A small horizontal bar with a teal segment on the left and an orange segment on the right, positioned above the title.

Illustrative Session on Image Generative Models with Dall.E Mini

Karthik Desingu | Anirudh A | Karthik Raja A

Machine Learning Research Group, SSN College of Engineering

Session 3 *of the* Image and Video Analysis Workshop

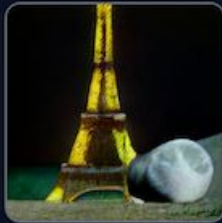
International Conference on Computational Intelligence in Data Science, 2023



AI model drawing images from any prompt!

the Eiffel tower landing on the moon

DRAW



Dall.E Mini — Text to Image

[Live Online Version of Dall.E Mini](#)



Part 1: Building Blocks of Dall.E Mini

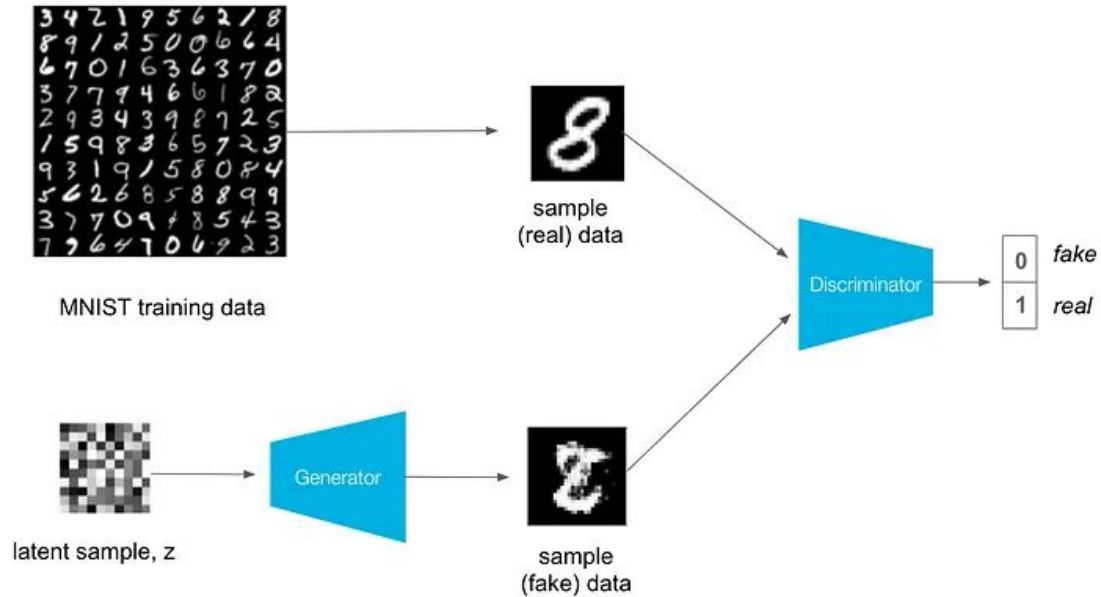
- **BART-based Encoder-Decoder:** Encodes captions as embedding vectors
- **VQ-GAN:** Decodes caption embeddings into Images
- **CLIP:** Evaluates caption-image relevance



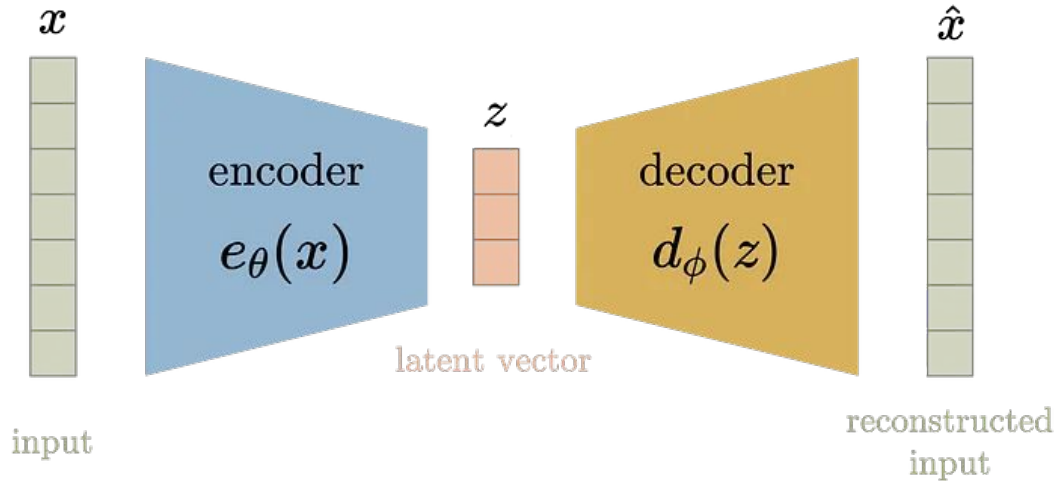
Part 2: Generative Adversarial Networks (GANs)

- Dall.E Mini uses a variant of GANs called **VQ-GANs**.
- The evolution of VQ-GANs,
 - Vanilla GAN
 - **Autoencoders** (AEs)
 - **Variational** Autoencoders (VAEs)
 - **Vector Quantized** Autoencoders (VQ-AEs)
 - Vector Quantized **GANs** (VQ-GANs)

Vanilla GAN



Autoencoder (AE)

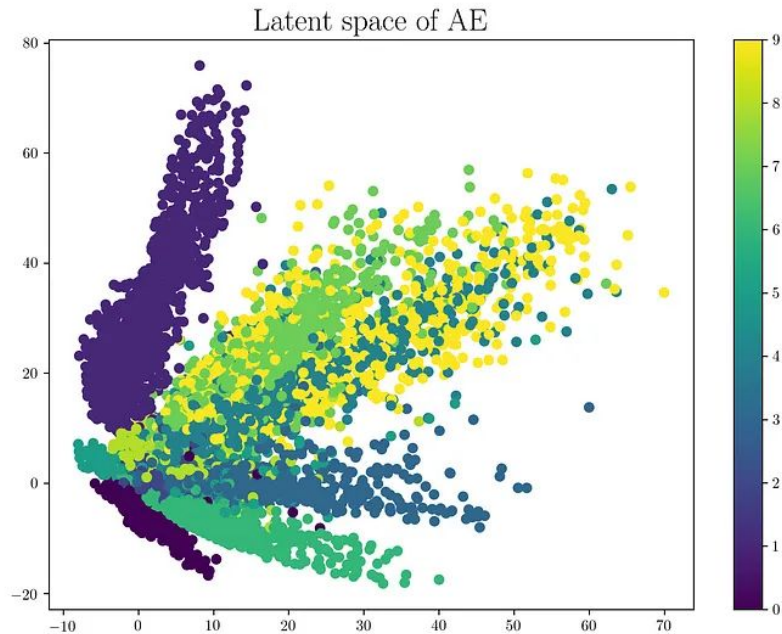


- The latent space is discontinuous and has significant “gaps”.

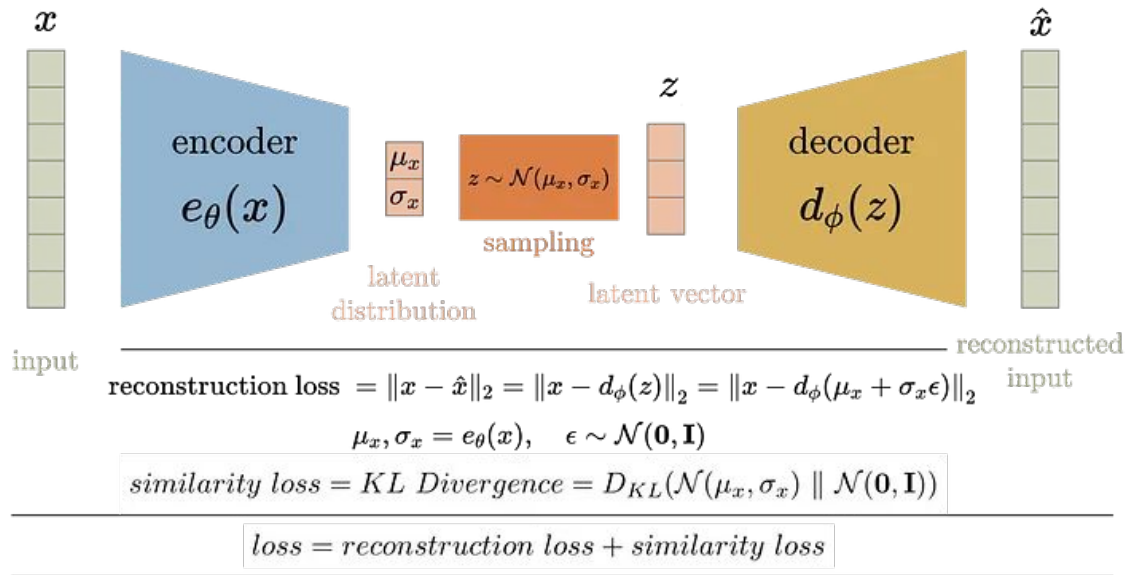
$$loss = \|x - \hat{x}\|_2 = \|x - d_{\phi}(z)\|_2 = \|x - d_{\phi}(e_{\theta}(x))\|_2$$



Autoencoder (AE)

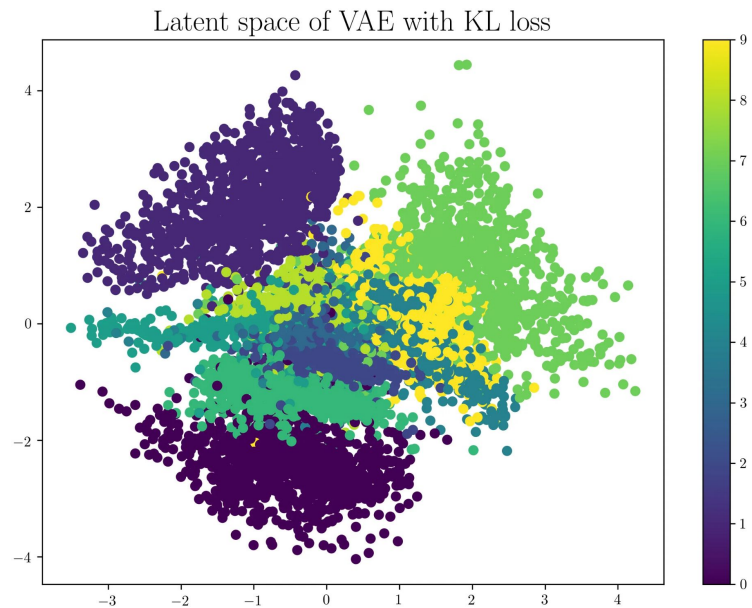


Variational Autoencoder (VAE)

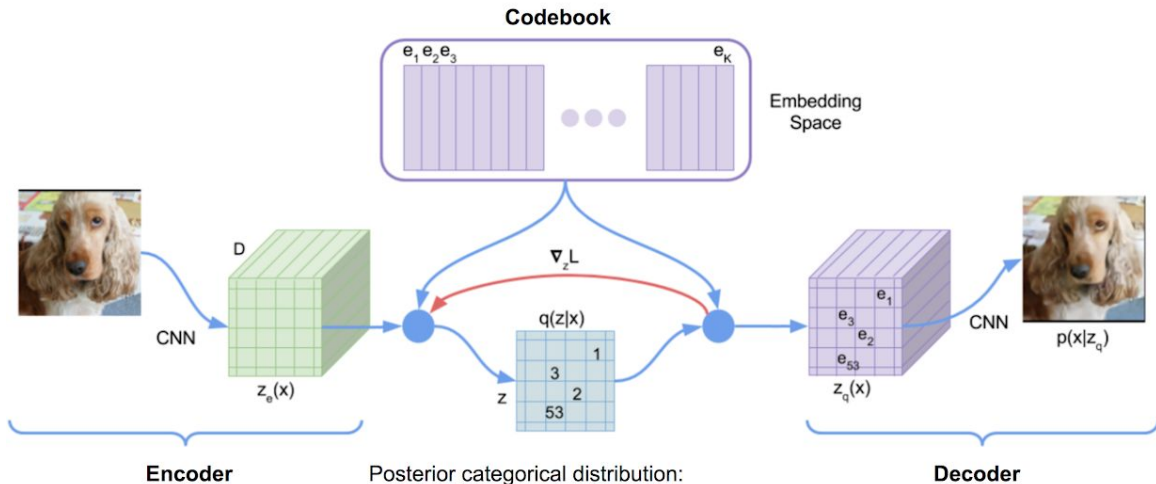


- The latent space is more cohesive – resembles the unit norm.
- Overlapping regions produce “morphed” images.

Variational Autoencoder (VAE)



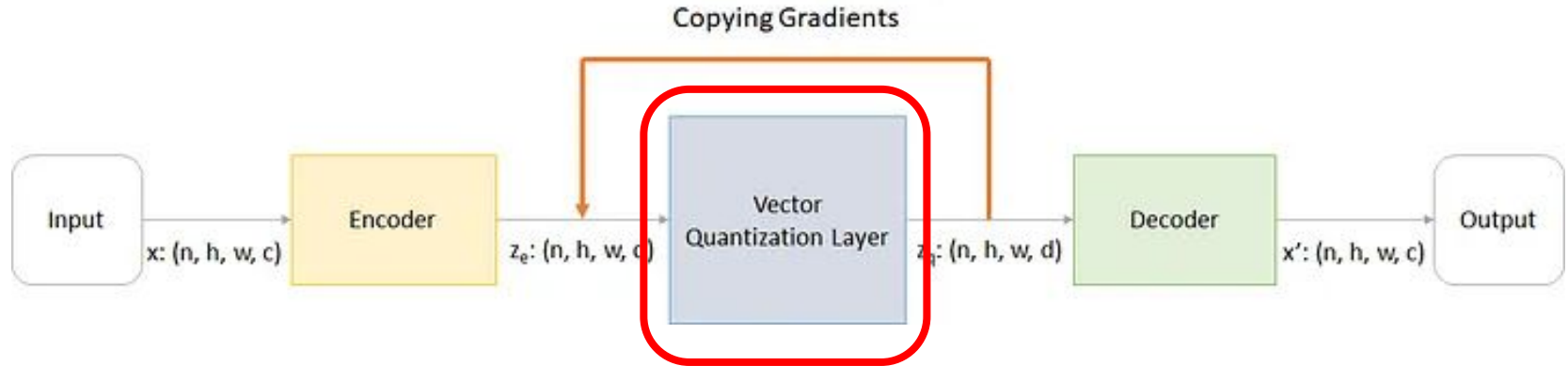
Vector-Quantized Variational Autoencoder (VQ-VAE)



- The latent space is discrete.
- No “morphed” outputs.
- Latent space has same dimensions as codebook.

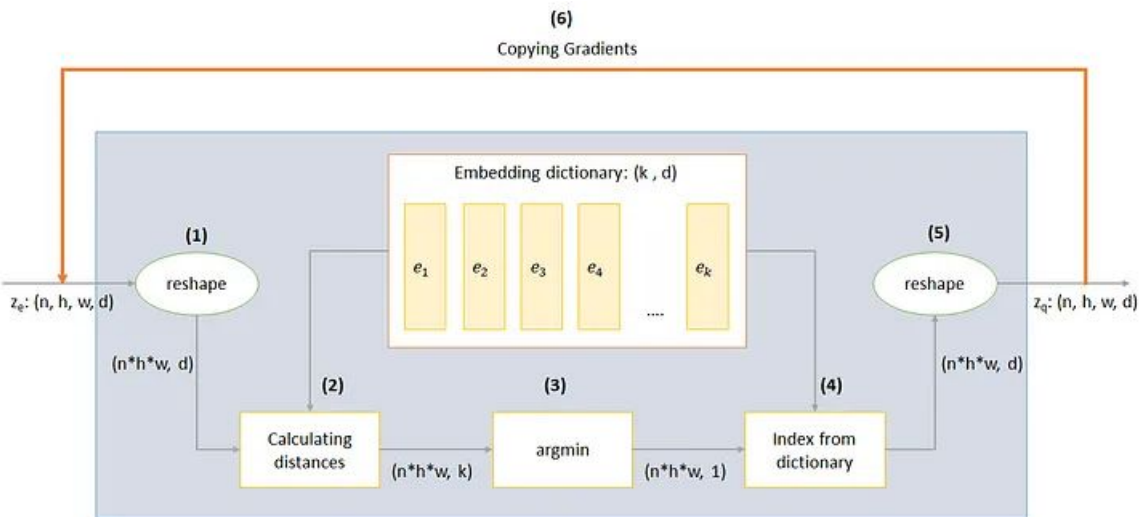
$$q(\mathbf{z} = \mathbf{e}_k | \mathbf{x}) = \begin{cases} 1 & \text{if } k = \arg \min_i \|\mathbf{z}_e(\mathbf{x}) - \mathbf{e}_i\|_2 \\ 0 & \text{otherwise.} \end{cases}$$

Vector-Quantized Variational Autoencoder (VQ-VAE)

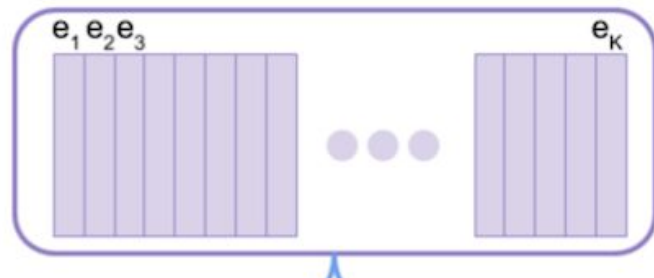


Adding the VQ layer to the AE

Vector-Quantized Variational Autoencoder (VQ-VAE)

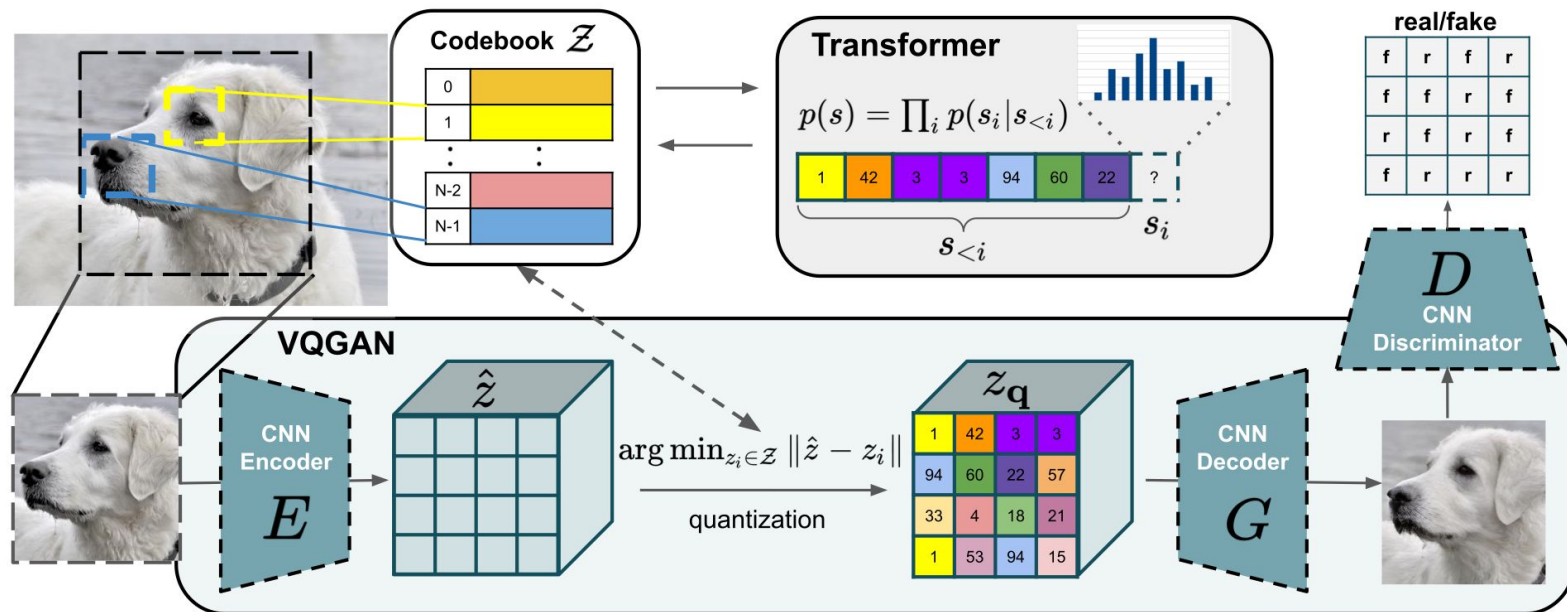


The Vector-Quantization Layer



The Codebook

Vector-Quantized Variational GAN (VQ-GAN)

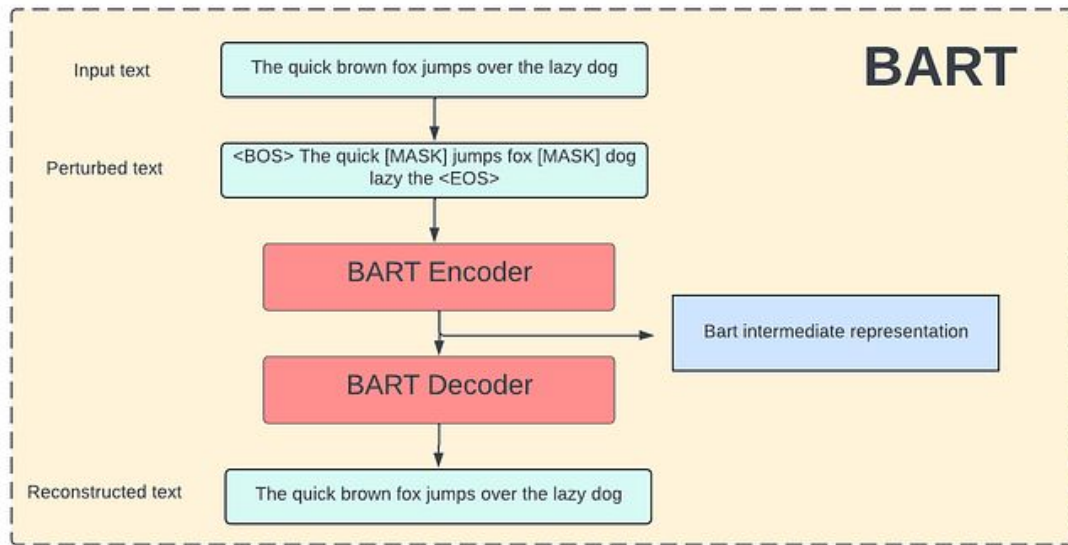




Part 3: BART Encoder-Decoder

- A BART model is pre-trained to “clean” text captions.
- For Dall.E Mini, the BART model **translates captions into the codebook vocabulary**.
- The codebook of VQ-GAN, in effect, maps text embeddings to image embeddings.

What BART does.



- Translates captions to codebook vocabulary.

Part 4: CLIP to Rank Images by Relevance



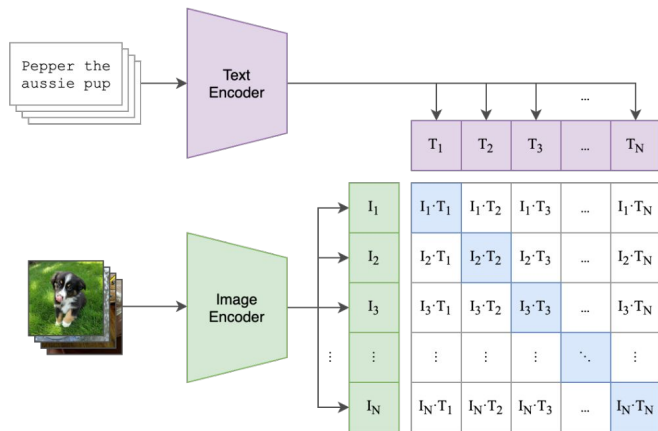
Python Code Demo

- CLIP is a neural network trained on a variety of (image, text) pairs
- It can be instructed in natural language to predict the most relevant text snippet, given an image (and vice versa), without directly optimizing for the task
- CLIP is thus similar to the zero-shot capabilities of GPT-2 and 3
- CLIP matches the performance of the original ResNet50 on ImageNet “zero-shot” without using any of the original 1.28M labeled examples

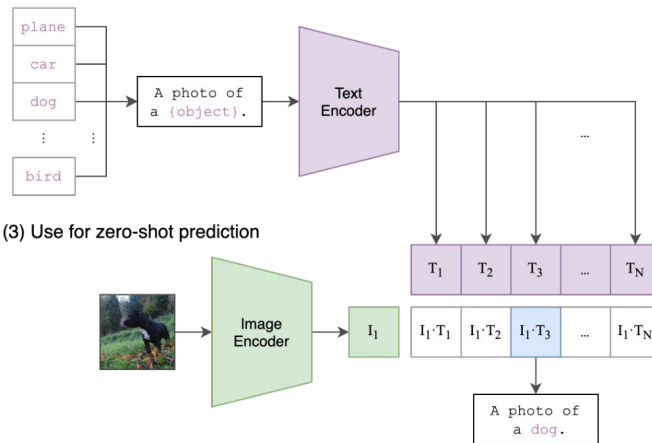
CLIP Architecture

Python Code Demo

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

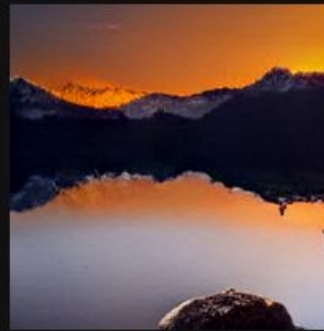
Contrastive pre-training is a type of self-supervised learning technique to learn representations of data that are useful for downstream tasks, such as image classification or natural language processing.

Relevance Scores

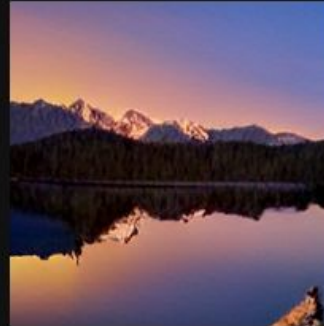
Prompt: sunset over a lake in the mountains



Score: 31.59

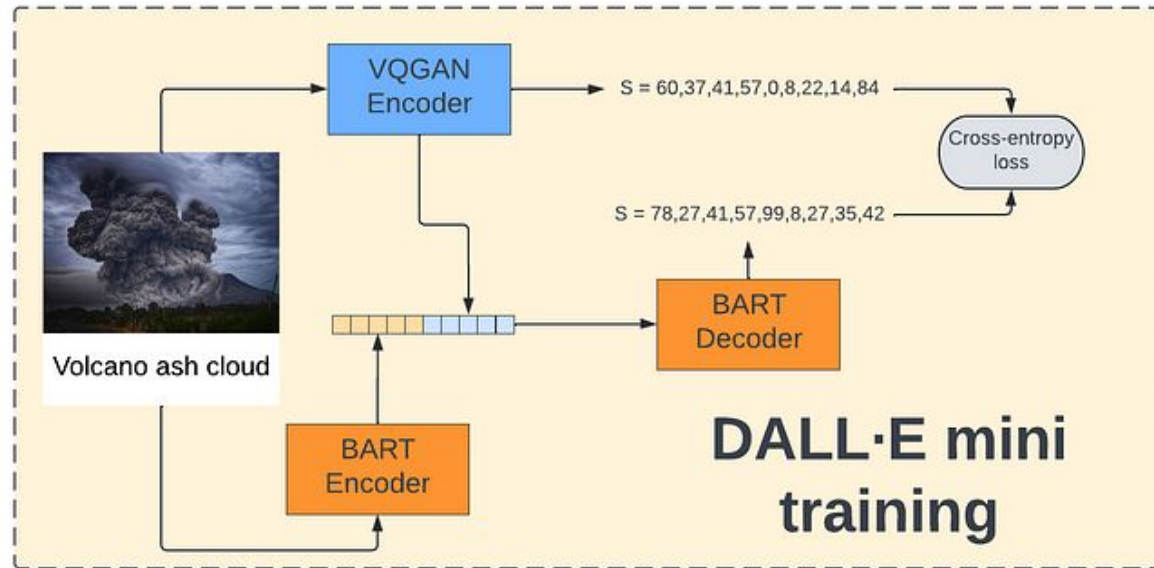


Score: 31.45

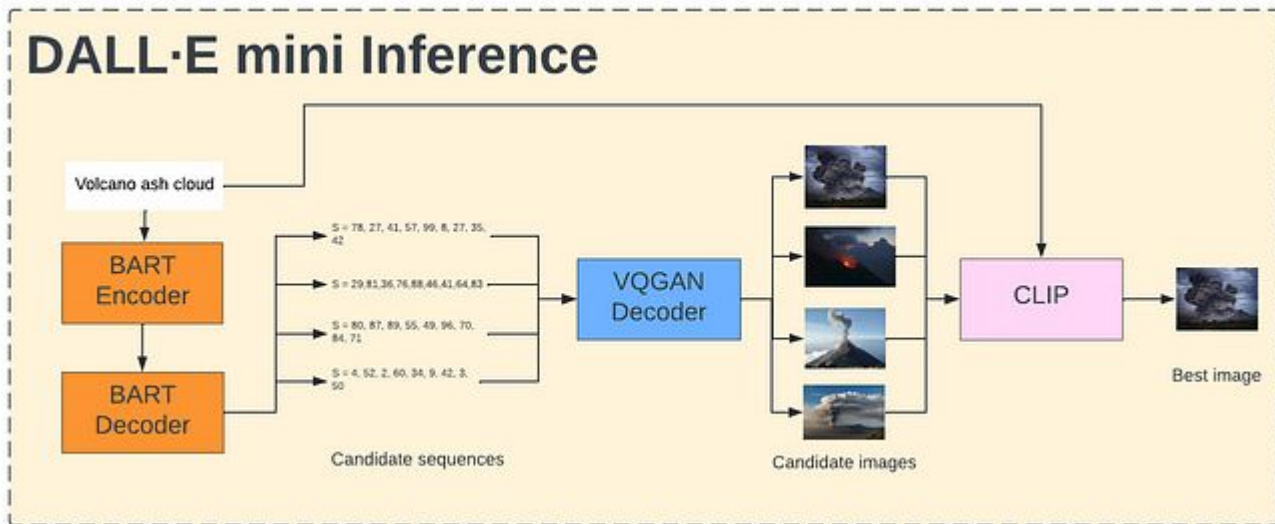


Score: 30.44

Part 5: Piecing the blocks together.



The Dall.E Mini Text-to-Image Pipeline.



Thank you for listening!

Questions?

Examples of Generated Images

TEXT PROMPT

a stained glass window with an image of a blue strawberry

AI-GENERATED
IMAGES



Examples of Generated Images



Examples of Generated Images

TEXT PROMPT a photo of the food of china

AI-GENERATED
IMAGES

