Where is the uncertainty in Bayesian Neural Networks?

Moritz Kniebel

June 12, 2020

1 Uncertainty in Neural Networks

Predictions of (deep) Neural Networks (NNs) are usually point estimates. This creates problems, as the associated uncertainties are unknown. Especially in safety-critical applications (e.g. self-driving cars) having a posterior distribution can improve decision-making.
Building on this problem, Bayesian Neural Networks describe methods that apply uncertainty to Neural Networks, by using probability distributions as weights. A probabilistic view on deep learning yields a posterior distribution of a models prediction. This gives every output and weight a bound of confidence. With that information a scientist, or system has a measurement to rely on the network for analysis or decision-making. Analyzing the uncertainty of the model can also help to improve it. The dataset on which the trained model relies on could be extended or the architecture of the model could be improved.
Using Bayes theorem the posterior of the weights can be formulated. But it contains the data-likelihood, which is analytically intractable due to size and non-linearity of NNs. Addressing this problem, there are Bayesian methods to approximate the posterior. These approximations are mostly normal distributions, commonly referenced as Gaussians. Two of the most common are the ones used in this thesis:

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

To this date, the location of the uncertainty in NNs hasn't been explored extensively. Though locating the uncertainty could have important implications and help to improve networks further. The training procedure can be improved, by focusing on more certain parts first. Also the models architecture can be improved. For example, if a weight, or even a layer doesn't get identified during training, maybe it shouldn't be part of the network at all. By knowing the uncertainty in specific parts of a network, conclusions on the training-dataset could be drawn. Maybe it's not diverse enough for the network to get high confidence?

2 Description of the main methods

2.1 Laplace approximation of the weights in Neural Networks

Training a Neural Network is an optimization-problem. The goal is to minimize the error function. This happens by adjusting the weights, to find a good representation of the dataset. Using Bayesian Neural Networks moves this to an integration-problem. To find the posterior, we need to consider all possible weight constellations. Using the Laplace approximation (MacKay 1992), the posterior can be approximated. This leads us to build a Gaussian around the mode (W_{MAP}) with a curvature that is given by the Hessian, evaluated at (W_{MAP}):

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

, where $\mathcal{N}$ is the normal distribution, W are the weights, $\mathcal{D}$ is the data, H is the Hessian and *MAP* is the maximum a posteriori.

2.1.1 Kronecker-factored Approximate Curvature (KFAC)

Looking at state of the art Neural Networks, with millions of weight parameters the method of (MacKay 1992) reaches its limit. The Hessian gets enormously large, which makes the computation infeasible. That's why (Ritter, Botev, and Barber 2018) combine the Laplace approximation with using a Kronecker product to approximate the covariance matrix. Doing this they can approximate the posterior distribution over the weights as a Gaussian. This is done by assuming independence between all network layers. The Hessian H_λ of each layer λ can be approximated with a Kronecker product of $\mathcal{Q}_\lambda$ and $\mathcal{H}_\lambda$. In this, $\mathcal{Q}_\lambda$ is the covariance of incoming activation of layer λ and $\mathcal{H}_\lambda$ is the pre-activation Hessian of layer λ. The posterior of the weights in layer λ is:

$$W_\lambda \sim \mathcal{MN}(W_\lambda^*, \bar{\mathcal{Q}}^{-1}, \bar{\mathcal{H}}^{-1})$$

, where $\mathcal{MN}$ is the matrix Gaussian distribution.

2.2 Variational Inference

This method, again, tackles the problem of intractability of the posterior $p(W|\mathcal{D})$, where W stands for the weights and $\mathcal{D}$ represents the data. The method approximates the posterior by using a variational distribution $q(W|\theta)$, parameterized with θ, of the same functional form. The approximating distribution $q(W|\theta)$ should be as close as possible to the posterior $p(w|\mathcal{D})$, making this an optimization-problem. This is done by minimizing the Kullback-Leibler divergence between $p(W|\mathcal{D})$ and $q(W|\theta)$, while the parameters in $q(W|\theta)$ are estimated.
Using this method, $q(W|\theta)$ learns a good representation of the data.

2.2.1 ELBO and reparametrization trick

The VI method can be optimized using ELBO and the reparametrization trick.

Trying to minimize the difference between $p(W|\mathcal{D})$ and $q(W|\theta)$ only using the KL depends on the intractable posterior, making the problem unsolvable. The evidence lower bound (ELBO) substantially achieves the same as the KL: minimizing the difference between $p(W|\mathcal{D})$ and $q(W|\theta)$. The ELBO also involves the posterior, but is dependent on the parameter θ, making it tractable. Hence the new objective is to maximize the ELBO.
The reparametrization trick 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.

2.3 Discussion of other possible Bayesian methods

There are also more possible Bayesian methods that yield uncertainty estimates from Neural Networks. In this thesis, only Laplace approximation and variational inference will used, since they are best suited for our purposes.

- (Lakshminarayanan, Pritzel, and Blundell 2017) are moving away from the usage of BNNs to integrate uncertainty into NNs. Instead, they are making use of NN ensembles, which yield *'high quality predictive uncertainty estimates'*. These ensembles consist of several NNs and their predictions are averaged to a joint output.

- (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 $\mathcal{F}$.

3 Research Goals/ Contents of the thesis

Main goals:

1. Making use of the above methods [(MacKay 1992) and (Graves 2011)] the uncertainty will be located in a more simpler network.

2. Next up is a more detailed look on the single layers of the network. This contains to observe them in terms of uncertainty.

3. To make this more understandable a visualization of the above will be elaborated. After that we will move on to a more complex network and try to transfer the findings from the previous steps.

Possible extensions:

- Afterward it would be interesting to see, how the uncertainty behaves during training.
- Adding a third method to receive uncertainty (e.g. KFAC)

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.