Bachelor's Thesis

Survey on Continual Learning

Department of Statistics
Ludwig-Maximilians-Universität München

Jörg Schantz

Munich, January 17$^{\text{th}}$, 2025

Submitted in partial fulfillment of the requirements for the degree of B. Sc.
Supervised by Dr. Julian Rodemann

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Bach and Maloof (n.d.)

Contents

1 Introduction

Bli bla bulb

2 Framework

We are interested in optimizing the parameters theta of a single neural network to perform well across multiple tasks $D_1, ..., D_T$, specifically finding a MAP estimate $\theta^* = \arg\max_\theta p(\theta|D_1, ..., D_T)$. However, the datasets arrive sequentially and we can only train on one of them at a time. In the following, we first discuss how Bayesian online learning solves this problem and introduce an approximate procedure for neural networks. We then review recent Kronecker factored approxima- tions to the curvature of neural networks and how to use them to obtain a better fit to the posterior. Finally, we introduce a hyperparameter that acts as a regularizer on the approximation to the posterior.
Bayesian online learning [31], or Assumed Density Filtering [25], is a framework for updating an approximate posterior when data arrive sequentially. Using Bayes' rule we would like to simply incorporate the most recent dataset D into the posterior as:

$$E = mc^2 \tag{1}$$

where we use the posterior D from the previously observed tasks as the prior over the parameters for the most recent task. As the posterior given the previous datasets is typically intractable, Bayesian online learning formulates a parametric approximate posterior q with parameters pi, which it iteratively updates in two steps: Update step In the update step, the approximate posterior q with parameters pi from the previous task is used as a prior to find the new posterior given the most recent data:

$$E = mc^2 \tag{2}$$

Projection step The projection step finds the distribution within the parametric family of the approximation that most closely resembles this posterior, i.e. sets pi such that:

$$E = mc^2 \tag{3}$$

Opper and Winther [31] suggest minimizing the KL-divergence between the approximate and the true posterior, however this is mostly appropriate for models where the update-step posterior and a solution to the KL-divergence are available in closed form. In the following, we therefore propose using a Laplace approximation to make Bayesian online learning tractable for neural networks:

2.1 Metrics

metrics are nice

2.2 Bayes

bayes is even nices

3 Conclusion

Blub bla bli

A Appendix

See all the extra material here.

B Electronic appendix

Data, code and figures are provided in electronic form.

References

Bach, S. H. and Maloof, M. A. (n.d.). *A Bayesian Approach to Concept Drift*, pp. 127–135.

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, January 17th, 2025

Name