Project Machine Learning
— Milestone 1 —

Konstantin Ausborn, Timon Palm, Marco Rosinus Serrano

November 16, 2024

1 Introduction

2 Vector-Quantized variational Autoencoder (VQ-VAE)

3 Dataset overview

Vectore-quantised variational autoencoders (VQ-VAE) are applicable to different modalities, like image, video or audio generation. By virtue of comparison with its predecessors (Autoencoder, VAE) and the paper itself, we will first focus on image generation only.

The performance of the VQ-VAE was showcased on three different image datasets: *ImageNet*, *CIFAR-10* and Video frames from *DeepMind Lab*.

In this project, we will train and test our implementation on ImageNet as well as CIFAR-10, which makes our results comparable to the original paper but also to other generative architures.

Image generation is an unsupervised learning task and therefore does not require labeled data. Potentially, any kind of image data could be used for training and testing. However, the quality of the generated images is highly dependent on the quality and diversity of the training data.

ImageNet and CIFAR-10 are among the most common image datasets. They are both created for image classification tasks, but we can also leverege their capacity by simply dropping their labels. Table 1 gives an overview of the datasets.

3.1 ImageNet

ImageNet is a widely used image datasets in the field of machine learning, primarily for the task of classification and detection. The full dataset consists of 14.197.122 hand-labeled photographs collected from flickr and other search engines Russakovsky et al. (2015). The images are distributed over 21841 *synonym sets* from the *WordNet* Fellbaum (1998) hierarchy.

When talking about ImageNet, most authors refer to the *ImageNet Large Scale Visual Recognition Challenge 2012* (ILSVRC) dataset Russakovsky et al. (2015), which is a subset of the full dataset. It contains 1.281.167 unique labeled training images and 100.000 labeled test images, which are distributed over 1000 classes. ILSVRC is an image classification/recognition challenge. It therefore additionally contains 50.000 validation images without labels as part of the benchmark. There are neither missing values nor duplicates in the dataset and every image belongs to exactly one class.

In the following, we will refer to the ILSVRC 2012 dataset as ImageNet if not stated otherwise.

As the number of samples and classes hint, the data can not be balanced. In fact, the number of images per class varies from 732 to 1300 images in the training set (1). Among the classes with the lowest number of images are *"black-and-tan coonhound"*, *"otterhound"* and *"English foxhound"* with 732, 738 and 754 number of samples, respectively.

Images in the ILSVRC dataset have a wide range of resolutions, from 75x56 pixels to 4288x2848 pixels and an average resolution of 469x387 pixels. While technically a VQ-VAE can handle different images sizes (see section 2), we will preprocess all images to a fixed size for training and testing.

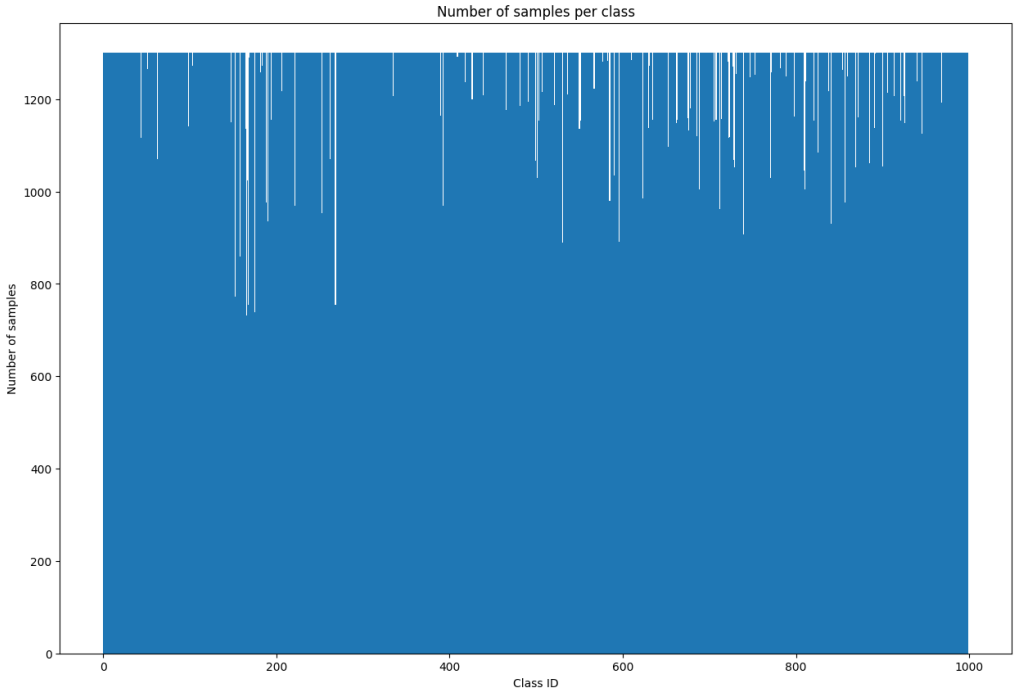

Figure 1: Distribution of images per class in the ImageNet dataset

Dataset	# train images	# test images	# classes	image size
ImageNet (ILSVRC)	1.281.167	100.000	1000	75x56px - 4288x2848px
CIFAR-10	50.000	10.000	10	32x32px

Table 1: Datset overview: ImageNet and CIFAR-10

3.1.1 Preprocessing

For training and testing, we will resize all images to 128x128 pixels, as it was also done in the paper. By virtue of the VQ-VAE architecture, this implies a latent space of 32x32x1 pixel.

There are different ways to resize the images, like cropping, padding or resizing. We will use random cropping to extract a square image and resize it to 128x128 pixels. This way, we can also handle smaller images without disturbing them or augmenting them with padding.

Although *data whitening/standardization* is not as important for Convoluational Neural Networks (CNNs), which VQ-VAE ultimately is build of, it is still a suggested procedure for ImageNet data. ImageNet whitening was also done for training the common *Resnet* architecture He et al. (2015) on ImageNet. Furthermore, it is in the standard tensorflow and pytorch preprocessing pipeline for ImageNet class.

Example images from ImageNet dataset are depicted in figure 2 (before whitening) and 3 (after whitening).

3.2 CIFAR-10

CIFAR-10 Krizhevsky (2012) is another popular image classification dataset, as well as the larger version CIFAR-100. CIFAR-10 consists of 60.000 32x32 pixel images, which are distributed over 10 classes. The dataset is split into 50.000 training images and 10.000 test images. The classes are mutually exclusive, which means that each image belongs to exactly one class. The classes are: *airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck*

The train set contains exactly 5000 images per class, while the test set contains 1000 images per class.

There is no further preprocessing necessary for CIFAR-10, as the images are already 32x32 pixels and have a fixed size.

4 Baseline method and evaluation

- basically just use negative log entropy as the factor to compare
- this describes the entropy of the picture generated? I do not have clarity here see: https://bjlkeng.io/posts/a-note-on-using-log-likelihood-for-generative-models/
- I need to understand PixelCNN better to continue
- see Shannon for theory on entropyShannon (1948)
- dont use parzen windowsTheis et al. (2015)
-
-

5 Discussion

:-)

References

C. Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. URL `https://mitpress.mit.edu/9780262561167/`.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL `https://arxiv.org/abs/1512.03385`.

A. Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252, 2015. doi: 10.1007/s11263-015-0816-y.

C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

L. Theis, A. v. d. Oord, and M. Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

6 Appendix

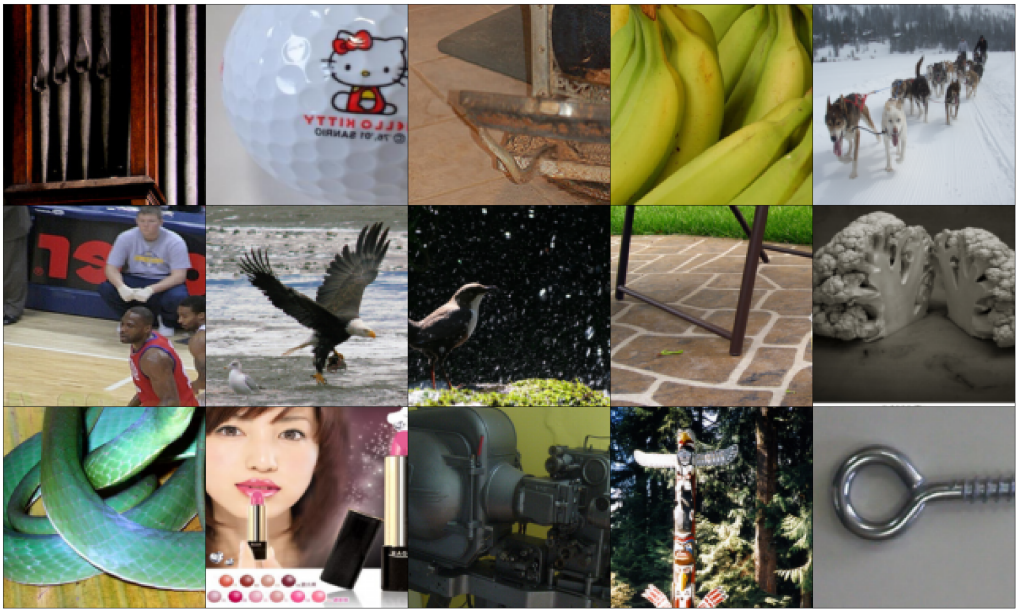

Figure 2: Example images from the ImageNet dataset randomly cropped and resized to 128x128 pixels

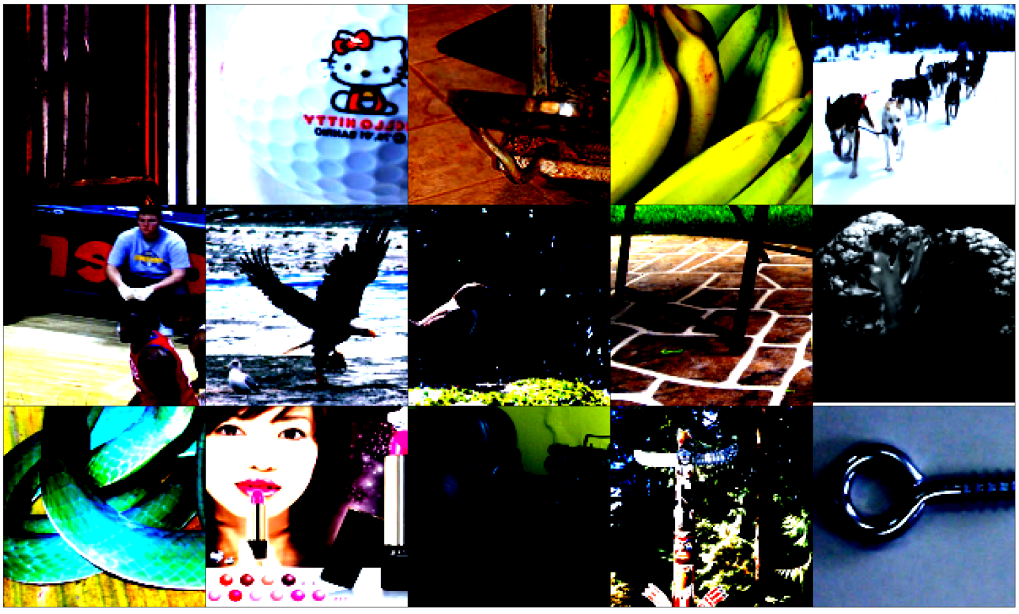

Figure 3: Example images from the ImageNet dataset randomly cropped and resized to 128x128 pixels and standardized