Supplemental: Investigating Permutation-Invariant Discrete Representation Learning for Spatially Aligned Images

Jamie Stirling[1], Noura Al-Moubayed[1], and Hubert P. H. Shum[1]

Durham University, United Kingdom

1 Derivation of Information Capacities

1.1 Nearest-Neighbour VQ with Limited per-Image Codebook Usage

In our experiments, we observed that applying naive nearest-neighbour vector-quantization [1] alongside PI-VQ results in a severely diminished per-image codebook usage K_{img}, the maximum of which over the whole training set was 49 (where the technical maximum is 512) when training on CelebA-HQ. This results in many repeated codes (since each image representation is of total length 512), and hence redundancy in the discrete latent representation, which severely restricts the capacity of PI-VQ information bottleneck.

Here we derive a precise upper limit on information capacity when using nearest-neighbor VQ, with an overall codebook usage (over the entire dataset) of K_{data}, representation length L, and maximum per-image codebook usage K_{img}. We use the same notation as in the main text.

In the best (maximum information) case, when representing an input image chosen at random from the dataset, the model selects K_{img} codebook elements at random, without replacement, from the K_{data} codebook elements. The space $\mathcal{S}$ of possible values that this "working subset" can take on is of cardinality:

$$|\mathcal{S}| = \binom{K_{data}}{K_{img}} \tag{1}$$

This is equivalent to stating that there are "K_{data} choose K_{img}" ways to choose K_{img} unique elements from a set of K_{data} unique elements.

Next, given a fixed working subset of length K_{img}, we determine the cardinality of the space $\mathcal{R}$ of representations of length L consisting of only K_{img} unique elements as:

$$|\mathcal{R}| = \binom{L + K_{img} - 1}{K_{img} - 1} \tag{2}$$

This is equivalent to the number of ways one can distribute L "bars" among K_{img} "stars" using the combinatoric "stars-and-bars" graphical aid [2].

Finally, we give an upper bound on the cardinality of the space $\mathcal{M}$ of all possible representations as:

$$|\mathcal{M}| < |\mathcal{S}| \times |\mathcal{R}| = \binom{K_{data}}{K_{img}} \times \binom{L + K_{img} - 1}{K_{img} - 1} \tag{3}$$

We note that $|M|$ is strictly smaller than the product of the two because there exist instances of overlap between representations from two distinct "working sets", e.g. two distinct working sets both containing codebook entry c can both represent an image as c repeated L times. This inequality is not an issue for our derivation since we are deriving an upper bound.

The upper bound on information capacity, in bits, is then:

$$\text{Capacity} = \log_2 \left[|\mathcal{S}| \times |\mathcal{R}| \right] \tag{4}$$

$$= \log_2 \left[\binom{K_{data}}{K_{img}} \times \binom{L + K_{img} - 1}{K_{img} - 1} \right] \tag{5}$$

1.2 Improved Capacity using Matching Quantization

The proposed matching quantization approach ensures that representations of length L always use exactly L unique elements from the overall codebook. The space $\mathcal{M}$ of possible representations under this approach is then of cardinality:

$$|\mathcal{M}| = \binom{K_{data}}{L} \tag{6}$$

This is because representing a given image is equivalent to choosing L elements from a set of size K_{data}, to achieve which there are exactly "K_{data} choose L" possible ways. We note this is equivalent to the naive quantization case with K_{img} fixed at L. Thus the associated information capacity in bits is:

$$\text{Capacity} = \log_2 |\mathcal{M}| \tag{7}$$

$$= \log_2 \left[\binom{K_{data}}{L} \right] \tag{8}$$

2 Training Details

We train the PI-VQ run for 1,400,000 iterations for the 256x256 datasets and 400,000 iterations for CelebA 64x64. Following earlier work with VQ-GAN [3, 4], we begin adversarial training part-way though the training run (iteration $500,000$ onward, or $100,000$ onward for CelebA 64x64).

Codebook initialization and reset: We set $T_q = 60,000$ for FFHQ and CelebA-HQ. Due to the smaller overall number of training iterations on CelebA 64x64, we set $T_q = 25,000$. In each case, T_q is chosen to take place before the onset of adversarial training, and after at least one complete iteration through the entire training set.

References

1. A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf
2. P. Flajolet and R. Sedgewick, *Analytic combinatorics*. cambridge University press, 2009.
3. P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.
4. S. Bond-Taylor, P. Hessey, H. Sasaki, T. P. Breckon, and C. G. Willcocks, "Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 170–188.