Project Proposal

Investigating Bayesian Inference in Continual Learning

M. E. Tong, supervised by S. Farquhar and Y. Gal

For real-world problems, it is vital that intelligent agents have the capacity to learn and remember a variety of tasks [Kirkpatrick et al., 2017]. Continual learning refers to online multi-task learning where tasks are learned sequentially, with datasets discarded after training. It is of particular importance for applications for which retaining old datasets is unethical, undesireable, illegal or imprudent [Farquhar & Gal, 2019, Farquhar & Gal, 2018].

The major problem in the field is that of catastrophic forgetting. New learning often causes neural networks to rapidly forget old learning [McCloskey & Cohen, 1989]. The challenge is to balance learning new tasks whilst remembering previous ones. Though promising results have been published in the field of continual learning, clear shortcomings have been identified in many of the current approaches and evaluations, such as a dependency on re-training on previous datasets or tasks [Farquhar & Gal, 2018].

The most recent advances in continual learning have favoured a prior-focused approach, such as variational continual learning [Nguyen et al., 2017], synaptic intelligence [Zenke et al., 2017], elastic weight consolidation [Kirkpatrick et al., 2017], Riemannian walk [Chaudhry et al., 2018], Kronecker factored online Laplace approximation [Ritter et al., 2018]. These prior-focused approaches use the posterior or other parameters from previous tasks as priors for new tasks. Likelihood-based approaches, such as Deep Generative Replay [Shin et al., 2017], are based on pseudo-rehearsal [Robins, 1995], simulating previous datasets in order to estimate their log-likelihood according to the new model. There have also been approaches based on dynamic architectures.

We suggest that further investigation into the variational continual learning (VCL) approach [Nguyen et al., 2017] is warranted. Currently, good performance is dependent on re-training on small coresets retained from previous datasets [Farquhar & Gal, 2018], but improvements to the approach may eliminate the need for these coresets, which would better reflect true continual learning.

VCL demonstrates a prior-focused Bayesian approach to continual learning, where the posterior for the previous tasks is used as the prior when training on the new task dataset. This is an intuitive way of allowing the previous tasks to strongly influence prediction, whilst allowing the parameters to adapt to the new task.

$$\underbrace{p(\boldsymbol{\theta}|\mathcal{D}_{1:t})}_{\text{new posterior}} \propto \underbrace{p(\boldsymbol{\theta}|\mathcal{D}_{1:t-1})}_{\substack{\text{previous posterior} \\ \text{(new prior)}}} \underbrace{p(\mathcal{D}_t|\boldsymbol{\theta})}_{\text{likelihood}} \tag{1}$$

However, the true posterior $p(\boldsymbol{\theta}|\mathcal{D}_{1:t})$ is computationally intractable, so we must approximate it. We suggest that improvements to this approximation may be key to performance.

The posterior approximation in VCL is performed using Kullback-Leibler (KL) minimisation, a variational method, over a model family of possible posteriors $\mathcal{Q}$, to yield a tractable normalised approximation $q_t(\boldsymbol{\theta})$ to the true posterior $p(\boldsymbol{\theta}|\mathcal{D}_{1:t})$. We define $q_0(\boldsymbol{\theta})$ to be the prior, $p(\boldsymbol{\theta})$.

$$\underbrace{p(\boldsymbol{\theta}|\mathcal{D}_{1:t})}_{\substack{\text{new true}\\ \text{posterior}}} \approx \underbrace{q_t(\boldsymbol{\theta})}_{\substack{\text{new posterior}\\ \text{approximation}}} = \underset{q\in\mathcal{Q}}{\operatorname{argmin}}\, \mathrm{KL}\left(q(\boldsymbol{\theta}) \,||\, \frac{1}{Z_t}\underbrace{q_{t-1}(\boldsymbol{\theta})}_{\substack{\text{previous posterior}\\ \text{approximation}}} p(\mathcal{D}_t|\boldsymbol{\theta})\right) \tag{2}$$

Here, $\boldsymbol{\theta}$ are the parameters, $\mathcal{D}_i$ is the i^{th} dataset, $q_i(\boldsymbol{\theta})$ is the posterior approximation following the i^{th} dataset, Z_t is the intractable normalisation constant which is irrelevant for minimisation.

Crucially, this is exact Bayesian inference, that is, $q_t(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathcal{D}_{1:t})\ \forall t$, if two criteria are met at every step:

1. The true posterior is a member of the model family $\mathcal{Q}$.
2. The optimisation achieves the minimum KL divergence.

Since the posterior approximation occurs after every task is learned, improving this approximation is important to prevent error propagation across different tasks. We therefore suggest that investigation into improvements to these two criteria should be the major aims of the project.

We suggest that an investigation into an improvement to the first criterion is the most promising, as we suspect that it is likely that the model family used in VCL is not sufficient to approximate the posterior well. The primary aim of the project is therefore to see if using a more adaptable model family results in better performance:

1. **Using full covariance Gaussian distributions for the model family $\mathcal{Q}$.**

   VCL uses a Gaussian mean-field approximate posterior, with diagonal covariance:

   $$q_t(\boldsymbol{\theta}) = \prod_{d=1}^{D} \mathcal{N}(\theta_{t,d}; \mu_{t,d}, \sigma_{t,d}^2)$$

   We expect that a full covariance Gaussian distribution will be able to approximate the posterior more accurately, as the covariance matrix has off-diagonal correlation elements. 'More flexible distributions ... simply give better approximations to the true posterior' [Barber & Bishop, 1998]. The minimum KL fit obtained with the full covariance model family should therefore be better than that obtained with the diagonal covariance model family.

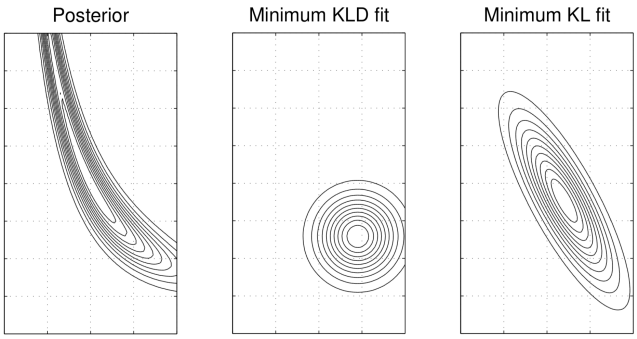

Figure 1: Comparison of posterior distribution approximations on a synthetic example with two parameters. A diagonal covariance Gaussian distribution (KLD) results in KL_{res} = 4.6, whereas a full covariance Gaussian distribution (KL) results in an improved KL_{res} = 3.9. [Barber & Bishop, 1998].

We suggest that an investigation into an improvement to the second criterion may also be promising. Approximate Bayesian inference is typically carried out with either variational or Markov Chain Monte Carlo (MCMC) inference. We also suggest looking into using a MCMC approach instead of variational approach such as KL minimisation. The secondary aim of the project is therefore to see if using MCMC results in better performance:

2. **Using a Markov Chain Monte Carlo method for posterior approximation.**

   We expect that Hamiltonian Monte Carlo (HMC) is a promising MCMC method to use for approximate inference. HMC is a principled Hamiltonian dynamics-based MCMC method which has been applied successfully to Bayesian neural networks [Neal, 1995]. It systematically and coherently traverses the state space, avoiding the slow exploration typical of random walk proposals [Neal, 2012]. This is particularly useful in high-dimensional spaces, where we must use information about the space geometry for an efficient exploration. Gradient-based algorithms such as HMC tend to be more robust and geometrically ergodic over a larger class of target distributions than non-gradient based algorithms, yielding 'stronger guarantees on the validity of the resulting estimators' [Betancourt, 2017].

Approximate Timeline

March	Write project proposal; review literature
April	Use non-continual BNN to compare diagonal and full model families
May	Continue with comparison; begin continual learning implementation
June	Continue with continual learning comparison implementation
July	Continue with continual learning comparison implementation; begin with Markov Chain Monte Carlo implementation
August	Continue with MCMC implementation
September 2	Hand-in date

Project supervisors' signatures

S. Farquhar _____

Y. Gal _____

References

[Barber & Bishop, 1998] D. Barber, C. M. Bishop, 1998. *Ensemble Learning in Bayesian Neural Networks*. Neural Networks and Machine Learning, Springer, 215-237. `https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/bishop-ensemble-nato-98.pdf`

[Betancourt, 2017] M. Betancourt, 2017. *A Conceptual Introduction to Hamiltonian Monte Carlo*. arXiv:1701.02434v2

[Chaudhry et al., 2018] A. Chaudhry, P. K. Dokania, T. Ajanthan, P. H. S. Torr, 2018. *Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence*. arXiv:1801.10112v3

[Farquhar & Gal, 2019] S. Farquhar, Y. Gal, 2019. *A Unifying Bayesian View of Continual Learning*. arXiv:1902.06494v1

[Farquhar & Gal, 2018] S. Farquhar, Y. Gal, 2018. *Towards Robust Evaluations of Continual Learning*. arXiv:1805.09733v2

[Kirkpatrick et al., 2017] J. Kirkpatrick, R. Pascanua, N. Rabinowitza, J. Venessa, G. Desjardinsa, A. A. Rusua, K. Milana, J. Quana, T. Ramalhoa, A. Grabska-Barwinskaa, D. Hassabisa, C. Clopathb, D. Kumarana, and R. Hadsella, 2017. *Overcoming catastrophic forgetting in neural networks*. arXiv:1612.00796v2

[McCloskey & Cohen, 1989] M. McCloskey, N. J. Cohen, 1989. *Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem*. Psychology of Learning and Motivation, Volume 24, 109-165.

[Neal, 2012] R. M. Neal, 2012. *MCMC using Hamiltonian dynamics*. arXiv:1206.1901v1

[Neal, 1995] R. M. Neal, 1995. *Bayesian Learning for Neural Networks*. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.446.9306&rep=rep1&type=pdf`

[Nguyen et al., 2017] C. V. Nguyen, Y. Li, T. D. Bui, R. E. Turner, 2017. *Variational Continual Learning*. arXiv:1710.10628v3

[Ritter et al., 2018] H. Ritter, A. Botev, D. Barber, 2018. *Online Structured Laplace Approximations For Overcoming Catastrophic Forgetting*. arXiv:1805.07810v1

[Robins, 1995] A. Robins. *Catastrophic forgetting, rehearsal, and pseudorehearsal*. Connection Science: Journal of Neural Computing, Artificial Intelligence and Cognitive Research, Volume 7, 123-146.

[Shin et al., 2017] H. Shin, J. K. Lee, J. Kim, J. Kim, 2017. *Continual Learning with Deep Generative Replay*. arXiv:1705.08690v3

[Zenke et al., 2017] F. Zenke, B. Poole, S. Ganguli, 2017. *Continual Learning Through Synaptic Intelligence*. arXiv:1703.04200v3