Author response for Continual Unsupervised Representation Learning

We would like to thank the reviewers for their insightful and constructive comments, which will certainly help to strengthen the paper. We have systematically addressed each of their concerns, which we summarise below.

Clarification on dynamic expansion, and thresholds c_{new} and N_{new} (R1, R2) c_{new} is similar to the alpha parameter in a Dirichlet process, controlling how easily additional components are added to the mixture. To analyse its effect, we have added a sweep over this parameter in the experiments: with lower threshold values, the model can use more components, leading to higher accuracies at the cost of additional capacity. In terms of N_{new}, the model is quite insensitive: we set the value to 100 but found the performance to be quite stable for values up to 1000. In the case where new classes are very different (asked by R2), our intuition was that the new class is still closer to existing classes than to randomly initialised weights (mentioned by R1), and we verified empirically that our approach yields better performance than this case.

Model details (R2) and theoretical aspects (R3) We only use one sample for z for simplicity/efficiency, but this can easily be extended in a similar fashion to Importance-Weighted Autoencoders - this is now mentioned in the paper. We have also added additional explanation and intuition around how Eqn. 4 (the supervised loss) is derived, and how each term corresponds to its equivalent in Eqn 3. Finally, we have also added a clarification of θ_{prev} leading up to Eqn. 6: it denotes the parameters of the previous model snapshot, specifically those of the prior $p(z|y)$ and decoder $p(x|z)$.

More in-depth discussion on results ("why") (R1) Interestingly, the performance of individual classes for CURL seems more dependent on similarity between classes: for example, 5 is very similar to 3, 8, and 9, which is likely why its performance is lower than the more distinguishable 0. This is in contrast to previous continual learning work, in which forgetting is often correlated with whether classes are learned earlier or later. We have added additional discussion in Section 4.2 to make this clear, as well as further class-specific plots in the appendix.

Experimental details (R1) The details around error bars and training/test splits are in Appendix C. We have modified this to clarify that the errors show the standard deviation over 5 runs. The time complexity has also been added to the Appendix: training time is around $15 - 20$ minutes for MNIST and $4 - 5$ hours for Omniglot (with additional time for analysis and evaluation on validation and test sets).

Comparison to clustering baselines (R2) Unfortunately, to the best of our knowledge, [hierarchical] clustering techniques still assume i.i.d data, while in our work we investigate clustering in a non-stationary setting. In the i.i.d case, we do include simple baselines in Section 4.4, reporting kNN error with raw pixels and random network encodings.

Low Omniglot accuracy (R2) The reviewer is correct that the general performance on Omniglot is quite poor - this is largely due to the more challenging task with 50 classes, but performance is still much better than random chance at 2%. Though the large error makes it a bit more difficult to separate different techniques, we still observe strong performance with respect to baselines in the ablation and external comparison.

Reproducibility of Figure 4 and Figure 2b (R2) We believe the reviewer is referring to the class-specific analysis in Figure 3b and Figure 4. This kind of behavior is very robust/stable across multiple seeds: similar classes are confused and similar numbers of components are used for each class.

Clarification of "Unsupervised i.i.d. learning" (R3) By i.i.d. learning, we refer to the setting where all classes are sampled with uniform probability from the beginning of training. We have improved its definition in Section 4.1.

Improved experiments

- We have managed to increase performance (generally across the board) with small architectural changes and further hyperparameter tuning. The numbers have been updated and details have been added to the appendix.
- We are running experiments with CIFAR-100 (suggested by R1) and hope to have this by the camera-ready deadline. Surprisingly, there is little past work demonstrating class-discriminative unsupervised learning with CIFAR-100, even in the i.i.d setting, so we focus on the supervised incremental CIFAR-100 benchmark.
- After correspondence with authors of related work, we found that the unsupervised i.i.d benchmark was performed with sampled latents, and redid the experiments accordingly. The analysis and conclusions drawn from the original submitted version still hold, so merely the numbers have slightly changed.
- We are currently in the process of open-sourcing our code.