**MPhil project proposal** **Marton Havasi**
Supervisor: José Miguel Hernández-Lobato Due Date: 17/03/16

# Designing neural network hardware accelerators with deep Gaussian processes

## Project description

Bayesian optimization methods are often employed in a hardware design setting where simulation of a given hardware configuration can be computationally very expensive. The problem gets more difficult when there are multiple objective functions to balance against each other, for instance, the power consumption and the accuracy of the hardware accelerator. The set of optimal configurations that are not outperformed in both objective functions is called the Pareto front. This project attempts to use deep Gaussian process models in order to give better prediction of the Pareto front of multiobjective optimization problems. The state-of-the-art algorithms use Gaussian processes to model each objective function. This approach can be inefficient because it is unable to capture the correlation between the different objectives. Deep Gaussian processes can improve on this by sharing multiple layers of Gaussian processes between the objective functions. By capturing these dependencies, we can increase the accuracy of the model and therefore increase the accuracy of the resulting Pareto front.

The project will have three major parts:

- Implementing deep Gaussian processes for multiple outputs. We have code available for deep Gaussian processes, however, this code needs to be extended to be able to cope with multiple outputs that share the hidden layers. This will be used to model the objective function. [1]

- Using the previously implemented model in Bayesian optimization. For this, we need to use a multiobjective acquisition function. Our current candidate is SMSego [2] which is a technique that was successfully employed with Gaussian processes. We aim to extend it with our deep Gaussian model.

  At this point, we can measure the performance of our model on toy problems. For example, we can generate test data by sampling objective functions from the deep Gaussian process model. We can use the traditional approach as the baseline for the experiments.

- Putting the deep Gaussian process model to the test in a real-world optimization problem. We will be using it to design hardware accelerators for neural networks. The simulation of hardware configurations is very expensive and therefore they are ideal candidates for optimization methods. We aim to accurately predict the Pareto front of the configurations. The performance of deep Gaussian processes can be compared to the performance of the already existing approaches. [3] [4]

## Success criteria

- Deep Gaussian processes for multiple objectives show improvement over the approach where the objectives are modeled independently on the previously described toy problems.

- The performance of deep Gaussian processes in the hardware design problem is on-par with the currently existing approaches.

## Extensions

- Extend the design of neural network hardware accelerators to larger datasets. Currently, most of the work is being done on the MNIST dataset because of its small size. However, there have been recent developments in the simulation software so it is within reach to test on larger datasets. We want to consider hardware accelerators for the ImageNet dataset.
- Printing the hardware design. If we manage to find a high-performing configuration, it can be realized for further testing.

## Workplan

- March 20. - April 30.

  **Focus:** Coursework and exams. Preliminary reading.

- May 1. - May 14.

  **Focus:** Literature review on deep Gaussian processes.

  **Deliverable:** Implementation for multiobjective deep Gaussian processes.

- May 15. - May 28.

  **Focus:** Literature review on multiobjective acquisition functions. Extending the implementation of SMSego with deep Gaussian models.

  **Deliverable:** Implementation of multiobjective Bayesian optimization methods with deep Gaussian models.

- May 29. - June 11.

  **Focus:** Evaluation of deep Gaussian models on toy problems.

  **Deliverable:** Report on the performance on toy problems.

- June 12. - June 25.

  **Focus:** Applying deep Gaussian models to the hardware design problem.

  **Deliverable:** Implementation of the deep Gaussian models to the hardware design problem.

- June 26. - July 9.

  **Focus:** Evaluating the performance of deep Gaussian models on the hardware design problem.

  **Deliverable:** Report on the performance of the deep Gaussian models on the hardware design problem.

- July 10. - July 23.

  **Focus:** This period is left for extensions if the project is progressing well. Otherwise this period will be used to catch up to schedule.

- July 24. - August 11.

  **Focus:** Finishing the write-up.

  **Deliverable:** Thesis.

## References


[1] Bui T. D., Hernández-Lobato J. M., Li Y., Hernández-Lobato D. and Turner R. E. Deep *Gaussian Processes for Regression using Approximate Expectation Propagation*, In ICML, 2016.

[2] W. Ponweiser, T. Wagner, D. Biermann, and M. Vincze, *Multiobjective optimization on a limited budget of evaluations using model-assisted s-metric selection* in Proc. Parallel Problem Solving from Nature (PPSN X) , Dortmund, Germany, Sept. 2008, pp. 784794.

[3] Hernández-Lobato J. M., Gelbart M. A., Reagen B., Adolf R., Hernández-Lobato D., Whatmough P., Brooks D., Wei G.-Y. and Adams R. P. *Designing Neural Network Hardware Accelerators with Decoupled Objective Evaluations*, In NIPS Workshop on Bayesian Optimization, Barcelona, Spain, 2016.

[4] Reagen B., Whatmough P. Adolf R., Rama S., Lee H., Lee S., Hernández-Lobato J. M., Wei G. Y. and Brooks D. *Minerva: Enabling Low-Power, High-Accuracy Deep Neural Network Accelerators*, In ISCA, 2016.