Step-unrolled Denoising Autoencoders for Fast Image Generation

Alex F. McKinney
Durham University
Durham, UK
alexander.f.mckinney@durham.ac.uk

Chris G. Willcocks
Durham University
Durham, UK
christopher.g.willcocks@durham.ac.uk

Abstract

An empty abstract

1. Introduction

foobar

2. Related Work

foobar

3. Method

3.1. Latent Dataset Generation

We use the standard two-stage scheme for vector-quantized image modelling [1,2,5,6] using VQ-GAN [2] as our feature extractor. Where such models are available, we use pretrained VQ-GANs for our experiments. For higher resolution experiments (for example, FFHQ-1024 [3]), pretrained models are not available and so training our own VQ-GAN was necessary (see §3.6).

The second stage is to learn a discrete prior model over these latent variables. To enable this, we must first build a latent dataset using our trained VQ-GAN. Formally, given a dataset of images $\mathcal{X}$, a VQ-GAN encoder E with downsample factor f, and vector-quantization codebook $\mathcal{C}$ trained on $\mathcal{X}$, we define our latent dataset $\mathcal{L}$ as:

$$\mathcal{L} = \{\mathcal{C}(E(\mathbf{x})) \mid \mathbf{x} \in \mathcal{X}\} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ is a single element of the image dataset and $\mathbf{z} = \mathcal{C}(E(\mathbf{x})) \in \{1, \ldots, |\mathcal{C}|\}^{h\times w}$ is the corresponding discrete latent representation. In other words, each $f \times f$ pixels in $\mathbf{x}$ is mapped to a single discrete value from 1 to $|\mathcal{C}|$ (which in turn, corresponds to a vector $\mathbf{e} \in \mathcal{C}$), resulting in a latent representation of shape $\frac{H}{f} \times \frac{W}{f} = h \times w$.

We then use $\mathcal{L}$ to train a discrete prior over the latents. Coupled with the VQ-GAN decoder G, we obtain a powerful generative model.

3.2. 2D-Aware Hourglass Transformer

Inspired by successes in hierarchical transformers for generative language modelling [4], we modify their architecture for use with discrete latent representations of image data. We will later use this architecture as the discrete prior over the VQ-GAN latents.

Hourglass transformers have been seen to efficiently handle long-sequences, outperform existing models using the same computational budget, and meet the same performance as existing models more efficiently by using an explicit hierarchical structure [4]. The same benefits should also apply to vector-quantized image modelling.

Our modifications are 2D-aware downsampling, axial rotary embeddings, and removal of causal modelling constraints.

2D-Aware Downsampling

foobar

Axial Rotary Embeddings

foobar

Removal of Causal Constraints

foobar

3.3. Non-Autoregressive Generator Training

3.4. Generating High-Resolution Images

3.5. Arbitrary Pattern Inpainting

3.6. Training a megapixel VQ-GAN

References

[1] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P. Breckon, and Chris G. Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes, 2021. 1

[2] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021. 1

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. 1

[4] Piotr Nawrot, Szymon Tworkowski, Michał Tyrolski, Łukasz Kaiser, Yuhuai Wu, Christian Szegedy, and Henryk Michalewski. Hierarchical transformers are more efficient language models, 2021. 1

[5] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2, 2019. 1

[6] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. 1