MEGAPIXEL IMAGE GENERATION WITH STEP-UNROLLED DENOISING AUTOENCODERS

Alex F. McKinney[1], **Chris G. Willcocks**[1,2]

[1]Department of Computer Science, Durham University
[2]Project Supervisor

Figure 1: 1024×1024 samples from our FFHQ1024 model. Resulting samples are diverse and of high-fidelity. Each sample was generated in ≈ 2 seconds on a consumer-grade GPU (GTX 1080Ti), in contrast to existing approaches at this resolution, which take minutes to generate. To our knowledge, this is the fastest sampling, non-adversarial generative framework at this resolution.

Background

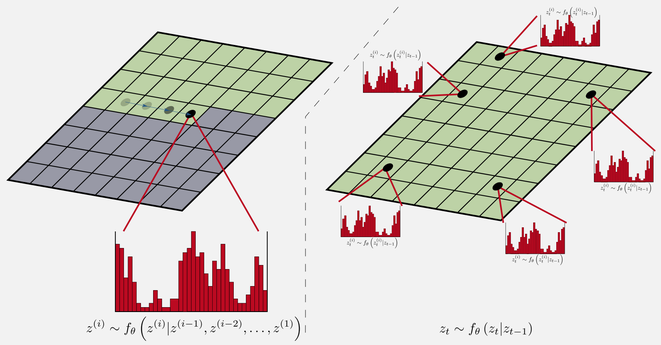

Figure 2: Autoregressive sampling (left) is defined in terms of the probabilistic chain rule, so sampling is done iteratively with complexity $\mathcal{O}(n)$. Non-autoregressive sampling (right) can sample an arbitrary number of elements in parallel, so number of steps does not scale with input size. It can also use the full context available to it, resulting in better quality samples and flexible inpainting. However, existing NAR methods still typically require many network evaluations to produce meaningful samples.

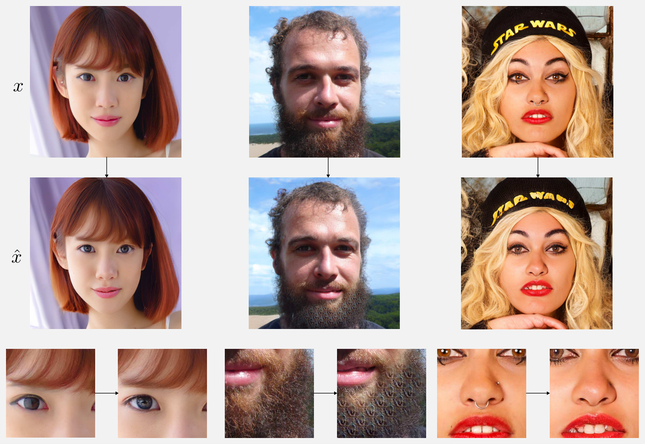

Figure 3: VQ-GAN is used to reduce computational requirements for training and sampling of generative models by acting as a compression model. VQ-GAN does not always faithfully reproduce the input image, though outputs are perceptually valid. For example, left shows a change in eye colour and the concealment of a piercing by adjusting hair position. Middle has its hair texture changed. Right has piercings removed and text corruption.

Proposed Method

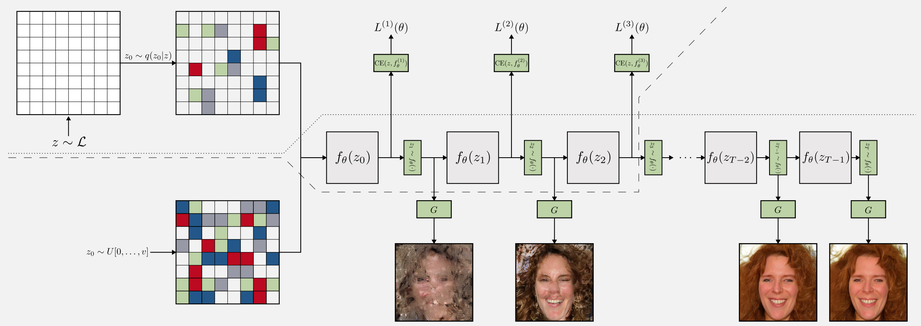

Figure 4: An overview of the SUNDAE training and sampling of discrete latent representations. Above the dashed line represents the process for training, whereas below the dotted line represents the sampling process. The training process begins by sampling $\mathbf{z} \sim \mathcal{L}$ and then sampling from the corruption distribution $q(\mathbf{z}_0|\mathbf{z})$. SUNDAE then denoises for 2 to 3 steps, computing the cross-entropy loss at each step in the chain which is subsequently averaged to produce a final loss. Sampling begins by obtaining $\mathbf{z}_0$ from a uniform prior and iteratively denoising. SUNDAE easily outspeeds prior autoregressive and (non-adversarial) non-autoregressive models, using typically 50-100 steps ($\approx$ 2 seconds) during sampling.

$$x \tag{1}$$

$$x \tag{2}$$

The objective of this project was to push the efficiency of generative models to their limit and expand the array of non-autoregressive methods for image generation. By combining various methods – each at the pinnacle of efficiency in their respective areas. These include VQGAN, a vector quantization model with an unparalleled rate of compression; hourglass transformers, a highly scaleable attention model; and SUNDAE, a fast, non-autoregressive text generative model.
Our primary contributions are as follows:

- The development of a non-autoregressive, non-adversarial generative modelling framework with extremely flexible sampling including self-correction and arbitrary inpainting pattern capabilities. The model can be directly configured for both low- and high-step sampling scenarios, resulting in high quality and diverse samples in mere seconds of sampling time.
- Modifications to SUNDAE and hourglass transformers to be more suited for the modelling of multi-dimensional discrete data. Though applied to discrete latents in our work, the modifications are also applicable in a wider context, such as to pixel-level modelling. We also demonstrate the superiority of hierarchical transformers – forming a key component in the scalability of our approach.
- The scaling of VQGAN to extremely high resolution images of human faces. 1024×1024 images far exceeds resolutions achieved in prior work. This ultimately allowed for the **generation of megapixel images in as few as two seconds** on a consumer-grade GPU when combined with our fast and scalable generative framework. This is in contrast to prior autoregressive methods and non-autoregressive diffusion methods that take minutes to generate, or have not scaled to such resolutions entirely.

Results

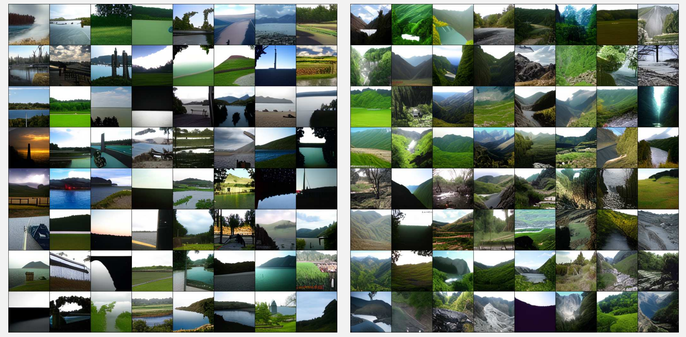

Figure 5: 256×256 class-conditioned samples on ImageNet. The left batch of samples are from the class “Lakeside” whereas the right batch are from the class “Valley”. We note, that our approach can be extended to use text prompts as a conditioning signal, yielding a text-to-image model.

Figure 6: Representative inpainting results on FFHQ1024 using our trained model. We demonstrate the superiority of NAR methods for inpainting by using arbitrary inpainting masks, including completely random (left) and large block masks (right). Such patterns are difficult to inpaint with using autoregressive models, and cannot utilise the full context available to them.

References

[1] Patrick Esser, Robin Rombach, and Björn Ommer. *Taming Transformers for High-Resolution Image Synthesis*. 2021. arXiv: 2012.09841 [cs.CV].

[2] Piotr Nawrot et al. *Hierarchical Transformers Are More Efficient Language Models*. 2021. arXiv: 2110.13711 [cs.LG].

[3] Nikolay Savinov et al. *Step-unrolled Denoising Autoencoders for Text Generation*. 2022. arXiv: 2112.06749 [cs.CL].