VQGAN+CLIP MODEL

Thursday, July 7, 2022 1:54 PM

VQGAN stands for Vector Quantized Generative Adversarial Network,

CLIP stands for Contrastive Image-Language Pretraining.

the way they work is that VQGAN generates the images, while CLIP judges how well an image matches our text prompt. This interaction guides our generator to produce more accurate images

VQGAN+CLIP is a combination of two neural network architectures: VQGAN and CLIP.

Basic idea:-

The text-to-image paradigm that VQGAN+CLIP popularized certainly opens up new ways to create synthetic media and maybe even democratizes “creativity”, by shifting the skillset from (graphical) execution or algorithmic instruction (programming) to nifty “prompt engineering”.

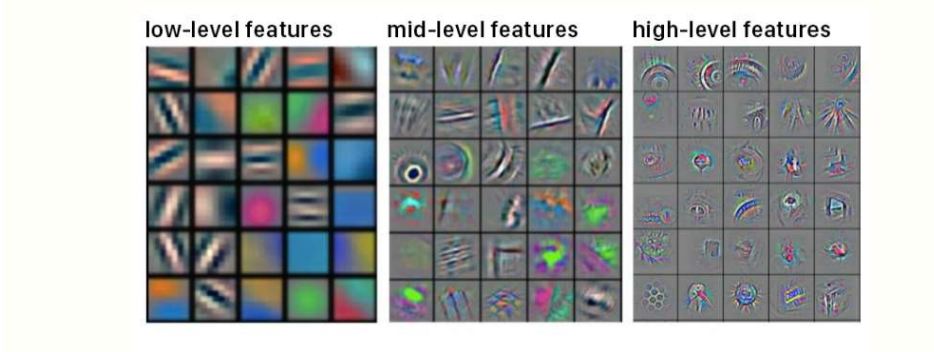
1. an interesting view of perception that allows us to model long-range dependencies by representing images discretely; and—
2. a pixel-based approach to learn local interactions and visual parts.

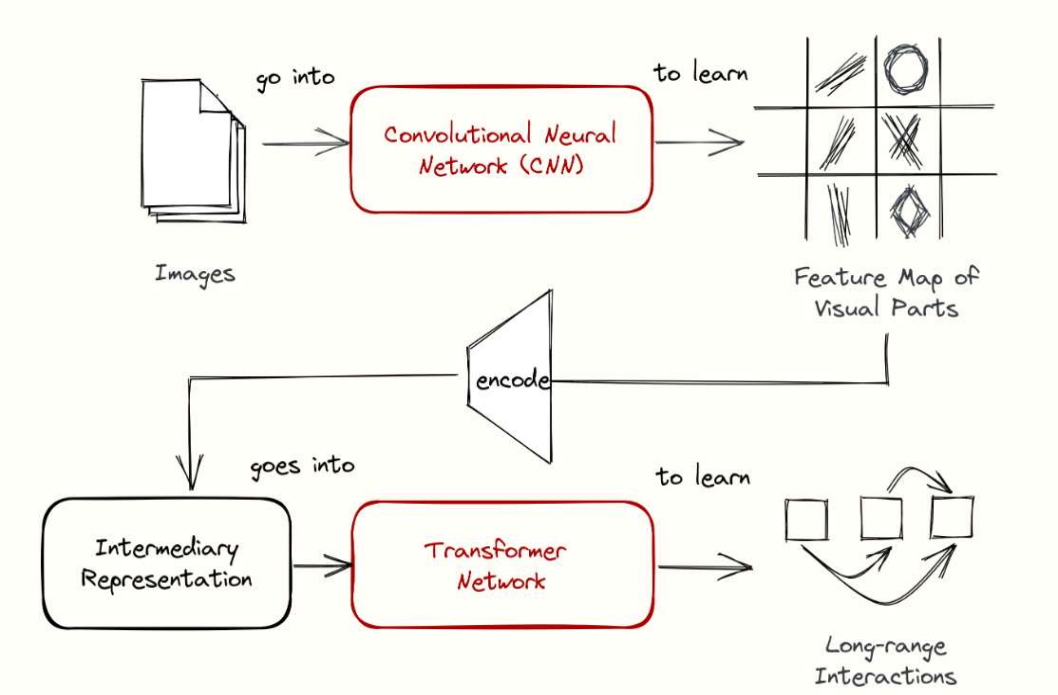
VQGAN was able to combine both of them. It can learn not only the (1) visual parts of an image, but also the (2) relationship (read: long-range dependencies) between these parts.
The codebook is generated through a process called vector quantization (VQ), i.e., the “VQ” part of “VQGAN.” Vector quantization is a signal processing technique for encoding vectors.

From <https://ljvmiranda921.github.io/notebook/2021/08/08/clip-vqgan/>

a CNN learned how to compose pixels at varying layers of abstraction: pixels become edges, edges become shapes, and shapes become parts.

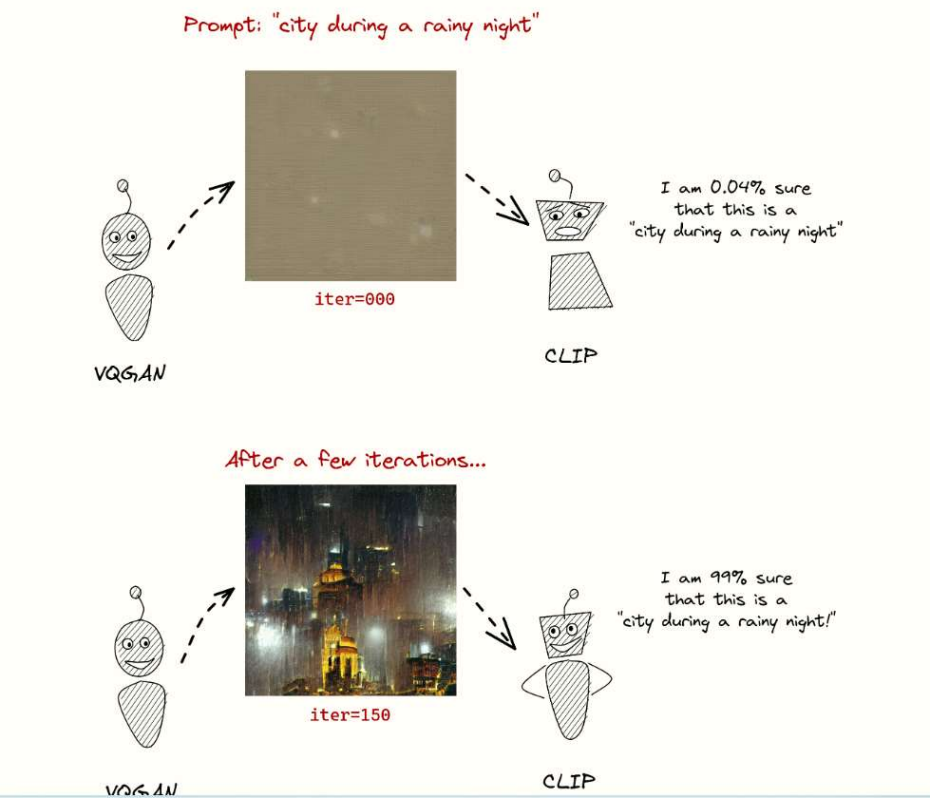
From <https://ljvmiranda921.github.io/notebook/2021/08/08/clip-vqgan/>





From <https://ljvmiranda921.github.io/notebook/2021/08/08/clip-vqgan/>

CLIP is the “Perceptor” and VQGAN is the “Generator”.



From <https://alexasteinbruck.medium.com/vqgan-clip-how-does-it-work-210a5dca5e52>