Author Response for "Generating Diverse High-Fidelity Images with VQVAE-2"

We thank the reviewers for the detailed and constructive feedback. All reviewers were impressed by the quality of our samples. R2 had positive remarks about the significance of our method and the thoroughness of our evaluation. R3 was satisfied with the clarity of the writing. There were however some concerns about additional results, architecture details and novelty of the approach. We hope that we have addressed the requested clarifications below, explaining how each will be improved in the final paper. We believe these clarifications will resolve all reviewers' concerns, but would be happy to consider any additional suggestions.

General response regarding to:

R1 - *Name of the method*: The name VQ-VAE was coined by previous authors, so we decided to adopt the same name, rather than using a new one. Moreover, it is possible to frame VQ-VAE in a variational framework as well (with a delta posterior and a uniform prior), which is discussed in the original VQ-VAE paper.

R1 - *h_bottom encoding everything*: **A** h_bottom is preceded by an Information-Bottleneck (similar to the KL of a regular VAE) so every latent can at most encode $\log_2(N)$ bits, where N is the size of the VQ codebook. In our case, this is 9bits per latent. There is one latent for every 16 pixels, each having 3 color channels with 8 bits each, so this yields a compression factor of 4*4*3*8/9=42.67. **B** Adding h_top results in much better reconstruction MSE and sharper reconstructions. We also visualize reconstruction from h_top only (Fig. 3 in the paper and Fig. 5 in the appendix), showing that h_top has indeed encoded quite a lot of information.

R1 - *Model remembering data*: The VQ-VAE is able to reconstruct test set images equally well as training images (MSE, and qualitatively), which demonstrates that it generalizes to unseen data. Similarly, for the PixelCNN, the NLLs of test data are comparable to those of training data.

R1 - *Interpolations*: There is indeed no simple way to do interpolations, which We will clarify in the final version.

R1 - *Speed*: Generation is slower than GANs, but faster than other autoregressive approaches that model images in the pixel space (about 45x faster). We also implemented incremental sampling (as in Paine et al. arxiv.org/abs/1611.09482) to cache intermediate activations that can considerably reduce sampling time. We will add a comparison in the final version.

R1 - *Objective with stop-gradient not elegant*: As noted in the paper, in equation 3, we use the Exponential Moving Average version which does not use stop-gradients. The loss in equation 2 is included for sake of completeness. These are both different neural implementations of the K-means algorithm, which has a long established track record in many areas of machine learning. We would also like to point out that elegance is a subjective matter, and simplicity is a form of elegance we strove for in this work: indeed, VQ-VAE is quite simple and can be implemented in just a few lines of code.

R2 - *Detailed architecture*: We agree that the architecture description could be more detailed. We will make sure that our architecture is thoroughly specified in our final version and we will include all details and hyperparameters.

R2 - We will fix minor details, also cite [A], [B] to emphasize the connection with lossy-compression.

R3 - *Novelty / This paper is not the first to achieve that*: We are not aware of any prior works that show comparable sample quality to BigGAN (which is SOTA) while having better diversity in any model class (let alone among likelihood-based methods). For faces, the best prior works (ie, Glow and SPN) used a much simpler dataset CelebaHQ with 256x256 resolution. Still their their samples look less realistic and have lower fidelity than our 1024x1024 samples from more complex FFHQ. The only other model to achieve this has been StyleGAN, which also has the discussed diversity limitations of GANs.

R3 - *Ablations wrt. model size*: We will add ablations of our model wrt. model size and batch size, but the result is that larger models get better results. Comparison with other works, however, shows that scaling up is necessary but not sufficient: our model gets better results compared to models with similar (or larger) size, batch-size, and compute requirements: Menick et al. 2018, Defauw et al. 2019. The same applies for BigGAN.

R3 - *No quantization*: The model does not work at all without quantization. We will add this ablation in the Appendix (with a reference from the main text).

R3 - *BPD measurements*: As R2 has noted, this model is inspired by lossy-compression where performance is usually characterized with rate-distortion curves. We apply log-likelihood based methods in a compressed lossy space, thus not having to model imperceptible details in images. This is where benefits like speed, global coherence, etc., come from. We do report BPD in the latent spaces. Trying to go back to the pixel-domain would defeat the main purpose of the method. That said, if there is truly interest in this metric, it is straightforward to estimate and add it to the final version of the paper.

R3 - *Classifier based rejection sampling*: All samples in the paper are without the rejection sampling except for Fig 8 and 9 in the appendix where we aim to illustrate the effect of various rejection thresholds. Similarly the numbers in Table 1 do not use this. CBRS is only used for the P/R and FID/IS curves in Fig. 5. of the main text.

R3 - *Nearest Neighbours*: As noted in the paper and in our response to R1, our model can be directly assessed for overfitting by comparing train and test NLL. Nevertheless, we will include nearest neighbours in the pixel and VGG spaces in the final version.