Chapter 7

Future Research

Bayesian modelling and deep learning have traditionally been regarded as fairly antipodal to each other: one pushed forward by theoreticians, while the other by practitioners. Bayesian modelling is based on the vast theory of Bayesian statistics, in which we aim to capture the processes assumed to have generated our data. This often results in interpretable models that can explain the data well, at least when we can perform inference in the models. Deep learning on the other hand is mostly driven by pragmatic developments of tractable models, and has fundamentally affected the way machine learning is used in real-world applications. But unlike Bayesian modelling, deep learning lacks a solid mathematical formalism, and many developments are very weakly mathematically justified. Deep learning's success is often explained by various metaphors which do not shed much light on the reasons models are built in certain ways.

In dropout for example the network's units are multiplied by Bernoulli random variables. This slows down training but circumvents over-fitting and improves model accuracy considerably. Such techniques have had tremendous success in deep learning and are used in almost all modern models [Srivastava et al., 2014]. But why do these work so well? In [Srivastava et al., 2014] dropout is suggested to work well following a sexual reproduction metaphor. But then why would multiplying a network's units by a Gaussian distribution $\mathcal{N}(1, 1)$ instead of Bernoulli random variables result in the same model performance?

Perhaps surprisingly, we gave a possible answer to the questions above using Bayesian statistics and variational inference. We have shown that dropout in deep NNs can be cast as a variational approximation in Bayesian neural networks. The implications of this result are far-reaching. Since many modern deep learning tools make use of some form of stochastic regularisation, this means that many modern deep learning systems perform approximate Bayesian inference, capturing the stochastic processes underlying the observed data. This link opens the vast literature of Bayesian statistics to deep learning, explaining many deep learning phenomena with a mathematically rigorous theory, and extending existing tools in a principled way. We can use variational inference in deep learning, combining deep learning tools and Bayesian models in a compositional fashion. We can even assess model uncertainty in deep learning and build interpretable deep learning tools.

There are many open leads for future research:

Deep learning can be extended in a principled way. Understanding the underlying principles leading to good models allows us to improve upon them. For example, alternative approximating distributions to the ones discussed above would translate to new stochastic regularisation techniques. These can range from simple distributions to complex ones. Model compression can be achieved by forcing weights to take values from a discrete distribution over a continuous base measure of "hyper-weights" for example.

Deep learning uncertainty. Initial research above assessed the performance of dropout in terms of the predictive mean and variance. Even though the Bernoulli approximating distribution is a crude one, the model outperformed its equivalents in the field. But different non-linearity–regularisation combinations correspond to different Gaussian process covariance functions, and these have different characteristics in terms of the predictive uncertainty. Understanding the behaviour of different model structures and the resulting predictive mean and variance are of crucial importance to practitioners making use of dropout's uncertainty.

Deep learning can make use of Bayesian models. A much more interesting application of the theory above is the combination of techniques from the two fields: deep learning and Bayesian modelling. Bayesian models, often used in data analysis, strive to describe data in an interpretable way—a property that most deep learning models lack. Using the theory above we can combine deep learning with interpretable Bayesian models and build hybrid models that draw from the best that both worlds have to offer. For example, in the fields of computational linguistics and language processing we often look for models that can explain the linguistic phenomena underlying our data. Current deep learning methods work well modelling the data and have improved considerably on previous research—partly due to their tractability and ability to go beyond the bag-of-words assumptions. But the models are extremely opaque and have not been able to explain the linguistic principles they use. Interleaving Bayesian models with deep ones we could answer many of these open problems.

Bayesian models can make use of deep learning. The field of Bayesian modelling can benefit immensely from the simple data representations obtained from deep learning models. Sequence data, image data, high dimensional data—these are structures that traditional Bayesian techniques find difficult to handle. Many unjustified simplifying assumptions are often used with these data structures: bag-of-words assumptions, reducing the dimensionality of the data, etc. By interpreting deep learning models as Bayesian ones, we can combine these easily and in a principled way. Further, models can be built in a compositional fashion by propagating derivatives, forming small building blocks that can be assembled together by non-experts.

Unsupervised deep learning. One last problem discussed here is the design of unsupervised models. Bayesian statistics lends itself naturally to data analysis and unsupervised data modelling. With the Bayesian interpretation of modern deep learning new horizons open and new tools become available to solve this laborious task.

It is my hope that the framework presented in this thesis will lay the foundations of a new and exciting field of study, combining modern deep learning and Bayesian techniques in a principled and practical way.