NeuroBayes: Convergence of Bayesian framework and deep neural networks in large-scale machine learning problems.

Dmitry Vetrov

National Research University Higher School of Economics

Abstract. In the paper we review several important directions on the combination of deep neural networks with Bayesian framework for solving large-scale machine learning problems.

Keywords. Deep learning, Bayesian framework, variational inference, stochastic optimization

First attempts to combine deep learning and Bayesian probabilistic modeling were presented in 2013-2015. Specifically, neural networks were used for approximate Bayesian inference in complex probabilistic models and Bayesian regularization helps to avoid the effect of model overfitting. These works borrow the concept of "evidence" from Bayesian statistics, then introduce a computationally tractable lower bound on the evidence and maximize the bound with respect to neural network parameters. Early results show that the neural-Bayesian fusion outperforms analogous algorithms and have better capabilities for modeling data distributions. As a result, the proposed approach may potentially allow the generation of objects that were always considered to be a product of higher nervous activity (e.g. drawing, handwriting forgery, image captioning, translation etc.). Undoubtedly, further development in this direction will be among the main trends in machine learning for the next couple of years. The most promising areas of application are following.

Bayesian regularization of deep neural networks. One of the ways to prevent machine learning algorithms from overfitting the training sample is to introduce so-called regularization which prohibits the weights of the algorithm from becoming too well-fitted. A possible approach to regularization is Bayesian regularization which imposes a prior probabilistic distribution of the weights of the algorithm; the

learning occurs during the process of Bayesian inference which combines prior restrictions with the impact of the training sample. With the increasing size of modern neural networks, the overfitting problem has also arisen in the problems of deep learning. At the end of 2015, it was shown that one of the popular heuristic procedures for preventing overfitting (so-called dropout) is a rough approximation of the Bayesian regularization of a special kind.

Bayesian regularization of neural networks theoretically makes it possible to perform so-called incremental learning; when new data is coming, the network continues learning instead of learning «from scratch». There are currently no procedures for incremental learning of neural networks. By introducing the prior distribution of the neural network weights and by performing (approximate) Bayesian inference, it is possible to obtain the posterior distribution of the weights which accumulates all the information about previously observed samples. Using this posterior distribution as a new prior distribution when new data arrives allows the network to continue learning without the need for re-using the old data.

Algorithms for building attention maps. One of the most significant results obtained in deep learning in 2015-2016 are the algorithms for building so-called attention maps which help the neural network to «concentrate» on informative fragments in the description of the data (e.g. on particular parts of an image or text). Attention maps have dramatically improved the quality of solving such complex tasks as image caption generation, machine translation etc. The methodology for building attention maps is based on models with latent variables which are a Bayesian mechanism for learning from incomplete data; this methodology is still developing today.

Search for compact representations of neural networks. One of important results in deep learning is the realization of the following fact: whenever there are no restrictions on the training data size (e.g. when training data may be generated like in the AlphaGo system which defeated the Go world champion in 2016), the more

neurons and layers the neural network has, the better will be the quality. Modern neural networks have several hundred million parameters and up to thousand layers. Their further growth is constrained by the lack of corresponding RAM sizes on modern PCs, not to mention mobile devices. On the other hand, it is becoming increasingly clear that modern neural network architectures contain much redundancy. One of the ways to eliminate such redundancy is tensor algebra and machinery of tensor decompositions. Its usage allows to compress particular parts of the neural network up to several hundred thousand times almost without any loss in quality and performance.

Stochastic optimization. Any machine learning problem may be reduced to solving a particular optimization problem, e.g., maximization of the likelihood of the training sample correct recognition. With the increasing volumes of training samples and transition to the analysis of big data, it became clear that traditional approaches for optimizing functions that arise in machine learning do not scale (i.e. become extremely ineffective and are not suitable for big data). The solution was a paradigm shift and transition to so-called stochastic optimization techniques which in some cases allowed to find an extremum of the function faster than a single evaluation of the function at a single point. All modern deep learning methods use stochastic optimization. Moreover, the rapid development of deep learning stimulated the development of more effective methods for stochastic optimization. Now several effective methods for stochastic optimization of convex functions are known. However, in the field of deep learning one has to optimize significantly multi-extremal non-convex functions.

Improving procedures for approximate Bayesian inference. The key point, which made the application of Bayesian methods and models in problems with big data (in particular, in deep learning) possible, was the development of scalable procedures for so-called variational Bayesian inference in 2014. First of all, this includes the variational auto-encoder and its numerous modifications proposed in 2015-2016. All of them use the transition from the Bayesian inference problem to

the problem of evidence lower bound optimization (ELBO), which allows to construct an approximate posterior distribution of the parameters interesting to user with the help of stochastic optimization methods. Generally speaking, ELBO is not the only possible variational lower bound. By using the Jensen inequality one can construct an infinite set of various variational lower bounds. At the same time, the more accurate the lower bound is, the better will be the quality of the approximate Bayesian inference.

Development of generative Bayesian models. One of the advantages of the Bayesian approach to the construction of data processing models is the possibility of building complex probabilistic models from the simpler ones. This is possible because the result of the Bayesian inference in one model (posterior distribution of unknown variables) may be used as a prior distribution in another model, and so on. Such complex models are known as probabilistic graphical models and are widely applicable in image/video processing, signal analysis, speech recognition, tracking, social network analysis etc. However, the potential for fitting such complex models have been quite low so far and, in practice, one has to restrict their learning to a relatively poor log-linear class. With the development of deep learning methods, it becomes possible to construct graphical models that use neural networks as building blocks.

References.

(Rezende14) Danilo Jimenez Rezende, Shakir Mohamed, Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. ICML 2014.

(Kingma13) Diederik P. Kingma, Max Welling. Auto-Encoding Variational Bayes. ICLR, 2013.

(Sohl-Dickstein14) J. Sohl-Dickstein, B. Poole, S. Ganguli. Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods. ICML, 2014.