Variational Continual Learning

UNIVERSITY OF CAMBRIDGE

Jiajun He, Johannes Vallikivi, Nineli Lashkarashvili

Introduction

Continual learning aims to learn incrementally when data arrive in a possibly non i.i.d. way whereby tasks may change over time, without revisiting all previous data.

Methods

Bayes Rule gives a well-defined way to perform Continual Learning:

$$\underbrace{p(\theta|\mathcal{D}_{1:T})}_{\textit{new posterior}} \propto \underbrace{p(\theta|\mathcal{D}_{1:T-1})}_{\textit{old posterior/new prior}}\underbrace{p(\mathcal{D}_T|\theta)}_{\textit{likelihood}}$$

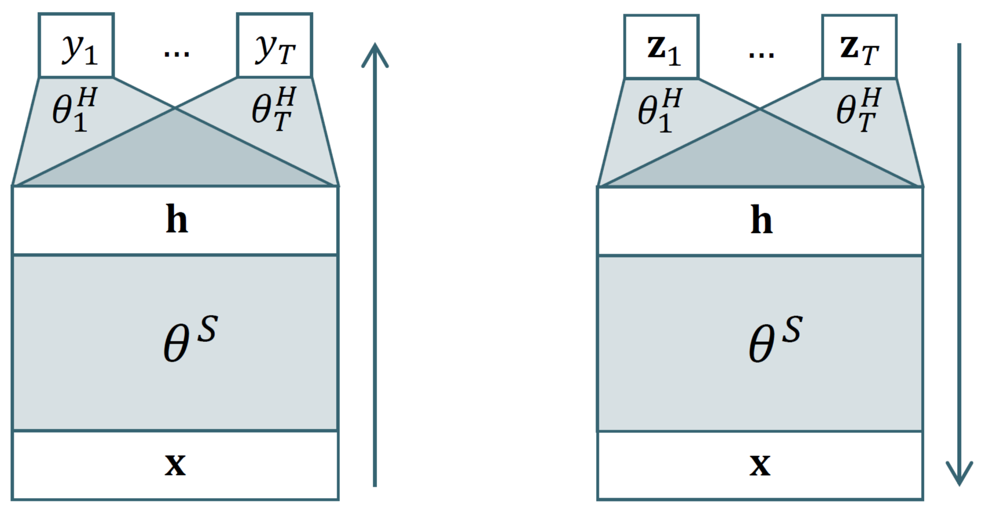
The intractability of posteriors is tackled by Variational Inference by $q(\theta) \approx p(\theta|\mathcal{D})$:

$$\log p(\mathcal{D}) = \underbrace{\mathbb{E}_{q(\theta)}\left[\log p(\mathcal{D}|\theta)\right] - \mathrm{KL}\left[q(\theta)||p(\theta)\right]}_{\mathrm{ELBO}} + \mathrm{KL}\left[q(\theta)||p(\theta|\mathcal{D})\right]$$

The error accumulated by sequential approximation is corrected by keeping a small "coreset" to avoid catastrophic forgetting:

$$\begin{tikzcd}
p(\theta) \arrow[r, "\text{propagation}", "\mathcal{D}_1\backslash\mathcal{C}_1"'] & q(\theta|\mathcal{D}_1\backslash\mathcal{C}_1) \arrow[r, "\text{propagation}", "\mathcal{D}_2\cup\mathcal{C}_1\backslash\mathcal{C}_2"'] \arrow[d, "\mathcal{C}_1"] & q(\theta|\mathcal{D}_1\cup\mathcal{D}_2\backslash\mathcal{C}_2) \arrow[r, "\text{propagation}"] \arrow[d, "\mathcal{C}_2"] & \cdots \\
 & q(\theta|\mathcal{D}_1) & q(\theta|\mathcal{D}_1\cup\mathcal{D}_2) &
\end{tikzcd}$$

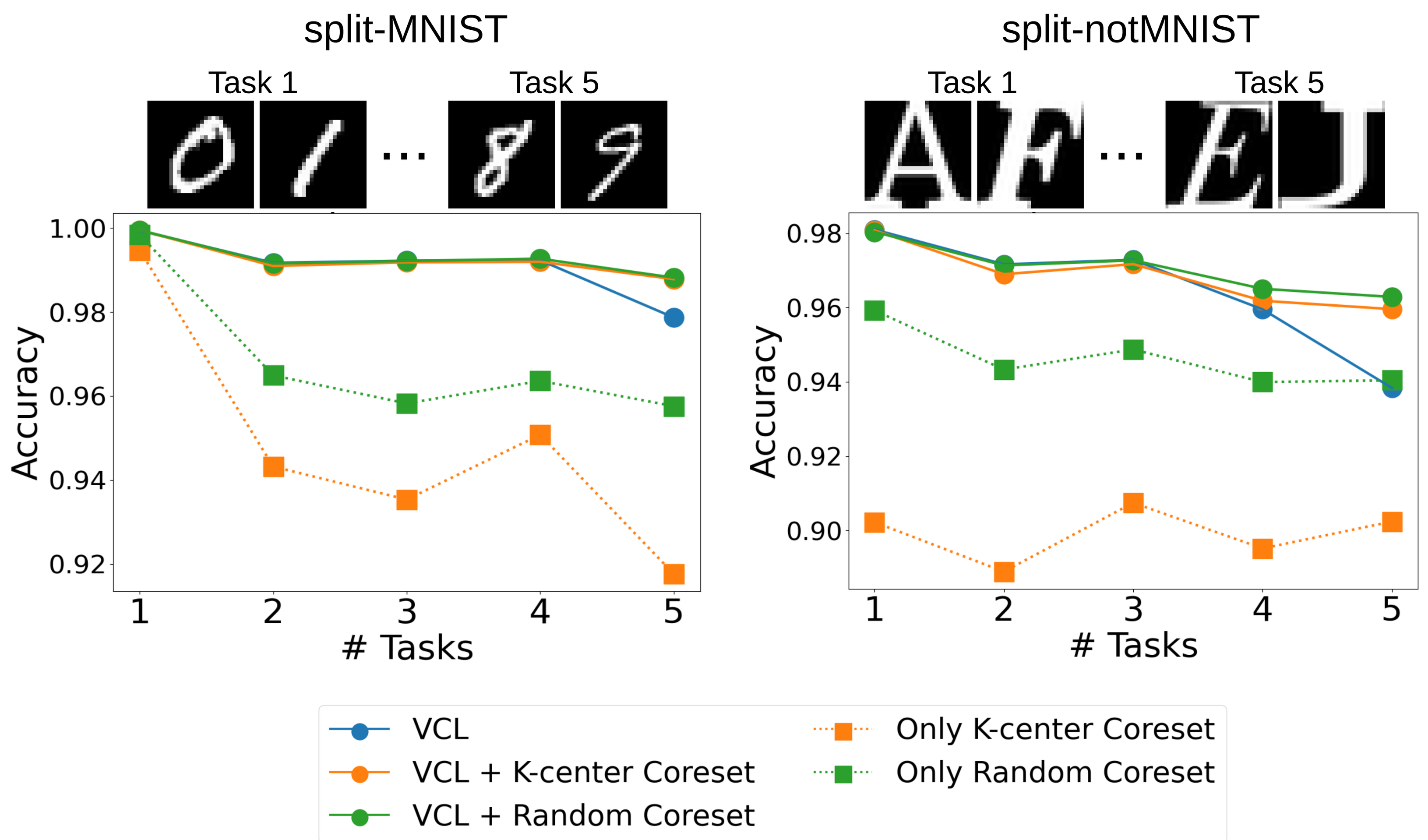
Bayesian Neural Networks with the following architectures are used:



Discriminative Model   Generative Model

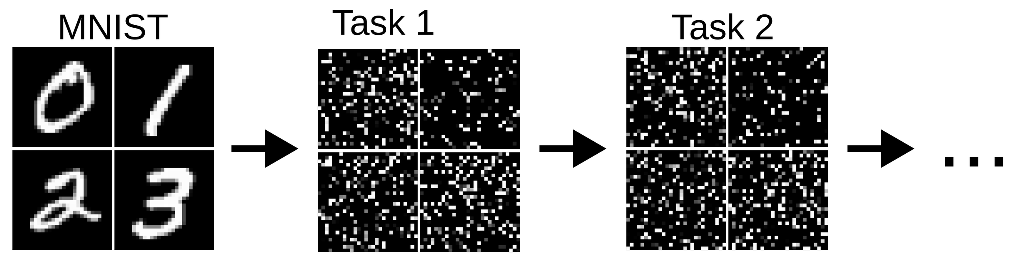
Split-MNIST/notMNIST

Datasets for each task are generated by splitting MNIST/notMNIST into subsets of two classes each.



Permuted MNIST

Datasets for each task are generated by permuting pixels of MNIST images.



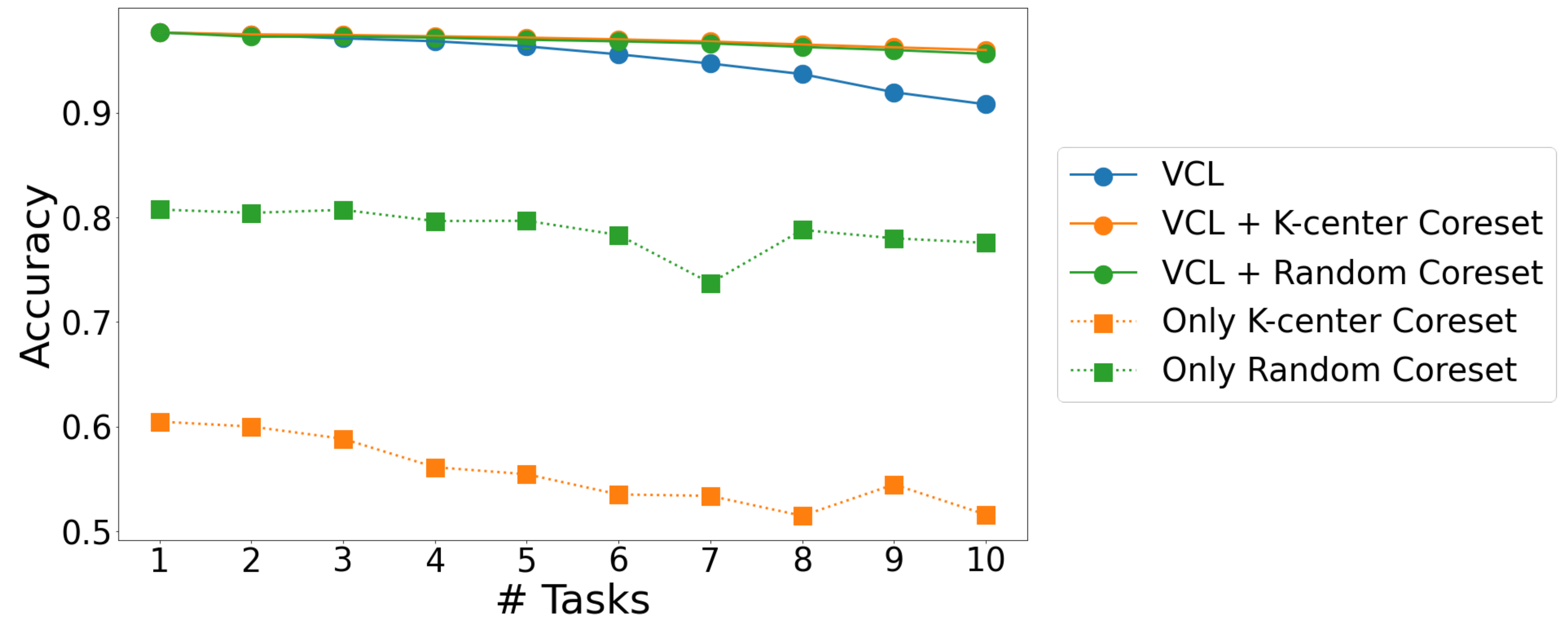
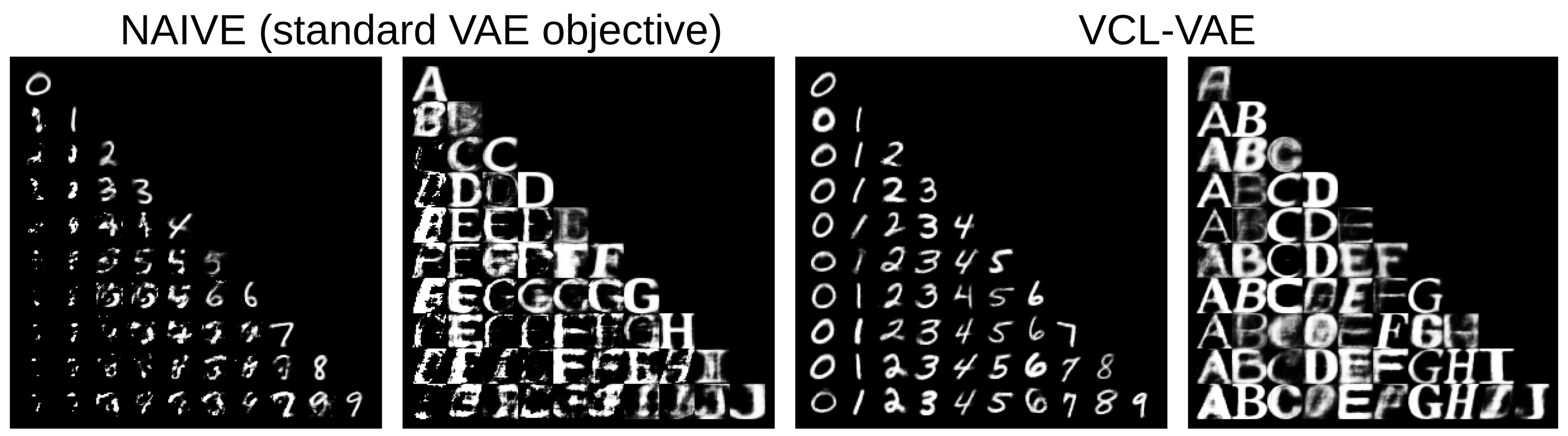
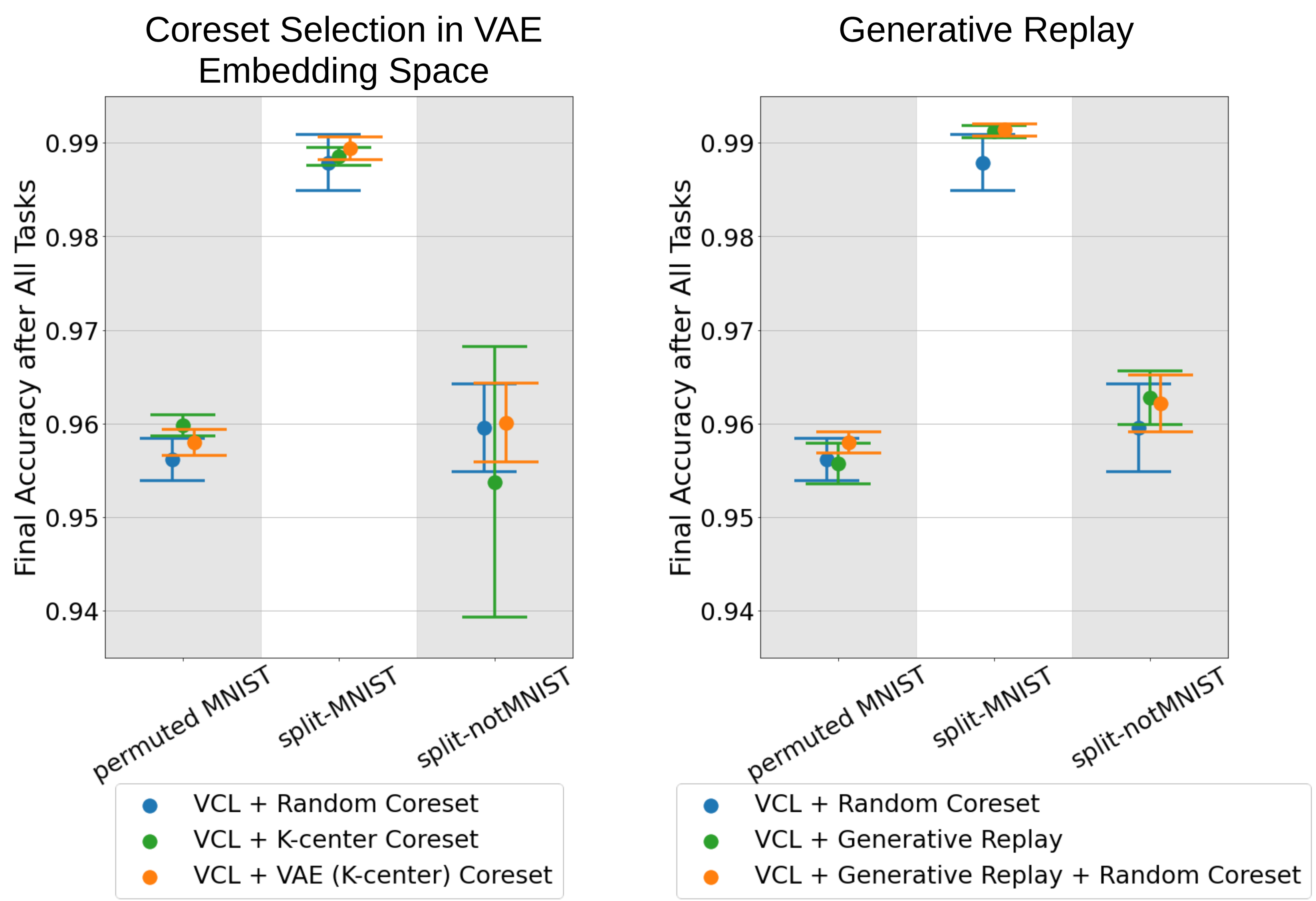


Image Generation (VCL-VAE)

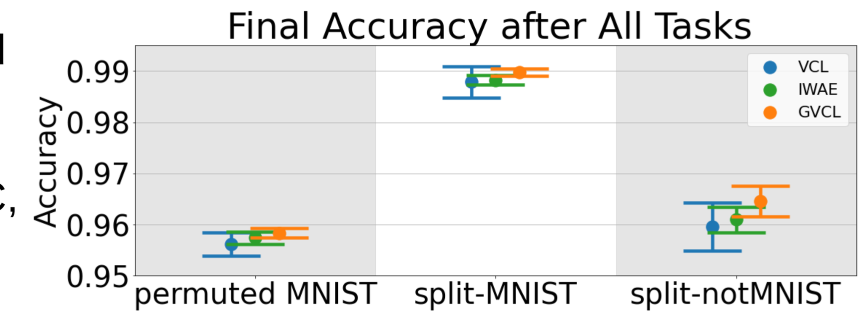


Extension: Memory Mechanism



Extension: Alternative Loss

- IWAE Bound provides a tighter bound of marginal likelihood than ELBO;
- GVCL Bound uniforms VCL and EWC, another continual learning algorithm.



Extension: VCL-GAN

VCL on GAN is tricky:
- It is difficult to balance the Bayesian-generator and discriminator;
- GAN loss is not a well-defined likelihood.

A successful VCL-GAN involves 3 key components:
(1) Wasserstein GAN Loss as the "negative log-likelihood" term
(2) Generalized-VCL Bound
(3) Bayesian body with task-specific non-Bayesian heads

Ablation study:

Standard GAN loss+(2)+(3)   (1)+VCL ELBO+(3)   (1)+(2)+Bayesian head

Conclusions

- VCL is a universal continual learning framework for both discriminative models and generative models;
- A more representative coreset tends to improve knowledge retention – Generative Replay combined with coreset provides a consistently better memory mechanism;
- Both IWAE Bound and GVCL Bound present superior results to the original;
- VCL was successfully applied to GAN. WGAN loss, GVCL Bound and non-Bayesian heads are key to fully functional VCL-GAN.

References

[1] Burda, Y., Grosse, R., & Salakhutdinov, R. (2015). Importance weighted autoencoders. arXiv preprint arXiv:1509.00519.
[2] Nguyen, C. V., Li, Y., Bui, T. D., & Turner, R. E. (2017). Variational continual learning. arXiv preprint arXiv:1710.10628.
[3] Shin, H., Lee, J. K., Kim, J., & Kim, J. (2017). Continual learning with deep generative replay. Advances in neural information processing systems, 30.
[4] Swaroop, S., Nguyen, C. V., Bui, T. D., & Turner, R. E. (2019). Improving and understanding variational continual learning. arXiv preprint arXiv:1905.02099.
[5] Loo, N., Swaroop, S., & Turner, R. E. (2020). Generalized variational continual learning. arXiv preprint arXiv:2011.12328.