Exposé BNN Uncertainty

Moritz Kniebel

June 10th, 2020

(Why?) Uncertainty

A trained (deep) Neural Network usually outputs point estimates, which are dependent on its weights. Those predicted outputs are often signed with intransparent confidence, making it bad to depend on, because the networks' performance is hard to assess.

Building on this problem, Bayesian Neural Networks use probability distributions as weight-values, making the network more focused on uncertainty. Also the output moves from a point estimate to a probability distribution. For small networks it is possible to compute the posterior distributions. But in more complex - state of the art - Neural Networks, this task turns into an intractable computation, resulting in the need for approximating the posterior distributions. Using Bayesian methods, this computation gets tractable, making it possible to assign uncertainty to the networks' weights and output:

- Laplace Approximations of weights (MacKay 1992, Laplace)
  - For more complex networks: A scalable Laplace approximation for Neural Networks (Ritter, Botev, and Barber 2018, KFAC)
- Variational Inference (Graves 2011, VI)

Why uncertainty? The networks' performance can be assessed better. Plus the training method of a network can be optimized using uncertainty. For example, if it is located in deeper parts of the network, the training method can focus on other parts at the beginning. Also the network architecture can be optimized using uncertainty, as it can show, when a layer doesn't get identified properly. If so, maybe it shouldn't be part of the network at all.

Bayesian Methods / Basic research

Laplace approximation of the weights in Neural Networks

In plain Neural Networks we end up in an optimization-problem, since we want to minimize the error function. Using Bayesian Neural Networks we move to an integration-problem, because probability distributions are assigned to the weights. Using Bayes theorem the posterior of the weights can be formulated, but is intractable. Using the Laplace's method (MacKay 1992) it can be approximated. This leads us to build a Gaussian around the mode (W_{MAP}) with a curvature that is given by the Hessian:

$$p(W|\mathcal{D}) \approx \mathcal{N}(W_{MAP}, H^{-1})$$

KFAC

Looking at state of the art Neural Networks, with millions of weight paramaters the method of (MacKay 1992) reaches its limit, because the Hessian gets enormously large. To tackle this problem, the Hessian gets approximated by using a Kronecker product of two smaller matrices. The scaled Laplace approximation ends up in a multivariate Gaussian and can be represented as:

$$\mathcal{N}(\text{vec}(W_{MAP}), (\mathcal{Q} \otimes \mathcal{H})^{-1})$$

Variational Inference

This method, again, tackles the problem of intractability of the posterior $p(w/\mathcal{D})$. The method essentially approximates the posterior by using a parameterized variate distribution $q(w|\theta)$ of the same functional form and minimizing the difference between $p(w/\mathcal{D})$ and $q(w|\theta)$, making this an optimization-problem. This is done by minimizing the Kullback-Leibler divergence between p and q, while the parameters in q are estimated.
Using this method, $q(w|\theta)$ learns a good representation of the data.

ELBO and reparametrization trick

The VI method can be further optimized using ELBO and the reparametrization trick.
The evidence lower bound (ELBO) substantially achieves the same as the KL: minimizing the difference between p and q. ELBO is maximal when p and q are the same. Using ELBO gives some advantages againts the KL, which will further discussed in the thesis.

The reparametrization trick basically makes it possible to optimize a models parameters, since we can not back-propagate through a stochastic node. By moving the parameters outside of the distribution function, a gradient can be calculated for the parameters.

Reference to other possible Bayesian methods

There are also more possible Bayesian methods that yield uncertainty estimates from Neural Networks. They will most likely not be a central part of my thesis, nevertheless its good to take note of them to show, that there are more possibilities in the field of uncertainty:

- (Lakshminarayanan, Pritzel, and Blundell 2017) are moving away from the usage of BNNs to integrate unvcertainty in NNs. Instead, they are making use of NN ensembles, which yield *'high quality predictive uncertainty estimates'*.
- (Gal and Ghahramani 2015): By developing a new theoretical framework, they use dropout NNs to model the uncertainty.
- (Blundell et al. 2015) developed a backprop-compatible algorithm, which regularizes the networks weights, by minimising the variational free energy $\mathcal{F}$.

Research Goals/ Contents of the thesis

Making use of the above methods [(MacKay 1992) and (Graves 2011)] I will try to locate the uncertainty in a more simpler network. Next up I will zoom into the network and observate a single layer in terms of uncertainty. To make this more understandable a visualization of the above will be elaborated. I will then move on to a more complex network and try to transfer the findings from the previous steps. Afterward it would be interesting to see, how the uncertainty bahaves during training (optional).

References

Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. *Weight Uncertainty in Neural Networks.* arXiv: `1505.05424 [stat.ML]`.

Gal, Yarin, and Zoubin Ghahramani. 2015. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.* arXiv: `1506.02142 [stat.ML]`.

Graves, Alex. 2011. "Practical Variational Inference for Neural Networks." In *Advances in Neural Information Processing Systems 24,* edited by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, 2348–2356. Curran Associates, Inc. `http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf`.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles." In *Advances in Neural Information Processing Systems 30,* edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 6402–6413. Curran Associates, Inc. `http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf`.

MacKay, David J. C. 1992. "A Practical Bayesian Framework for Backpropagation Networks." *Neural Computation* 4 (3): 448–472. doi:10.1162/neco.1992.4.3.448. eprint: https://doi.org/10.1162/neco.1992.4.3.448. https://doi.org/10.1162/neco.1992.4.3.448.

Ritter, Hippolyt, Aleksandar Botev, and David Barber. 2018. "A Scalable Laplace Approximation for Neural Networks." In *International Conference on Learning Representations.* https://openreview.net/forum?id=Skdvd2xAZ.