Author Response for "Exemplar VAEs for Exemplar based Generation and Data Augmentation"

We thank the reviewers for their valuable feedback. We respond to the main comments below, starting with the most critical review.

R4: *"My main concern is about the novelty and significance of this work..."* The reviewer's assertion about limited novelty may stem from a characterization of novelty in terms of modeling modifications. We claim no credit for the idea of using a mixture of variational posteriors as the prior, since it is known in the varitional inference community that the Bayes optimal prior for a VAE is a mixture of variational posteriors *a.k.a.,* aggregated posterior [2]. That said, previous work [3, 1] makes the observation that the use of the aggregated posterior as a VAE prior yields poor performance and massive overfitting, hence remedies such as learning a limited number of pseudo inputs were proposed [3]. This paper shows that the aggregated posterior is indeed an excellent prior for VAEs when simple yet **novel regularizers** are used. We propose a **new log marginal likelihood lower bound** based on kNN retrieval to enable the use of a large number of mixture components in the VAE's prior. Importantly, we propose a **novel application of VAEs to data augmentation** to improve supervised learning and show that the nonparametric nature of the Exemplar VAE is effective for improving classification error. Demonstrating the effectiveness of generative data augmentation is important as no prior work on VAEs has shown similar gains. These contributions are significant for the generative modeling community as they bridge the gap between likelihood based generative models, nonparameteric exemplar based approaches, and data augmentation strategies.

R4: *"...not surprising that VampVAE (Exemplar VAE) beats VAE with a standard Gaussian prior. A better baseline is VAE with a Gaussian mixture prior ... studied for clustering and data visualization ..."* Note that VAE with a VampPrior [3] outperforms VAE with a Guassian mixture prior, so we compared to VampPrior. But we'll include a comparison with MoG prior in the final revision.

R2: *"...I tend to see the introduced exemplar-based prior as a way of constraining model towards training data and reducing generalization"* Our empirical results suggest that Exemplar VAEs outperform predominant VAE variants in terms of generalization to held-out test sets as measured by log-likelihood. An intuitive explanation of this result is that learning to augment existing examples into new ones is easier than learning to generate examples from scratch.

R2: *"did a great job introducing regularization techniques, but might it be that these techniques would also boost original VAE and VAE with VampPrior?"* Good question. Our most important regularizer, leave-one-out, does not apply to parameteric priors such as VamPrior. Our preliminary experiments suggest that exemplar subsampling in a VampPrior has a similar effect to reducing the number of pseudo-inputs, but we will include an ablation and a discussion in the paper to address this issue.

R2, R3: *"...not discussed how one should choose k and M hyperparameters"* Section 5.1 (line 223) includes an ablation study to analyze the impact of M proportional to the dataset size N. We conclude that $M = N/2$ is a reasonable choice. We select k based on our computational budget to match the training cost of related work. Ignoring computation cost, a larger value of k is preferred.

R2: *Misc.* Thank you for such detailed comments. 1) To replicate VampPrior's results, we used the publicly available official repository, which gives consistent numbers on dynamic MNIST, but results in some discrepancy on Omniglot. The use of a small validation set (2K images) for early stopping can explain the discrepancy on Omniglot, and it is possible that VampPrior used a different procedure for early stopping on Omniglot. Nevertheless, our comparison is fair and all techniques use the same early stopping, training, and validation procedures. Importantly, note that for the ConvHVAE architecture, which has the best likelihood numbers, our replicated VampPrior numbers are better than the original paper. 2) As observed by prior work [4], VampPrior on CelebA didn't converge to a good solution in our experiments, which is the reason we didn't report VampPrior's numbers. The problem may be due to the initialization of pseudo inputs, which is a limitation of VampPrior. It's common to decrease the temperature of the model to improve sample quality and FID scores. 3) The choice of $M < N$ results in a consistent improvement on MNIST and Omniglot, so exemplar subsampling is helpful. 4) Agreed that Exemplar VAE augmentation can be combined with other approaches to reach a better classification error. MLPs are not competitive on permutation invariant MNIST without label smoothing. We'll fix the typos and include references to sections of the appendix. Part 9 of the appendix presents the pseudo-code.

R1: Thank you for bringing Graves *et al.* (2018) to our attention. We'll discuss in the final revision. Graves *et al.* learn an ordering of the data points focusing on autoregressive decoders to decrease the description length of transmitted codes. They propose a conditional prior that is difficult to compare against without an ordering. They also propose an unconditional prior that does not yield any likelihood gains. By contrast, we define a generic prior, which can be used to define an ordering to achieve the goal of Graves *et al.* and provides likelihood gains. Our unsupervised classification score outperforms Graves *et al.* on MNIST (98.5 *vs.* 98.87), which suggests our representations are higher quality. Finally, the summary of our contributions above is orthogonal to Graves *et al.*

R1: *"wall-clock time"* The cost of training Exemplar VAE is similar to VampPrior when the number of exemplars per minibatch is equal to the number of pseudo inputs, *e.g.,* for ConvHVAE on Omniglot with a minibatch size of 100 on a single GPU, VampPrior with 1000 pseudo inputs takes *58s/epoch* and Exemplar VAE with 10-NNs takes *51s/epoch*. ConvHVAE on MNIST & FashionMNIST with VampPrior takes *82s/epoch vs. 107s/epoch* for Exemplar VAE, since VampPrior uses 500 pseudo inputs here.

R1: *present samples from CelebA, but no bpd is reported...* ELBO numbers for CelebA are reported in the appendix, part 2. We can transform these numbers to bpd if that's more desirable. Also, there is some similarity between pseudo-likelihood and leave-one-out in Exemplar VAE, but pseudo-likelihood is one dimension given the rest, whereas leave-one-out is one example given the rest.

R3: *Misc.* The generation process is explained at the beginning of Section 3. For exemplar data augmentation we indeed sample from $r(z \mid x_i)$. CelebA has close to 200k 64x64 images, so we validated the effectiveness of method on a decent scale dataset.

References

[1] Bornschein, Mnih, Zoran, and Rezende. Variational memory addressing in generative models. *NIPS*, 2017.

[2] Hoffman and Johnson. ELBO surgery: yet another way to carve up the variational evidence lower bound. *NIPS Workshop*, 2016.

[3] Tomczak and Welling. VAE with a VampPrior. *AISTATS*, 2018.

[4] Xu, Luo, Henao, Shah, and Carin. Learning autoencoders with relational regularization. *arXiv:2002.02913*, 2020.