EVCL: Elastic Variational Continual Learning with Weight Consolidation

Hunar Batra [1] **Ronald Clark** [1]

Abstract

Continual learning aims to allow models to learn new tasks without forgetting what has been learned before. This work introduces Elastic Variational Continual Learning with Weight Consolidation (EVCL), a novel hybrid model that integrates the variational posterior approximation mechanism of Variational Continual Learning (VCL) with the regularization-based parameter-protection strategy of Elastic Weight Consolidation (EWC). By combining the strengths of both methods, EVCL effectively mitigates catastrophic forgetting and enables better capture of dependencies between model parameters and task-specific data. Evaluated on five discriminative tasks, EVCL consistently outperforms existing baselines in both domain-incremental and task-incremental learning scenarios for deep discriminative models.

1. Introduction

Continual Learning focuses on training models on non-stationary data from sequential tasks, where tasks change over time or new tasks can emerge (Schlimmer & Fisher, 1986; SUTTON, 1993; Ring, 1997). The challenge is to balance between adapting to new data (plasticity) and retaining existing knowledge (stability). Given a sequence of tasks $D \in \{D_1, \ldots, D_T\}$, where each task $D_t = \{(x_i^t, y_i^t)\}_{i=1}^{N_t}$; $x_i^t \in \mathcal{X}$; $y_i^t \in \mathcal{Y}$, Continual Learning optimises a model $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by θ, and aims to infer a distribution over predictive functions that performs well on the set of all tasks D, with access to only a single task at a time. However, when trained sequentially, neural networks (NNs) often suffer from catastrophic forgetting, where knowledge of previously learned tasks is abruptly lost as the model adapts to new tasks (McCloskey & Cohen, 1989; Ratcliff, 1990; Goodfellow et al., 2013; Nguyen et al., 2019).

[1]Department of Computer Science, University of Oxford. Correspondence to: Hunar Batra <hunar.batra@cs.ox.ac.uk>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria. Copyright 2024 by the author(s).

Various approaches have been proposed to mitigate catastrophic forgetting, including inference based approaches (Nguyen et al., 2017; Swaroop et al., 2019; Kirkpatrick et al., 2017), regularisation-based methods (Kirkpatrick et al., 2017; Zenke et al., 2017; Pan et al., 2020), replay-based memory contextualisation (Rolnick et al., 2019; Shin et al., 2017), and architectural methods (Lee et al., 2017; Rusu et al., 2016; Masse et al., 2018; Loo et al., 2020a). Amongst the probabilistic inference methods, Variational Continual Learning (VCL) (Nguyen et al., 2017) and Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) are the two main approaches. VCL employs variational inference to approximate the posterior $p(\theta|D_t)$ based on a prior distribution $p(\theta)$. While, EWC penalises changes to parameters important for previous tasks by using the Fisher information matrix F_t as a proxy for the posterior distribution $p(\theta|D_t)$, constraining the model to stay close to the optimal parameters θ_t^* for prior tasks.

Despite their individual strengths, both VCL and EWC have limitations in effectively mitigating catastrophic forgetting, with VCL suffering from accumulated errors due to approximate posterior alignment (Nguyen et al., 2017) and EWC's laplace approximation local estimate leading to underestimation of the importance of certain parameters (Huszár, 2018; Kirkpatrick et al., 2017). This work introduces a novel integration of the variational inference mechanism of VCL with the EWC regularisation. Our proposed Variational Continual Learning with Weight Consolidation (EVCL) hybrid model combines the variational posterior approximation capabilities of VCL with the parameter-protection strategy of EWC, making it capture complex dependencies between model parameters and task-specific data better. Moreover, incorporating EWC's regularisation into VCL allows for direct optimization of the model parameters for each task by penalising changes to important parameters, as identified by the Fisher information matrix, mitigating the need for additional memory-intensive coresets. This work finds that EVCL effectively outperforms existing baselines on a variety of continual learning tasks for deep discriminative models. We release our code at https://github.com/hunarbatra/elastic-variational-continual-learning.

The contributions of this work are as follows:

- We propose Elastic Variational Continual Learning

(EVCL), a novel hybrid method that integrates the strengths of Variational Continual Learning (VCL) and Elastic Weight Consolidation (EWC), combining the variational posterior approximation mechanism of VCL with the regularization-based parameter-protection strategy of EWC, effectively mitigating catastrophic forgetting.

- EVCL outperforms existing baselines in both domain-incremental and task-incremental learning scenarios, achieving higher average test accuracies on a variety of continual learning tasks for deep discriminative models.
- EVCL demonstrates significantly less degradation in average accuracy as the number of tasks increases compared to other methods, underscoring its superior performance and stability in managing catastrophic forgetting.

2. Related Work

A variety of approaches have been proposed to mitigate catastrophic forgetting, including regularisation, memory-based, and model architectural strategies. Our work on integrating VCL and EWC falls within the context of inference-based regularisation methods and parameter adaptation techniques, and falls within the broader landscape of regularisation, inference, and hybrid methods.

Regularisation Approaches. EWC (Kirkpatrick et al., 2017) and Synaptic Intelligence (SI) (Zenke et al., 2017) are foundational regularisation approaches that preserve critical parameters by penalising changes to significant weights. However, EWC's reliance on the Laplace approximation can lead to underestimation of parameter importance (Huszár, 2018). This work addresses this limitation by integrating VCL with EWC, providing a more robust framework for continual learning.

Inference and Bayesian Approaches. VCL (Nguyen et al., 2017) tackles catastrophic forgetting by approximating the posterior distributions of model parameters across sequential tasks. Likelihood-tempered VCL (Zhang et al., 2018; Osawa et al., 2019) modifies the approach by down-weighting the KL-divergence, but still faces challenges with posterior alignment. Our combined EVCL model enhances VCL's adaptation ability and provides a better posterior approximation.

Replay and Rehearsal Methods. Replay-based methods, such as Experience Replay (ER) (Rolnick et al., 2019) and Deep Generative Replay (DGR) (Shin et al., 2017), complement the current training data with representative data from past observations or generated pseudo-data. VCL with coreset (Nguyen et al., 2017) extends VCL by incorporating a coreset memory to regularise the variational posterior. Our approach is more memory-efficient and scalable as it does not require storing past data or generating pseudo-data.

Hybrid Model Approaches. Recent works have explored combining different approaches to improve continual learning performance. The Progress and Compress model (Schwarz et al., 2018) uses concepts from Progressive Neural Networks while employing EWC to protect important weights. Generalized VCL (GVCL) (Loo et al., 2020b) integrates multi-task FiLM architecture and likelihood-tempered variational inference, bridging VCL and Online EWC. Our approach aims for a computationally efficient integration of VCL and EWC without significantly increasing model complexity or memory demands.

3. Preliminaries

3.1. Variational Continual Learning (VCL)

VCL (Nguyen et al., 2017) is a Bayesian approach to continual learning that approximates the posterior distribution of the model parameters given the current task data and the accumulated approximate posteriors from previous tasks as the prior. The objective function of VCL for task t is derived from the variational lower bound (ELBO) of the log-likelihood of the data (Appendix B Lemma 1), where we maximise the variational lower bound (i.e negative online variational free energy) to the online marginal likelihood, or equivalently minimising Kullback-Leibler divergence $\mathrm{KL}(\cdot\|\cdot)$:

$$\mathcal{L}^t_{\mathrm{VCL}}\left(q_t(\boldsymbol{\theta})\right) = \sum_{n=1}^{N_t} \mathbb{E}_{\boldsymbol{\theta}\sim q_t(\boldsymbol{\theta})}\left[\log p\left(y_t^{(n)} \mid \boldsymbol{\theta}, \mathbf{x}_t^{(n)}\right)\right] - \mathrm{KL}\left(q_t(\boldsymbol{\theta})\|q_{t-1}(\boldsymbol{\theta})\right) \quad (1)$$

where $q_t(\boldsymbol{\theta})$ is the variational approximation of the posterior distribution at task t, N_t is the number of data points in task t, $y_t^{(n)}$ and $\mathbf{x}_t^{(n)}$ are the target and input for the n-th data point in task t.

3.2. Elastic Weight Consolidation (EWC)

EWC (Kirkpatrick et al., 2017) is a regularization-based approach that preserves important parameters for previous tasks while allowing the model to adapt to new tasks. The EWC loss function is given by $\mathcal{L}_{\mathrm{EWC}}(\boldsymbol{\theta}) = \sum_i \frac{\lambda}{2} F_i^{t-1}\left(\theta_i - \theta^*_{t-1,i}\right)^2$, where λ is a hyperparameter controlling the strength of the EWC regularization, F_i^{t-1} is the Fisher information matrix computed from the previous task, θ_i are the current model parameters, and $\theta^*_{t-1,i}$ are the optimal model parameters for the previous task. The Fisher Information Matrix (FIM) $F_{t-1,i} = \mathbb{E}_{p(\mathcal{D}_{t-1}|\boldsymbol{\theta})}\left[\left(\frac{\partial \log p(\mathcal{D}_{t-1}|\boldsymbol{\theta})}{\partial \theta_i}\right)^2\right]$ captures the importance of each parameter θ_i for the previous task $t-1$.

4. Elastic Variational Continual Learning with Weight Consolidation (EVCL)

Variational Continual Learning (VCL) approximates the posterior distribution of model parameters, capturing uncertainty and facilitating knowledge transfer across tasks. However, VCL suffers from catastrophic forgetting due to the divergence between the approximate and true posteriors, leading to accumulated errors. Additionally, VCL's reliance on coresets and additional episodic memory can limit its scalability and flexibility (Nguyen et al., 2017).

To address these limitations, we propose Elastic Variational Continual Learning with Weight Consolidation (EVCL), which integrates Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) into the VCL framework. EWC is a regularisation-based approach that penalises changes to important parameters for previous tasks, as determined by the Fisher Information Matrix (FIM). By incorporating EWC's regularisation term into the variational objective, EVCL balances the performance on the current task with the retention of knowledge from previous tasks.

EVCL leverages the strengths of both techniques, using VCL to approximate the posterior distribution and EWC to identify and preserve important parameters. The parameter space is shared among tasks with controlled overlaps, allowing for a form of parameter space decomposition that shields critical parameters from large updates. This ensures that EVCL does not diverge significantly from the parameter configurations crucial for previously learned tasks, thereby reducing the compounding of approximation errors across tasks.

The Fisher Information Matrix with diagonal approximation in EWC captures the importance of each parameter for the previous tasks, guiding the variational posterior to retain crucial knowledge. By leveraging historical parameter importance, EVCL minimises the need for episodic memory. The regularisation provided by the Fisher Information also guides the variational approximation to remain closer to the true posterior, enhancing the overall fidelity of the model across multiple tasks.

EVCL combines the variational approximation methods of VCL with the stability-enhancing regularisation strategies of EWC, addressing the drawbacks associated with each method when used independently. In the following section we describe how we integrate VCL and EWC into a unified objective function and then empirically demonstrating its effectiveness in mitigating catastrophic forgetting across various continual learning benchmarks for deep discriminative models.

4.1. Approach

This section presents our approach EVCL, combining VCL and EWC to mitigate catastrophic forgetting in continual learning.

For EVCL, to integrate EWC into the variational framework of VCL, we add the EWC penalty term to the VCL loss function. The EWC penalty term is computed as the sum of the squared differences between the models current variational parameters and those of the previous task, weighted by the Fisher Information Matrix F_i^{t-1} and the regularization strength λ.

By adding the EWC penalty term to the VCL loss, we obtain the combined EVCL loss function:

$$\mathcal{L}_{\text{EVCL}}^{t}\left(q_t(\boldsymbol{\theta})\right) = \mathcal{L}_{\text{VCL}}^{t}\left(q_t(\boldsymbol{\theta})\right) + \sum_i \frac{\lambda}{2} F_i^{t-1} \left[\left(\mu_{t,i} - \mu_{t-1,i}\right)^2 + \left(\sigma_{t,i}^2 - \sigma_{t-1,i}^2\right)^2\right] \quad (2)$$

where $\mathcal{L}_{\text{VCL}}^{t}\left(q_t(\boldsymbol{\theta})\right)$ is computed using a Gaussian mean-field approximate posterior $q_t(\theta) = \prod_{d=1}^{D} \mathcal{N}(\theta_{t,d}; \mu_{t,d}, \sigma_{t,d}^2)$ to allow analytical computation of the KL divergence and expectations. Here, $\mu_{t,i}$ and $\sigma_{t,i}^2$ represents the mean and variance of the variational posterior for parameter θ_i at task t respectively; and $\mu_{t-1,i}$ and $\sigma_{t-1,i}^2$ represents the mean and variance of the variational posterior for parameter θ_i at the previous task $(t-1)$ respectively.

Expected log-likelihood is approximated using simple Monte Carlo and local reparameterisation (Kingma et al., 2015; Salimans & Knowles, 2013; Kingma & Welling, 2014). At the first time step, $q_0(\theta)$ prior distribution is chosen as a multivariate Gaussian distribution (Graves, 2011; Blundell et al., 2015; Nguyen et al., 2017). The integration of the Fisher Information Matrix in EVCL employs a diagonal approximation to efficiently capture parameter importance without significantly increasing computational complexity.

The combined loss function encourages the model to find a variational approximation that not only fits the current task well but also integrates information from previous tasks through the VCL term, while the EWC term penalizes significant deviations in parameters crucial for previous tasks, ensuring stability and retaining knowledge.

The hyperparameter λ, which we set to 100, controls the strength of the EWC regularization for EVCL, allowing for a trade-off between plasticity and stability. Notably, there is no explicit weighting between the VCL and EWC components in the combined EVCL loss; instead, the balance is controlled implicitly through the variational framework and the regularization strength parameter λ of EWC.

By incorporating the EWC penalty term into the variational framework of VCL, we effectively regularize the variational approximation to preserve the important parameters for the previous tasks while allowing the model to adapt to new tasks, mitigating catastrophic forgetting in continual learning scenarios.

An overview of the full algorithm can be found in Appendix A.

5. Experiments

We apply the proposed EVCL framework to discriminative models, specifically fully-connected neural network classifiers, and evaluate its performance on five tasks: PermutedMNIST, SplitMNIST, SplitNotMNIST, SplitFashionMNIST, and SplitCIFAR-10.

For PermutedMNIST, we test the domain incremental context of continual learning using a single-head MLP. For the other tasks, we employ a multi-head MLP with separate output heads but shared model parameters to test the incremental continual learning contexts. EVCL is compared against multiple baselines including VCL, VCL with Random Coreset, VCL with K-center coreset, EWC, and standalone Coreset models.

The EWC penalty used in EVCL and standalone EWC model, is set to $\lambda = 100$. The Fisher Information Matrix is estimated using 5000 samples from the data, and the coreset size is set to 200. We train each of these models for 100 epochs with a batch size of 256. During evaluation, after training on each task, we test on all tasks seen so far and compute the average test accuracies across these tasks, with results averaged over 3 runs.

For EVCL, during training, the loss function is minimized with respect to the variational parameters of $q_t(\boldsymbol{\theta})$ using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 1e-3. The gradients of the EWC term used in EVCL are computed using the reparameterization trick (Kingma et al., 2015; Kingma & Welling, 2014), enabling end-to-end training of the model.

5.1. Task setup and Results

PermutedMNIST: This is a domain-incremental learning task consisting of labeled MNIST images whose pixels have undergone a fixed random permutation (Goodfellow et al., 2013; Kirkpatrick et al., 2017; Zenke et al., 2017). This simple setup tests the robustness of models to domain shifts. For this task, we use a single-head two-layer MLP with 100 hidden units per layer and ReLU activations. The average test accuracy after training on 5 tasks shows that EVCL achieves 93.5%, significantly outperforming VCL at 91.5%, VCL with Random-Coreset at 91.68%, VCL with K-Center Coreset at 92%, and EWC at 65%.

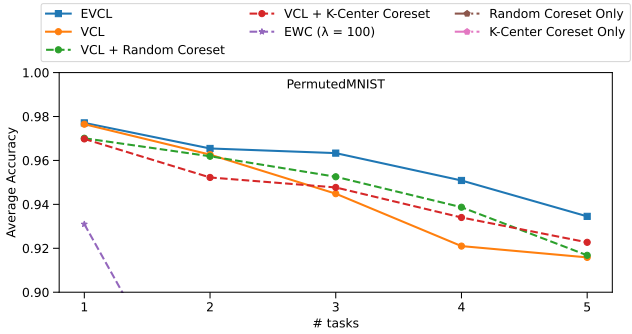

Figure 1. Test set average accuracy over PermutedMNIST for EVCL and baseline models.

SplitMNIST: This task comprises sequential binary classification tasks derived from MNIST digits, specifically 0/1, 2/3, 4/5, 6/7, and 8/9 (Zenke et al., 2017). A multi-head MLP network with 256 hidden units per layer is employed for this task. The average test accuracy after training on all 5 tasks in the SplitMNIST experiment shows that EVCL leads with 98.4%, compared to VCL at 94%, VCL with Random Coreset at 96%, VCL with K-Center Coreset at 94.4%, and EWC at 88%.

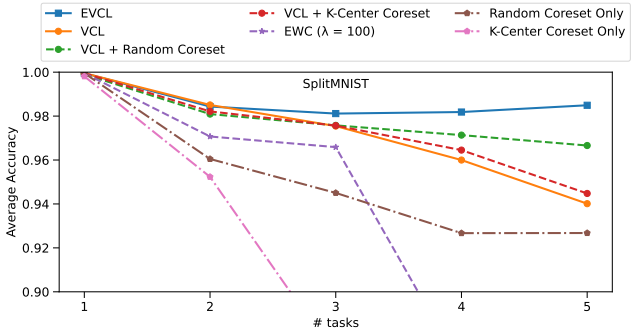

Figure 2. Test set average accuracy over SplitMNIST for EVCL and baseline models.

SplitNotMNIST: This experiment used by (Nguyen et al., 2017) challenges models with character classification from A to J across various fonts, split over 5 binary classification tasks: A/F, B/G, C/H, D/I, and E/J. A deeper multi-head network with four layers of 150 hidden units per layer and shared parameters is utilized. The average test accuracy after training on all tasks shows that EVCL records 91.7%, outstripping VCL at 89.7%, VCL with Random Coreset at 86%, VCL with K-Center Coreset at 82.7%, and EWC at 62.9%.

SplitFashionMNIST: This experiment involves classifying fashion items into five categories: top/trouser,

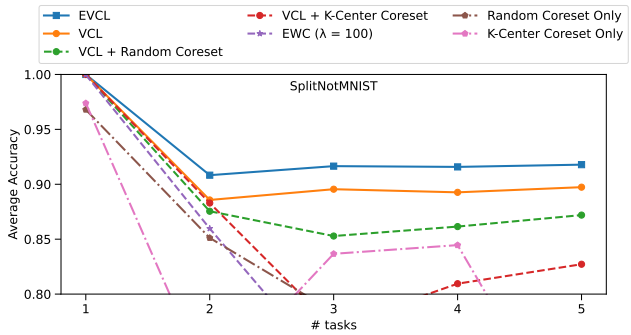

Figure 3. Test set average accuracy over SplitNotMNIST for EVCL and baseline models.

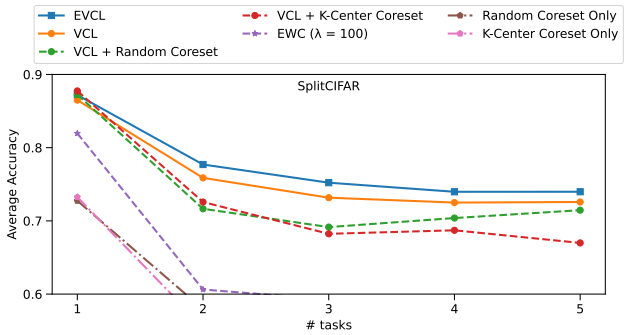

Figure 5. Test set average accuracy over SplitCIFAR-10 for EVCL and baseline models.

pullover/dress, coat/sandal, shirt/sneaker, and bag/ankle boots (Xiao et al., 2017). The architecture mirrors that of SplitNotMNIST, using a deeper multi-head network with four layers of 150 hidden units and shared parameters. The average test accuracy after training on all tasks shows that EVCL attains 96.2%, exceeding VCL at 90%, VCL with Random Coreset at 86%, VCL with K-Center Coreset at 86.3%, and EWC at 74%.

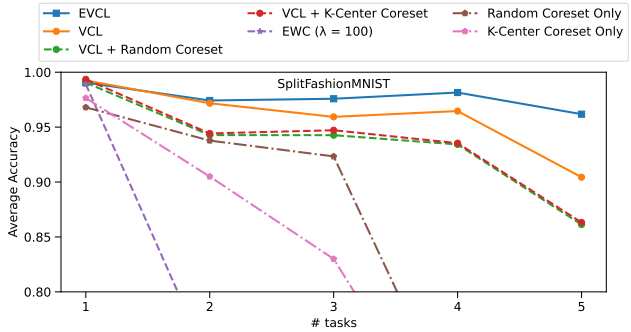

Figure 4. Test set average accuracy over SplitFashionMNIST for EVCL and baseline models.

SplitCIFAR-10: This experiment tests classification across complex and diverse images such as airplanes, automobiles, birds, and cats, split over five binary classification tasks: airplane/automobile, bird/cat, deer/dog, frog/horse, and ship/truck (Krizhevsky et al., 2009). The average test accuracy after training on all tasks shows that EVCL achieves 74%, surpassing VCL at 72%, VCL with Random Coreset at 71.5%, VCL with K-Center Coreset at 67%, and EWC at 59%.

Across all tasks, EVCL consistently outperforms traditional VCL, VCL augmented with random and k-center coresets, EWC, and coreset-only approaches, demonstrating its effectiveness in managing catastrophic forgetting in complex continual learning scenarios. This improvement underscores the benefits of integrating EWC within VCL, especially in multi-task settings where model stability and adaptability are crucial. While all methods exhibit some degradation in accuracy as the number of tasks increases, EVCL shows significantly less degradation compared to the other methods, highlighting its robustness and superior performance.

6. Conclusion

Our proposed Elastic Variational Continual Learning (EVCL) model integrates Elastic Weight Consolidation (EWC) into the Variational Continual Learning (VCL) framework to mitigate catastrophic forgetting. EVCL consistently outperforms existing baselines for discriminative tasks by effectively balancing plasticity and stability. Additionally, EVCL exhibits significantly less degradation in performance as the number of tasks increases compared to other methods.

The model's performance could be further enhanced by incorporating natural gradient methods, such as Kronecker-factored Approximate Curvature (K-FAC) and Online Natural Gradient Descent (Martens & Grosse, 2015; Ollivier, 2018), to better approximate the Fisher Information Matrix (FIM) beyond the diagonal approximation, thus capturing the curvature of the parameter space more accurately and addressing EWC's limitations (Zhang et al., 2018).

Future work could explore the application of EVCL to generative models and reinforcement learning, as well as extending the framework to handle complex task structures and model architectures (Sankararaman et al., 2022). Additionally, incorporating experimental replay mechanisms (Rolnick et al., 2019; Shin et al., 2017; Nguyen et al., 2017), sparse coding techniques (Sarfraz et al., 2023), and parameter efficient fine-tuning approaches (Hu et al., 2021; Sharma et al., 2023) could potentially lead to more conclusive results in the field while improving the scalability and robustness of the approach.

References

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Graves, A. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Huszár, F. Note on the quadratic penalties in elastic weight consolidation. *Proceedings of the National Academy of Sciences*, 115(11):E2496–E2497, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Stochastic gradient vb and the variational auto-encoder. In *Second international conference on learning representations, ICLR*, volume 19, pp. 121, 2014.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. *Semantic Scholar preprint CorpusID:18268744*, 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Lee, S.-W., Kim, J.-H., Jun, J., Ha, J.-W., and Zhang, B.-T. Overcoming catastrophic forgetting by incremental moment matching. *Advances in neural information processing systems*, 30, 2017.

Loo, N., Swaroop, S., and Turner, R. E. Combining variational continual learning with film layers. In *4th Lifelong Machine Learning Workshop at ICML 2020*, 2020a.

Loo, N., Swaroop, S., and Turner, R. E. Generalized variational continual learning. *arXiv preprint arXiv:2011.12328*, 2020b.

Martens, J. and Grosse, R. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

Masse, N. Y., Grant, G. D., and Freedman, D. J. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences*, 115(44):E10467–E10475, 2018.

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

Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017.

Nguyen, C. V., Achille, A., Lam, M., Hassner, T., Mahadevan, V., and Soatto, S. Toward understanding catastrophic forgetting in continual learning. *arXiv preprint arXiv:1908.01091*, 2019.

Ollivier, Y. Online natural gradient as a kalman filter. *arXiv preprint arXiv:1703.00209*, 2018.

Osawa, K., Swaroop, S., Khan, M. E. E., Jain, A., Eschenhagen, R., Turner, R. E., and Yokota, R. Practical deep learning with bayesian principles. *Advances in neural information processing systems*, 32, 2019.

Pan, P., Swaroop, S., Immer, A., Eschenhagen, R., Turner, R., and Khan, M. E. E. Continual deep learning by functional regularisation of memorable past. *Advances in neural information processing systems*, 33:4453–4464, 2020.

Ratcliff, R. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

Ring, M. B. Child: A first step towards continual learning. *Machine Learning*, 28(1):77–104, 1997.

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., and Wayne, G. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Salimans, T. and Knowles, D. A. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis, Volume 8, Number 4 (2013), 837-882*, 2013.

Sankararaman, K. A., Wang, S., and Fang, H. Bayesformer: Transformer with uncertainty estimation. *arXiv preprint arXiv:2206.00826*, 2022.

Sarfraz, F., Arani, E., and Zonooz, B. Sparse coding in a dual memory system for lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9714–9722, 2023.

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

Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y. W., Pascanu, R., and Hadsell, R. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pp. 4528–4537. PMLR, 2018.

Sharma, P., Ash, J. T., and Misra, D. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. *arXiv preprint arXiv:2312.13558*, 2023.

Shin, H., Lee, J. K., Kim, J., and Kim, J. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

SUTTON, R. Online learning with random representations. In *Proc. 10th Int. Conf. on Machine Learning (ML-93)*, pp. 314–321, 1993.

Swaroop, S., Nguyen, C. V., Bui, T. D., and Turner, R. E. Improving and understanding variational continual learning. *arXiv preprint arXiv:1905.02099*, 2019.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. PMLR, 2017.

Zhang, G., Sun, S., Duvenaud, D., and Grosse, R. Noisy natural gradient as variational inference. In *International conference on machine learning*, pp. 5852–5861. PMLR, 2018.

Appendix

A. Elastic Variational Continual Learning with Weight Consolidation (EVCL) Algorithm

Algorithm 1: Elastic Variational Continual Learning with Weight Consolidation (EVCL)

Input: Dataset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_T\}$, learning rate α, EWC strength λ
Output: Variational parameters ϕ_t for each task t
Initialize variational parameters ϕ_0; Initialize prior $p(\boldsymbol{\theta}|\mathcal{D}_0) \leftarrow q_{\phi_0}(\boldsymbol{\theta})$
for $t = 1, \ldots, T$ **do**
  Initialize $\phi_t \leftarrow \phi_{t-1}$
  for *each batch* $\mathcal{B} \subset \mathcal{D}_t$ **do**
    Compute VCL loss: $\mathcal{L}^t_{\text{VCL}} = \frac{1}{|\mathcal{B}|}\sum_{n=1}^{|\mathcal{B}|} \mathbb{E}_{q_{\phi_t}(\boldsymbol{\theta})}[\log p(y_n|\boldsymbol{\theta}, \mathbf{x}_n)] - \text{KL}(q_{\phi_t}(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D}_{1:t-1}))$
    if $t > 1$ **then**
      Compute EWC loss: $\mathcal{L}^t_{\text{EWC}} = \sum_i \frac{\lambda}{2} F_i^{t-1} \left[(\mu_{t,i} - \mu_{t-1,i})^2 + (\sigma^2_{t,i} - \sigma^2_{t-1,i})^2\right]$
    else
      $\mathcal{L}^t_{\text{EWC}} = 0$
    end
    Compute total loss: $\mathcal{L}^t_{\text{EVCL}} = \mathcal{L}^t_{\text{VCL}} + \mathcal{L}^t_{\text{EWC}}$
    Update variational parameters: $\phi_t \leftarrow \phi_t - \alpha \nabla_{\phi_t} \mathcal{L}^t_{\text{EVCL}}$
  end
  Store model parameters mean and variance: $\mu_{t-1,i} = \mathbb{E}_{q_{\phi_t}(\boldsymbol{\theta})}[\theta_i]$ and $\sigma^2_{t-1,i} = \text{Var}_{q_{\phi_t}(\boldsymbol{\theta})}[\theta_i]$ for all i
  Compute Fisher Information Matrix: $F_i^t = \mathbb{E}_{p(\mathcal{D}_t|\boldsymbol{\theta})}\left[\left(\frac{\partial \log p(\mathcal{D}_t|\boldsymbol{\theta})}{\partial \theta_i}\right)^2\right]$ for all i
  Set prior for next task: $p(\boldsymbol{\theta}|\mathcal{D}_{1:t}) \leftarrow q_{\phi_t}(\boldsymbol{\theta})$
end

B. Lemmas

Lemma 1 (Evidence Lower Bound): Let $p(x|\theta)$ be the likelihood of the data x given the parameters θ, and let $q_\lambda(z)$ be a variational distribution parameterized by λ. The log marginal likelihood $\log p(x|\theta)$ can be decomposed as:

$$\log p(x|\theta) = \text{ELBO}_{\theta,\lambda} + \text{KL}(q_\lambda(z)||p(z|x,\theta)), \tag{3}$$

where the Evidence Lower Bound (ELBO) is defined as: $\text{ELBO}_{\theta,\lambda} = \int q_\lambda(z) \log \frac{p(z,x|\theta)}{q_\lambda(z)} dz$.

Proof: Starting from the definition of the ELBO, we have:

$$\begin{aligned}
\log p(x|\theta) - \text{KL}(q_\lambda(z)||p(z|x,\theta)) &= \log p(x|\theta) \int q_\lambda(z) dz - \int q_\lambda(z) \log \frac{q_\lambda(z)}{p(z|x,\theta)} dz && (4)\\
&= \int q_\lambda(z) \log p(x|\theta) dz + \int q_\lambda(z) \log \frac{p(z|x,\theta)}{q_\lambda(z)} dz && (5)\\
&= \int q_\lambda(z) \log \frac{p(z,x|\theta)}{q_\lambda(z)} dz && (6)\\
&= \int q_\lambda(z) \log p(x,z|\theta) dz - \int q_\lambda(z) \log q_\lambda(z) dz && (7)\\
&= \text{ELBO}_{\theta,\lambda}. && (8)
\end{aligned}$$

Since $\text{KL}(\cdot||\cdot) \geq 0$, we have $\log p(x|\theta) \geq \text{ELBO}_{\theta,\lambda}$ for any θ, λ, and q_λ. Hence, $\text{ELBO}_{\theta,\lambda}$ is called the Evidence Lower Bound, and

$$\arg\min_\lambda \text{KL}(q_\lambda(z)||p(z|x,\theta)) = \arg\max_\lambda \text{ELBO}_{\theta,\lambda}. \tag{9}$$