Paper Review

Deep Kernel Learning

Federico D'Agostino

ucabfd0@ucl.com

University College London

May 10, 2022

Gaussian Processes (GPs) are highly versatile probabilistic machine learning models whose desirable properties make them the gold-standard choice in uncertainty estimation problems. GPs are flexible function approximators, relatively simple to implement, fully probabilistic and readily interpretable in the covariance matrix (Rasmussen, 2004). However, a long-standing drawback of GPs is their $\mathcal{O}(n^3)$ computational demand at inference time, along with an $\mathcal{O}(n^2)$ memory requirement to respectively invert and store the full kernel matrix. Due to these limitations, "vanilla" GPs have been limited to implementation scenarios with at most a few thousand training points (Quiñonero-Candela and Rasmussen, 2005). Furthermore, the expressivity and performance of these models is directly linked - and in some way limited to - the choice of covariance kernel, which makes it pivotal for the practitioner to select a kernel that can best learn hidden representations in the data.

Naturally, all contemporary research on GPs has focused on accommodating both their computational limitations and expressivity issues. In this review we will focus on a paper by Wilson and colleagues (2016) which tries to address both at the same time.

1 Overview

Wilson and colleagues (2016) propose a new framework, named Deep Kernel Learning (DKL), which aims to combine the structural properties, inductive biases, and expressiveness of Deep Learning (DL) models with the nonparametric flexibility of kernel methods. Specifically, given any non-linear mapping induced by a neural network $g(\mathbf{x}, \mathbf{w})$ parametrised by $\mathbf{w}$ and any kernel $k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta})$ parametrised by $\boldsymbol{\theta}$ a "deep kernel" is defined by the authors as:

$$k\left(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\theta}\right) \rightarrow k\left(g\left(\mathbf{x}_i, \mathbf{w}\right), g\left(\mathbf{x}_j, \mathbf{w}\right) \mid \boldsymbol{\theta}, \mathbf{w}\right) \tag{1}$$

and is then used as the covariance function of a Gaussian Process. The parameters $\mathbf{w}$ of the neural network g are influenced by the specific architecture chosen and are extracted after pretraining. Interestingly, however, all kernel hyperparameters $\boldsymbol{\gamma} = \{\mathbf{w}, \boldsymbol{\theta}\}$, including the neural network ones, are then jointly learned in closed form during the log marginal likelihood maximisation step of Gaussian Process training.
The authors directly rely on "Structured Kernel Interpolation" (SKI; Wilson and Nickisch, 2015) to make this step computationally tractable and scalable to big-data regimes. SKI is a framework that generalises popular inducing point methods (Quiñonero-Candela and Rasmussen, 2005) to approximate the kernel matrix of Gaussian Processes such that Kronecker and Toeplitz algebraic properties can be applied to kernel interpolation without the requirement of grid data. In practice, this means that the kernel matrix $K_{\boldsymbol{\gamma}}$ is replaced by an approximation:

$$K_{\boldsymbol{\gamma}} \approx M K_{U,U}^{\text{deep}} M^{\top} := K_{\text{KISS}} \tag{2}$$

Where M is a sparse matrix of interpolation weights and $K_{U,U}^{\text{deep}}$ is a kernel matrix evaluated over a set of $U = \{u_i\}_{i=1\ldots m}$ inducing points placed on a regular multidimensional lattice so that $K_{U,U}$ can be decomposed to a Kroenecker product of Toeplitz matrices. As a result, solving $K_{\text{KISS}}^{-1}\mathbf{y}$ at inference time will be *linear* in n using linear conjugate gradients.

On the one hand, it is clear how this work aligns with and benefits from extensive research on fast kernel learning and approximation. This includes both traditional research on inducing points methods for GPs, as well as more recent work on fast kernel expansion approximations (Le et al., 2013; Yang et al., 2015).

On the other hand, DKL is somehow also reminiscent of another fruitful strand of research on GPs, that is Deep Gaussian Processes (Damianou and Lawrence, 2013). Yet, importantly, although the two approaches both aim to extend the expressivity and potential of GPs, they do so in markedly different ways. Whereas Deep GPs consist of hierarchically stacked Gaussian Processes trained by approximate variational marginalisation, only borrowing the idea of stacked, hierarchical layers from DL, Deep Kernel Learning involves explicitly incorporating DL-derived representations into the covariance function of GPs, which can be intuitively interpreted as applying a GP to the final hidden layer of a deep network.

2 Strengths and Weaknesses

We can briefly summarise the salient strong points of DKL as follows:

- Conceptually, there is noteworthy novelty and creativity in the way different approaches have been combined: spectral mixture kernels (Wilson and Adams, 2013), inducing point approaches (Quiñonero-Candela and Rasmussen, 2005; particularly the SKI framework: Wilson and Nickisch, 2015) and deep architectures. Only the careful composition of these elements allows the method to work in the form it is introduced.

- All kernel hyperparameters $\boldsymbol{\gamma} = \{\mathbf{w}, \boldsymbol{\theta}\}$ are learned jointly. Specifically, the loss $\mathcal{L}$ is the marginal likelihood of the gaussian process:

$$\mathcal{L} = \log p(\mathbf{y} \mid \boldsymbol{\gamma}, X) \propto -\left[\mathbf{y}^\top \left(K_{\boldsymbol{\gamma}} + \sigma^2 I\right)^{-1} \mathbf{y} + \log \left|K_{\boldsymbol{\gamma}} + \sigma^2 I\right|\right] \tag{3}$$

  on which backpropagation is run directly. This is intriguing given how Marginal Likelihood maximisation in GPs is known to implicitly regularise the training objective (though we will comment on this in a moment; Rasmussen and Williams, 2005, Ch. 5).

- Approximation of the Kernel matrix and subsequent evaluations at test time are impressively fast. The performance here is entirely inherited from the SKI framework, precisely the "KISS-GP" approximation (Wilson and Nickisch, 2015). Substituting K_{KISS} for $K_{\boldsymbol{\gamma}}$ in equation 3 results in $\mathcal{O}(n)$ training and $\mathcal{O}(1)$ testing, and a total run time for the model in the experiments which is comparable to non-Bayesian Deep Neural Networks (DNN).

- All experiments show equal or better performance for DKL over GPs and DNNs, demonstrating the model's potential even on datasets with over 500k training points.

Despite the many strong points, some apparent weaknesses can also be recognized in the paper:

- Even though GP training through the marginal likelihood is known to conveniently decompose the error into fit and complexity terms, it is not guaranteed that this protects from overfitting as the number of kernel parameters grow, as it has been recognised elsewhere (Rasmussen and Williams, 2005; Wilson and Adams, 2013). Since here numerous

neural network parameters are added in the GP training process, the way the authors gloss over this is unjustified. Experiments on the overfitting potential of DKL are needed.

- DKL is only introduced for regression problems. It is left unclear how the model could be generalised with non-Gaussian likelihoods. While the authors were able to avoid approximate Bayesian inference in the regression case, that is not possible for classification.
- Relative to the previous point, there is also no discussion of possible DKL implementations outside SKI, neither in terms of computational demands nor of overall tractability.
- Finally, even though the initial intent of the authors was to unify the non-parametric flexibility of GPs with DL, it is clear that the final model can hardly be called non-parametric considering that the need to come up with "good" neural network architectures for the "deep" part of the kernel is not eliminated.

3 Recommendation

All things considered, our recommendation is to accept the paper for publication. The idea of combining non-linear transformations from deep learning models into the kernel of GPs for joint hyperparameter optimisation is no doubt inventive and fruitful. Implementations specifics aside, the idea has remarkable potential to drive future research. Therefore, even if some weaknesses and unexplored questions can be recognised in the approach, these can reasonably be the subject of follow-up papers that scrutinize this new framework more attentively.
We are confident in this judgement.

4 Further feedback

A couple of additional points may be added to the publication to help the reader.
First of all, given how heavily the model in the present exposition relies on SKI, some more details on the method could be added where relevant. Without the context given by the original paper (Wilson and Nickisch, 2015), the model's performance in terms of computational complexity might strike as being almost miraculous.
Secondly, it would be informative to see how DKL compares to Deep Gaussian Processes (DGP), given the overlap between the two method's rationale, and the fact that DGPs have been recently made scalable to big data (Hensman and Lawrence, 2014).
This could be included in the experiments section.

5 Follow ups

This newly introduced framework opens many directions for future research. The most natural route is undoubtedly extending the model to classification problems. It is unavoidable that the authors will need to adopt some form of variational inference for this to happen. Fortunately, recent work shows promising applications of stochastic variational inference to GPs on big data (Hensman et al., 2013), even in classification settings (Hensman et al., 2015).
We are sure the authors are already eagerly working on this.

References

Damianou, A., & Lawrence, N. D. (2013). Deep Gaussian Processes [ISSN: 1938-7228]. *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, 207–215.

Hensman, J., Fusi, N., & Lawrence, N. D. (2013). Gaussian processes for Big data. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 282–290.

Hensman, J., & Lawrence, N. D. (2014). Nested Variational Compression in Deep Gaussian Processes [arXiv: 1412.1370]. *arXiv:1412.1370 [stat]*.

Hensman, J., Matthews, A., & Ghahramani, Z. (2015). Scalable Variational Gaussian Process Classification [ISSN: 1938-7228]. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 351–360.

Le, Q., Sarlos, T., & Smola, A. (2013). Fastfood - Approximating Kernel Expansions in Loglinear Time. *30th International Conference on Machine Learning (ICML)*.

Quiñonero-Candela, J., & Rasmussen, C. E. (2005). A Unifying View of Sparse Approximate Gaussian Process Regression. *Journal of Machine Learning Research*, *6*(65), 1939–1959.

Rasmussen, C. E. (2004). Gaussian Processes in Machine Learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures* (pp. 63–71). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_4

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning* (F. Bach, Ed.). MIT Press.

Wilson, A., & Adams, R. (2013). Gaussian Process Kernels for Pattern Discovery and Extrapolation [ISSN: 1938-7228]. *Proceedings of the 30th International Conference on Machine Learning*, 1067–1075.

Wilson, A., Hu, Z., Salakhutdinov, R., & Xing, E. P. (2016). Deep Kernel Learning [ISSN: 1938-7228]. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 370–378.

Wilson, A., & Nickisch, H. (2015). Kernel interpolation for scalable structured Gaussian processes (KISS-GP). *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, 1775–1784.

Yang, Z., Wilson, A., Smola, A., & Song, L. (2015). A la Carte – Learning Fast Kernels [ISSN: 1938-7228]. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 1098–1106.