Variational Continual Learning with and without Wasserstein Distance Minimisation VI Objective

1 Introduction

Continual learning (Schlimmer and Fisher [1986]) is a paradigm for machine learning in a context where datasets arrive in a continuous, online manner ($\mathcal{D}_t = \{\mathbf{x}_t^{(n)}, y_t^{(n)}\}_{n=1}^{N_t}$ where N_t represents the number of samples that arrived at time t). The data is treated as a set of different "tasks" (i.e. subsequent datasets might require inference for a problem that is separate from the previous one, or practically speaking- will have different output classes).

Standard neural networks demonstrate the problem of "catastrophic forgetting", when, during sequential training, as the number of tasks increases, the performance of the neural network on previous tasks is eroded completely (McCloskey and Cohen [1989]). Variational Continual Learning (Nguyen et al. [2017]) is a technique to mitigate this. This paper attempts to reproduce some of its experiments and extend them in two ways. First, by modifying the training objective to be one of regression, as compared to classification. Second, using a different metric based on the Wasserstein-2 distance for variational inference. The regression based model gives performance similar to standard VCL while the alternative variational inference model gives slighly lower accuracy with a significant speed-up in training time. Code can be found here.

2 Background

Replay based methods use ideas about experience replay buffers originally developed in the reinforcement learning community (Mnih et al. [2015]) to mitigate the effects of catastrophic forgetting. The general framework involves choosing a subset of examples from previous tasks to inform the training process during later tasks. Optimisation based techniques (Lopez-Paz and Ranzato [2017]) used previous examples to find suitable directions to project gradients onto. Another direction of this research involves finding an optimal set of previous examples to add to the replay buffer.

Architecture based techniques attempt to design neural networks in a manner that enforces parameters to adapt to previous and current tasks. Recent examples include the work by Serra et al. [2018] which involves learning a per-task attention mask for weights that is then used to scale the gradient updates of each neuron so that neurons that are crucial for previous tasks are not affected significantly by gradient updates. VCL itself uses a set of "shared" and separate (task specific) weights.

The main emphasis of VCL and the proposed extensions is on regularisation. In the general framework, model parameters for current tasks (θ) are learnt with a regularisation penalty that ensures that they do not stray too far from the weights for previous tasks (θ_{t-1}). This involves minimising the following objective:

$$\mathcal{L}^t(\theta) = \sum_{n=1}^{N_t} \log p(y_t^{(n)}|\theta, \mathbf{x}_t^{(n)}) - \frac{1}{2}\lambda_t(\theta - \theta_{t-1})^T \Sigma_{t-1}^{-1}(\theta - \theta_{t-1}) \tag{1}$$

A challenging problem here is the choice Σ. Some popular choices include a Laplace approximation (Smola et al. [2004]), Fisher information matrix (Kirkpatrick et al. [2017]) (EWC), and other heuristics based on how much certain weights affect gradients (Zenke et al. [2017]):

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Post VCL, there have been some augmentations to the original framework. Loo et al. [2020] generalise the idea of VCL to show how it can be modified to recover EWC with a flexible regularisation term for the KL divergence and also propose adding FiLM layers (Perez et al. [2018]) to improve performance.

More recently, Melo et al. [2024] connect continual learning to literature from the reinforcement learning community and derive an n-step VCL framework that is able to focus on more recent updates for regularisation and prevent compounding of errors. To the same effect they also propose a version of the loss function based on the TD(λ) (Sutton et al. [1998]) target.

3 Methodologies

The VCL paper (Nguyen et al. [2017]) first identifies a Bayesian recursion for learning weight posteriors after seeing T tasks:

$$p(\theta|\mathcal{D}_{1:T}) \propto p(\theta)p(\mathcal{D}_{1:T}|\theta) = p(\theta)p(\mathcal{D}_{1:T-1}|\theta)p(\mathcal{D}_T|\theta) \propto p(\theta|\mathcal{D}_{1:T-1})p(\mathcal{D}_T|\theta) \tag{2}$$

At each task T the previous posterior is approximated using a tractable (Gaussian) approximation $q_{T-1}(\theta)$ and the KL divergence between the current and (approximate) previous posterior is minimised yielding a simple Bayesian update framework:

$$\min_{q\in Q} \mathrm{KL}(q(\theta), \frac{1}{Z_t} q_{t-1}(\theta)p(\mathcal{D}_t|\theta)) \Rightarrow \min_{q\in Q} \sum_{n=1}^{N_t} -\mathbb{E}_q \left[\log p(y_t^{(n)}|\theta, \mathbf{x}_t^{(n)})\right] + \mathrm{KL}(q(\theta), q_{t-1}(\theta)) \tag{3}$$

The expectation of the likelihood is computed using Monte-Carlo methods and coresets (replay buffers) are used for retraining at inference time on small sets of samples from previous training tasks. As far as the architecture goes, the model utilises a shared backbone coupled with task specific classification heads.

4 Novel Extension

4.1 Regression

Instead of treating target vectors as class labels, I treat target vectors as continuous vectors in $\mathbb{R}^{c_t}$ where c_t represents the number of classes in task t. This changes the likelihood term in the minimisation objective defined above. Since the model outputs vectors in $\mathbb{R}^{c_t}$:

$$\mathbf{y_t} = f^\theta(\mathbf{x}_t) + \epsilon \quad \text{WHERE} \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma) \tag{4}$$

Based on this, $-\log p(\mathbf{y_t}|\theta, \mathbf{x_t}) \propto \left(\mathbf{y}_t - f^\theta(\mathbf{x}_t)\right)^T \Sigma^{-1} \left(\mathbf{y}_t - f^\theta(\mathbf{x}_t)\right)$, Since there is no information about class variances and the assumption that Σ is a diagonal matrix with identical values allows me to use the same setup as before, I use this assumption and $-\log p(\mathbf{y_t}|\theta, \mathbf{x_t}) \propto ||\mathbf{y}_t - f^\theta(\mathbf{x}_t)||_2^2$, yielding the MSE loss as a valid negative log likelihood function (I found this to give better results experimentally as well).

4.2 Modified VI

The KL divergence which is not a true distance metric in the sense that it does not respect the triangle inequality, is anti-symmetric, and most importantly for this work, is difficult to optimise. The anti-symmetry is problematic because this means that reversing the order of tasks changes our loss landscape which shouldn't necessarily matter since tasks can be completely independent of each other. Finally, it is difficult to optimise the KL term practically since gradients can grow very quickly (e.g. small terms in the denominator of the closed form of the KL divergence for two Gaussians).

In optimal transport, the optimal cost between two distributions $q(\theta), q_{t-1}(\theta')$ is given as follows:

$$\inf_{\gamma\in\Gamma(q,q_{t-1})} \mathbb{E}_{\theta,\theta'\sim\gamma}\left[c(\theta, \theta')\right] \tag{5}$$

Where c is the cost function. I define the cost function as a weighted sum of the log-likelihood of the data based on the current parameters and L2 distance between the weights:

$$c(\theta, \theta') = \sum_{n=1}^{N_t} -\log p(y_t^{(n)}|\theta, \mathbf{x}_t^{(n)}) + \lambda(\theta - \theta'_{t-1})^2$$

This objective to be minimised can be reduced to the following form (proven here 7.1):

$$\mathbb{E}_q \left[\sum_{n=1}^{N_t} -\log p(y_t^{(n)} | \theta, \mathbf{x}_t^{(n)}) \right] + \lambda \left((\mu - \mu_{t-1})^2 + (\sigma - \sigma_{t-1})^2 \right) \tag{6}$$

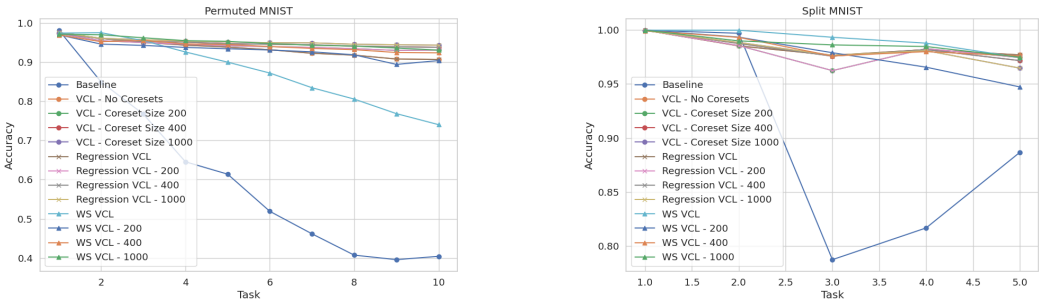
The final loss function for our model is:

$$\mathcal{L}^t(\{\mu_i, \sigma_i\}_i) = \sum_{n=1}^{N_t} -\log p(y_t^{(n)} | \{\mu_i, \sigma_i\}_i, \mathbf{x}_t^{(n)}) - \lambda \sum_i \left(\mu_i - \mu_{i_{t-1}}\right)^2 + \left(\sigma_i - \sigma_{i_{t-1}}\right)^2 \tag{7}$$

The regularisation term is symmetric, motivated by theory and gradients do not diverge.

5 Experiments

The Permuted and Split MNIST experiments from the paper have been reproduced. As a baseline, a Bayesian Network with the same number of weights and trained with the same hyperparameters as the network for comparison on the same task is used. RMSE plots are here 7.2.



(a) Permuted MNIST (b) Split MNIST

Figure 1: Classification/Regression Experiments on Permuted and Split MNIST

(a) Permuted MNIST

Method	Epochs	LR	λ	Accuracy
VCL	100	1e-3	2e-2	**0.943**
Regression	20	3e-4	1e-4	**0.943**
WS	5	1e-3	1.0	0.931

(b) Experiment B

Method	Epochs	LR	λ	Accuracy
VCL	25	3e-4	1e-1	0.965
Regression	25	3e-4	1.0	0.965
WS	5	1e-3	1.0	**0.975**

Table 1: Experiment Hyperparameters and Accuracies (Coreset Size = 1000)

6 Analysis and Conclusions

The results of the experiment with the modified variational inference term are quite interesting. While the accuracy is not as high as those for VCL (especially without the use of coresets), the performance of this model is based on only 5 epochs of training as compared to the other's 100. This is a huge saving in training time, both when retraining the model initially and while training on coresets - probably a useful application where models have to be retrained on coresets on the go.

Regression targets attain almost exactly the same accuracy as the model trained on discrete class based targets. Intuitively, this makes sense: the regularisation penalty does not change and both, the softmax and MSE based NLL maximise the likelihood. The regression model also took much fewer epochs to train(20) although still more than the other proposed method.

To conclude, variational continual learning is an extremely useful methodology for mitigating catastrophic forgetting. However, when an easier optimisation target is needed and the priors are set to be Gaussians, a moment matching approach inspired by the Wasserstein loss proves itself to be far more computationally feasible.

7 Supplementary Material

7.1 Proof of 6

In my setup, the marginals of the joint distribution γ over each set of weights ($\theta \sim q, \theta' \sim q_{t-1}$) is Gaussian:

$$\int_{\theta} \gamma(\theta, \theta')\, d\theta = q_{t-1}(\theta') = \mathcal{N}(\theta'; \mu_{t-1}, \sigma_{t-1}^2)$$
$$\int_{\theta}^{'} \gamma(\theta, \theta')\, d\theta' = q(\theta) = \mathcal{N}(\theta; \mu, \sigma^2)$$

$$\begin{aligned}
&\mathbb{E}_{\theta,\theta'\sim\gamma}\left[\sum_{n=1}^{N_t} -\log p(y_t^{(n)}|\theta, \mathbf{x}_t^{(n)}) + \lambda(\theta - \theta')^2\right] \\
&= \iint \sum_{n=1}^{N_t} -\log p(y_t^{(n)}|\theta, \mathbf{x}_t^{(n)})\gamma(\theta, \theta')\, d\theta' d\theta + \lambda \mathbb{E}_{\theta,\theta'\sim\gamma}\left[\theta^2 + \theta'^2 - 2\theta\theta'\right] \\
&= \int \sum_{n=1}^{N_t} -\log p(y_t^{(n)}|\theta, \mathbf{x}_t^{(n)}) q(\theta)\, d\theta + \lambda\left(\mathbb{E}_{\theta,\theta'\sim\gamma}[\theta^2] + \mathbb{E}_{\theta,\theta'\sim\gamma}[\theta'^2] - 2\mathbb{E}_{\theta,\theta'\sim\gamma}[\theta\theta']\right) \\
&= \mathbb{E}_q\left[\sum_{n=1}^{N_t} -\log p(y_t^{(n)}|\theta, \mathbf{x}_t^{(n)})\right] + \lambda\left((\mu^2 + \sigma^2) + (\mu_{t-1}^2 + \sigma_{t-1}^2) - 2(\rho_{\theta\theta'}\sigma\sigma_{t-1} + \mu\mu_{t-1})\right)
\end{aligned}$$

Where the last term follows from $\rho_{XY} = \frac{\mathbb{C}\text{ov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X - \mathbb{E}[\mathbb{X}])(Y - \mathbb{E}[Y])]}{\sigma_X \sigma_Y}$ for two random variables X, Y.

Finally, since both distributions are from the same family, for the optimisation objective of aligning them, I take the case when the correlation is 1, giving the following optimisation objective after completing the square:

$$\mathbb{E}_q\left[\sum_{n=1}^{N_t} -\log p(y_t^{(n)}|\theta, \mathbf{x}_t^{(n)})\right] + \lambda\left((\mu - \mu_{t-1})^2 + (\sigma - \sigma_{t-1})^2\right)$$

7.2 RMSE Plots for Regression Tasks

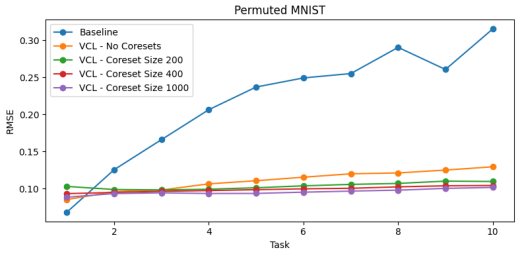


Figure 2: RMSE on Permuted MNIST

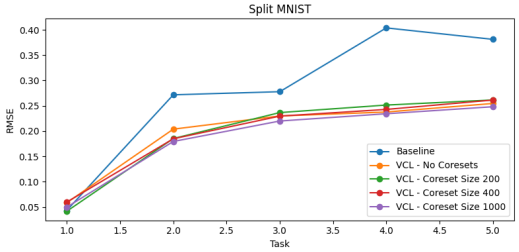


Figure 3: RMSE on Split MNIST

References

Jeffrey C Schlimmer and Douglas Fisher. A case study of incremental concept induction. In *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence*, pages 496–501, 1986.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pages 4548–4557. PMLR, 2018.

Alex J Smola, SVN Vishwanathan, and Eleazar Eskin. Laplace propagation. In *Advances in Neural Inf. Proc. Systems (NIPS)*, pages 441–448, 2004.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.

Noel Loo, Siddharth Swaroop, and Richard E Turner. Generalized variational continual learning. *arXiv preprint arXiv:2011.12328*, 2020.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Luckeciano C Melo, Alessandro Abate, and Yarin Gal. Temporal-difference variational continual learning. *arXiv preprint arXiv:2410.07812*, 2024.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.