Tanishq Gautam
05 Jul 21 • 4 min read

Overview

- Introducing a convolutional VQGAN, which learns a codebook of context-rich visual parts
- This approach is readily applied to conditional synthesis tasks, where both non-spatial information, such as object classes, and spatial information, such as segmentation which can control the generated image.

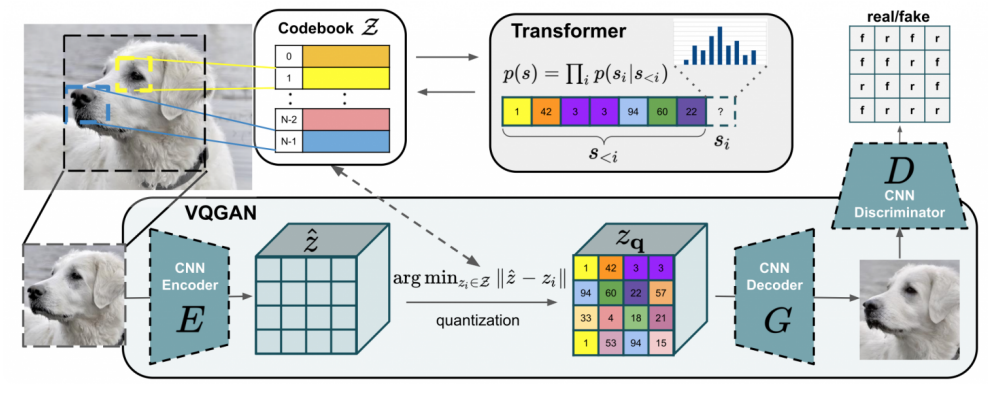
Introduction

Source

This approach enables transformers to synthesize high-resolution images like this one!

Transformers are on the rise and are taking over as the de-facto state-of-the-art architecture in all language-related tasks and other domains such as audio and vision. CNN's have shown to be vital but have been designed to exploit prior knowledge about strong local correlations within images whereas transformers are free to understand and learn the complex relationships among the inputs. One key goal is to obtain an effective and expressive model that combines convolutional and transformer architectures and can model the compositional nature of the computer visual world.

Understanding the Process



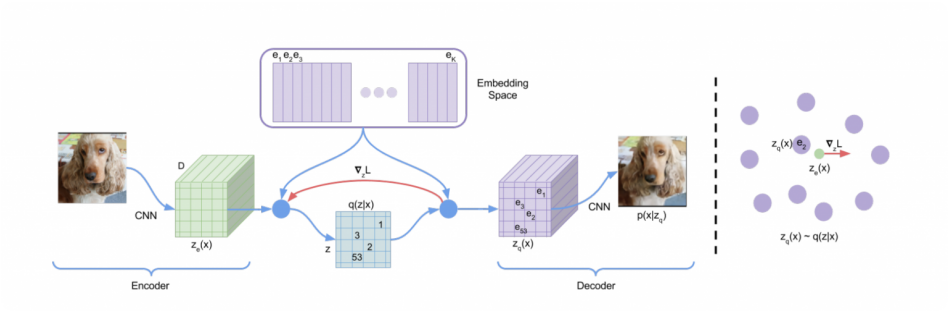
Source

modeled with an autoregressive transformer architecture. The codebook provides the interface between these architectures and a discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer-based high-resolution image synthesis.

To use transformers to synthesize higher resolution images the semantics of an image must be presented cleverly. Using pixel representation is not going to work as the number of pixels increases quadratically with a 2x increase in image resolution. Therefore, instead of representing an image with pixels, it is represented as a composition of perceptually rich image constituents from a codebook.

To do so, a new architecture VQGAN is proposed, a variant of the original VQVAE, and uses a discriminator and perceptual loss to keep good perceptual quality at an increased compression rate. Let's understand VQVAE a bit before diving any deeper.

Vector Quantized Variational Autoencoders (VQ-VAE)



Source

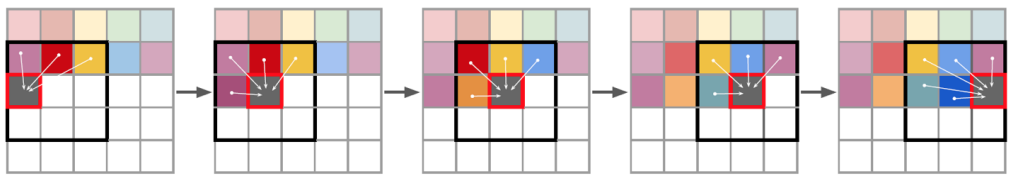
VQ-VAE consists of an encoder that maps observations/images onto a sequence of discrete latent variables, and a decoder that reconstructs the observations from these discrete variables. They use a shared codebook.

"latent space" which is simply a space of compressed image data in which similar data points are closer together. This is then quantized based on its distance to the code vectors such that each vector is replaced by the index of the nearest code vector in the codebook. The same is used by the decoder for reconstruction.

VQ-GAN

The authors here have used VQ-GAN, which is a variant of the original VQ-VAE that uses a discriminator and perpetual loss to keep good perceptual quality at an increased compression rate. It was in a two-step manner by firstly, training the VQ-GAN and learning the quantized codebook then training the autoregressive transformer using the quantized codebook as sequential input to the transformer.

Discrete Latent Representation is inspired by JPEG lossy compression of the image. JPEG encoding removes more than 80% of the data without noticeably changing the perceived image quality. Thus, training a generative model with less noise tends to work better.



Source

The attention mechanism of the transformer puts limits on the sequence length. While the number of downsampling blocks can be adapted by

To generate images in the megapixel manner, it has to be done patch-wise and crop images to restrict the length to a maximally feasible size during training. To sample images, then make use of the transformer in a sliding window manner as illustrated in the above figure.

The VQGAN ensures that the available context is still sufficient to faithfully model images, as long as either the statistics of the dataset are approximately spatially invariant or spatial conditioning information is available.

Ending notes

This approach addresses the fundamental challenges that previously confined transformers to low-resolution images by representing images as a composition of rich image constituents and thereby overcomes the quadratic complexity when modeling images directly in pixel space.

The approach taps into the full potential allow us to represent the first results on high-resolution image synthesis with a transformer-based architecture.

Taming Transformers | VQ-VAE | VQGAN

Tanishq Gautam
05 Jul 21 • 4 min read

Advanced | Computer Vision | Image | Image Analysis

Leave a Reply

What are your thoughts?...

Notify me of follow-up comments by email.

Notify me of new posts by email.

Submit reply