Skolkovo Institute of Science and Technology

MSc Data Science, 2nd year, Bayesian Methods of Machine Learning

Project Proposal

THE DEEP WEIGHT PRIOR

Leonid Matyushin

MOSCOW

2019

1 Paper

In this project, we are going to reproduce experiments from the following paper [1]:

https://arxiv.org/pdf/1810.06943.pdf

Abstract

Bayesian inference is known to provide a general framework for incorporating prior knowledge or specific properties into machine learning models via carefully choosing a prior distribution. In this work, we propose a new type of prior distributions for convolutional neural networks, deep weight prior (**dwp**), that exploit generative models to encourage a specific structure of trained convolutional filters e.g., spatial correlations of weights. We define **dwp** in the form of an implicit distribution and propose a method for variational inference with such type of implicit priors. In experiments, we show that **dwp** improves the performance of Bayesian neural networks when training data are limited, and initialization of weights with samples from **dwp** accelerates training of conventional convolutional neural networks.

2 Background and Related Work

Bayesian inference allows us to use prior knowledge during a machine learning model fitting. Often it allows to transforms a prior knowledge about parameters to a posterior distribution over parameters. Stochastic variational inference [2], [3] often allows us to obtain an approximation of Bayesian inference. On practice, variational inference leads to the variational lower bound maximization:

$$\mathcal{L}(\theta) = \mathbb{E}_{w \sim q_\theta(w)} \log p(D|w) - D_{KL}(q_\theta(w)||p(w))$$

Where

- w – network parameters
- D – data
- q_θ – posterior distribution approximation (often it is a finite-dimensional family. For example, multivariate normal distribution with independent components)
- p – prior distribution
- $\log p(D|w)$ – log-likelihood

In paper [4] proposed a technique of stochastic optimization of such expressions. The same technique has been used by students in HW2 at this BMML-2019 course.

As a result of such approximation, we got a posterior distribution over weights of a deep neural network. Prior distributions over neural network weights allow us to perform different moves, that are unavailable in deterministic non-bayesian case. As an example, it is known that prior distributions play an important role in sparsification [5]. We can consider log-uniform distributions over model weights and get a sparse neural network. However, almost all such approaches are limited, since prior distributions are supposed to be fully factorised.

Authors consider following family of non-factorised over spatial dimensions distributions over CNN filters:

$$p(W) = \overset{\text{prod over layers}}{\prod_l} \quad \overset{\text{prod over input channels}}{\prod_i} \quad \overset{\text{prod over output channels}}{\prod_j} \quad p_l(w_{ijl})$$

These is so-called source kernel distributions, distributions over convolutional kernels of convolutional neural networks, that is trained on the different data. Authors suppose that we can approximate this distribution (via training on the subset of the original dataset, or, say, via training in the different dataset of the same nature), and after this effectively use this approximation for the main training. In the considered paper, the authors propose a method that estimates the source kernel distribution in an implicit form and allows us to perform variational inference with the specific type of implicit priors. More concrete, they suppose that source kernel distribution can be approximated using a small subset of problems from a specific data domain.

Authors claim that they improve the performance of Bayesian neural networks when training data are limited and accelerate the training of conventional convolutional neural networks via **dwp**.

3 Data and Experiments Overview

Besides reproducing the results on suggested data from the original paper, to explore possible ways to explore the methods and compare times and qualities for different cases, we are planning to follow authors and use several popular datasets like MNIST, Not-MNIST, CIFAR-10, and CIFAR-100. For example, the CIFAR-10 dataset contains 60,000 32x32 color images in 10 different classes (6,000 images for each class). Ten different classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks.

3.1 Classification

In this experiment, authors performed variational inference over weights of a CNN with three different prior distributions for the weights of the convolutional layers: **dwp**, standard normal, and log-uniform. Authors found that **dwp** was the most effective one, it leads to better mean test accuracy, in comparison to log-uniform and standard normal prior distributions. Surprisingly, the difference gets more significant as the training set gets smaller.

3.2 Random Feature Extraction and Convergence

In this experiment, the authors study the influence of initialization of neural networks on the training. They compare three different initialization of weights:

- learned filters (network with simple pre-trained weights)
- samples from deep weight prior
- samples form Xavier distribution [6]

After this they provide two experiment schemes:

1. Train only FC layers
2. Train whole network

In both cases learned filters overperform **dwp**, and **dwp** overperform Xavier initialization. Also, it is important that the difference in quality between learned filters and **dwp** is not very big, plus **dwp**-approach is less memory consuming (since authors train an individual prior distribution for each convolutional layer). So, probably, there exist situations, when this approach of initialization makes sense (in comparison to the just pre-trained network, which is large)

4 Scope

We are going to reproduce considered experiments and prove (or disprove) that for **dwp** we have that

- **dwp** improve the performance of Bayesian neural networks in case of limited data
- initialization of weights with samples from **dwp** accelerates training of conventional convolutional neural networks, such that training procedure becomes faster than, say, uniform initialization case, or Xavier initialization case

References

[1] Andrei Atanov, Arsenii Ashukha, Kirill Struminsky, Dmitry Vetrov, Max Welling. The Deep Weight Prior *Published as a conference paper at ICLR 2019*

[2] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference *The Journal of Machine Learning Research*

[3] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introductionto variational methods for graphical models. *Machine learning*

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes *arXiv preprintarXiv:1312.6114*

[5] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neuralnetworks *In Proceedings of the 34th International Conference on Machine Learning*

[6] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neuralnetworks *InProceedings of the thirteenth international conference on artificial intelligence andstatistics*