more of the data. A discriminative model learns the conditional probability ("the probability of y given x"), whereas a generative model learns the joint probability of the two events happening together. The former deals only with the existing set of data. The latter can generate missing data or compress the data. A generative classifier learns the "rule" that generates the data. Once it "understood" what is the rule, it can generate data that have not happened but could happen and belong to the same class as the ones that did happen. On the other hand, a discriminative classifier sticks to the observed data and yields a simpler rule. Discriminative models include: logistic regression, support vector machines (SVMs), nearest neighbor, conditional random fields, and traditional neural networks. Generative methods include naive Bayes, hidden Markov models, restricted Boltzmann machines, generative adversarial networks.

It is natural to assume that generative classifiers are better, but in 2001 the young Andrew Ng, when he was studying with Michael Jordan at UC Berkeley, compared linear regression and naive Bayes and concluded that discriminative classifiers are usually a better choice ("On Discriminative vs Generative classifiers", 2001).

There are three kinds of generative models (or "density modeling") in the age of neural networks: generative adversarial networks (a game between two networks), variational autoencoders (methods that maximize the ELBO) and autoregressive models.

The probabilistic interpretation of autoencoders was pioneered by LeCun's student Marc'Aurelio Ranzato at New York University ("Efficient Learning of Sparse Representations with an Energy-based Model", 2007) and by Bengio's student Pascal Vincent at the University of Montreal ("A Connection Between Score Matching and Denoising Autoencoders", 2011).

It was time to rediscover Hinton's "wake-sleep" approach. Variational autoencoders were introduced in 2013 by Max Welling and his student Diederik Kingma at the University of Amsterdam in the Netherlands ("Auto-encoding Variational Bayes", 2013). They then worked with Dutch data scientist Tim Salimans on bridging the gap between MCMC and variational inference ("Markov Chain Monte Carlo and Variational Inference", 2015). Salimans was soon hired by OpenAI to work on generative adversarial networks with Goodfellow and Radford, as well to improve the accuracy of variational autoencoders with Kingma, also hired by OpenAI ("Improving Variational Inference with Inverse Autoregressive Flow", 2016).

A more general and efficient model of variational autoencoders was introduced in 2014 by Daan Wierstra's team at DeepMind, namely Danilo Rezende and Shakir Mohamed, by fusing elements of deep learning and of probabilistic inference, and they used it to generate realistic images ("Stochastic Backpropagation and Approximate Inference in Deep Generative Models", 2014).

These two projects yielded a new class of powerful generative models, called "Deep latent Gaussian models" (DLGM).

In 2015 Daan Wierstra's team at DeepMind unveiled the Deep Recurrent Attentive Writer or DRAW, which is a neural network built around a framework of variational autoencoders ("A Recurrent Neural Network for Image Generation", 2015). DRAW extended the variational autoencoder with two techniques ("progressive refinement" and "spatial attention") that greatly improved its efficiency, thereby enabling the generation of larger and more complex images.

Systems like DRAW (that Wierstra named "sequential generative models") show that inference, generation and generalization are different aspects of the same process. In fact, the same Rezende-Mohamed team built a system capable of one-shot generalization, of generalizing a concept after just one encounter ("One-Shot Generalization in Deep Generative Models", 2016).

Kihyuk Sohn, Honglak Lee, and Xinchen Yan of NEC Laboratories in Detroit built their hybrid "conditional variational autoencoder" VAEGAN (which stands for "Variational

Autoencoder + Generative Adversarial Net") that can generate faces "conditioned" on parameters such as "young", "blonde", etc ("Learning Structured Output Representation using Deep Conditional Generative Models", 2015).

The inventions of generative adversarial networks and of variational autoencoders were the events that sparked renewed interest in probabilistic thinking.

In 2015 Brendan Frey's student Alireza Makhzani at the University of Toronto in collaboration with Goodfellow (then at OpenAI) developed "adversarial autoencoders", i.e. a combination of probabilistic autoencoder and generative adversarial networks to perform variational inference.

In 2018 Juergen Schmidhuber of IDSIA and David Ha (a former Wall Street managing director now working at Google) published a deep reinforcement learning algorithm ("World Models", 2018) that solved the "car racing" problem in which an agent has to drive a car along a racetrack as fast as possible. The solution consisted of three components: a variational autoencoder that creates a compact representation of the situation (the car relative to the environment, e.g. a bend approaching); an LSTM recurrent neural network with 256 hidden units that predicts the next situation based on the current actions (steering, accelerating and braking); and a densely connected single-layer neural network that chooses the next action, which is a combination of three actions (steering, accelerating and braking). Based on random interactions with the environment, the network builds a mental model of how the world works (i.e., its physical laws) and of how its own actions affect the state of the world. At this point the agent can learn the optimal driving strategy without actually driving. To be fair, PILCO (Probabilistic Inference for Learning Control), developed in 2011 by Marc Deisenroth at the University of Washington and Carl-Edward Rasmussen at Cambridge University achieved similar results using a simpler method: a Gaussian process turns the data from the environment into a model of the system, and then uses this model to learn to perform complex control tasks like riding a unicycle.

The third kind of generative models was mostly the work of Google's division DeepMind. Autoregression is a forecasting method in which future values of a time series are estimated solely based on a weighted sum (a linear combination) of past values of the series.

In 2011 Iain Murray of the University of Edinburgh and Hugo Larochelle of the University of Toronto built NADE, which stands for "Neural Autoregressive Distribution Estimator", and used it to generate handwritten digits. In theory, the restricted Boltzmann machine is not suitable for estimating joint probabilities but they basically converted the restricted Boltzmann machine into a Bayesian network ("The Neural Autoregressive Distribution Estimator", 2011). In 2015 NADE evolved into MADE or "Masked Autoencoder for Distribution Estimation", that added a process called "masking" (similar to the widely used method of "dropout" training).

Then, in 2014, Andriy Mnih and Karol Gregor of Wierstra's team at DeepMind designed a new kind of autoencoder, the "deep autoregressive network" or DARN, based again on Rissanen's "minimum description length principle" and they similarly used it to generate handwritten digits ("Deep Autoregressive Networks", 2014).

The next autoregressive network to come out of DeepMind was PixelRNN, developed by a team led by Koray Kavukcuoglu (another former student of LeCun at New York University). Schmidhuber's student Alex Graves at the Technical University of Munich had introduced multidimensional Long Short-Term Memory networks in 2009 ("Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks", 2009) and DeepMind designed an architecture of twelve two-dimensional LSTM layers with the novelty of convolutions applied both horizontally and diagonally ("Pixel Recurrent Neural Networks", 2016). The same team implemented PixelCNN ("Conditional Image Generation with PixelCNN Decoders", 2016) and then WaveNet (2016), based on PixelCNN, that can generate speech and music. A team at OpenAI (Tim Salimans, Diederik Kingma, Andrej Karpathy) then turned PixelCNN into an autoregressive model, PixelCNN++, that uses the previous nine generated pixels to calculate how to generate the next pixel.

Scene understanding (what is going on in a picture, which objects are represented and what are they doing) is easy for animals but hard for machines. "Vision as inverse graphics" is a way to understand a scene by attempting to generate it: what caused these objects to be there and in those positions? The program has to generate the lines and circles that constitute the scene. Once the program has discovered how to generate the scene, it can reason about it and find out what the scene is about. This approach reverse-engineers the physical process that produced the scene: computer vision is the "inverse" of computer graphics. Therefore the "vision as inverse graphics" method involves a generator of images and then a predictor of objects. The prediction is inference. This method harkens back to the Swedish statistician Ulf Grenander in the 1970s. After DRAW, DeepMind (Ali Eslami, Nicolas Heess and others) turned to scene understanding. Their AIR ("Attend-Infer-Repeat", 2016) model, which was again a combination of variational inference and deep learning, inferred objects in images by treating inference as a repetitive process, implemented as a LSTM that processed (i.e., attended to) one object at a time. Lukasz Romaszko at the University of Edinburgh later improved this idea with his Probabilistic HoughNets ("Vision-as-Inverse-Graphics", 2017), similar to the "de-rendering" used by Jiajun Wu at MIT ("Neural Scene De-rendering", 2017).

Ali Eslami and Danilo Rezende at DeepMind developed an unsupervised model to derive 3D structures from 2D images of them via probabilistic inference ("Unsupervised Learning of 3D Structure from Images", 2016). Based on that work, in June 2018 they introduced a whole new paradism: the Generative Query Network (GQN). The goal was to have a neural network learn the layout of a room after observing it from different perspectives, and then have it display the scene viewed from a novel perspective. The system was a combination of a representation network (that learns a description of the scene, counting, localizing and classifying objects) and a generation network (that produces a new description of the scene).