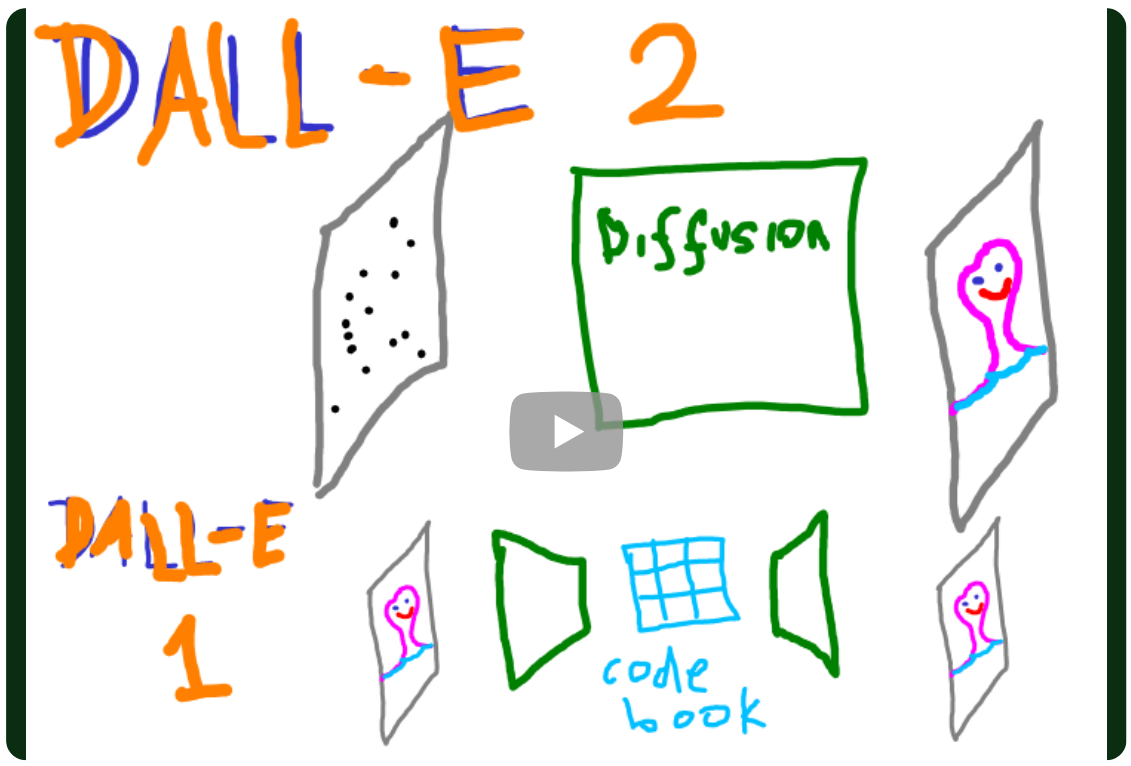
Vaclav Kosar
Welcome back to my blog.

OpenAI's DALL-E 2 and DALL-E 1 Explained

Compare of text-to-image generation models DALL-E 1, 2, and understand related models VQ-VAE, CLIP, and GLIDE



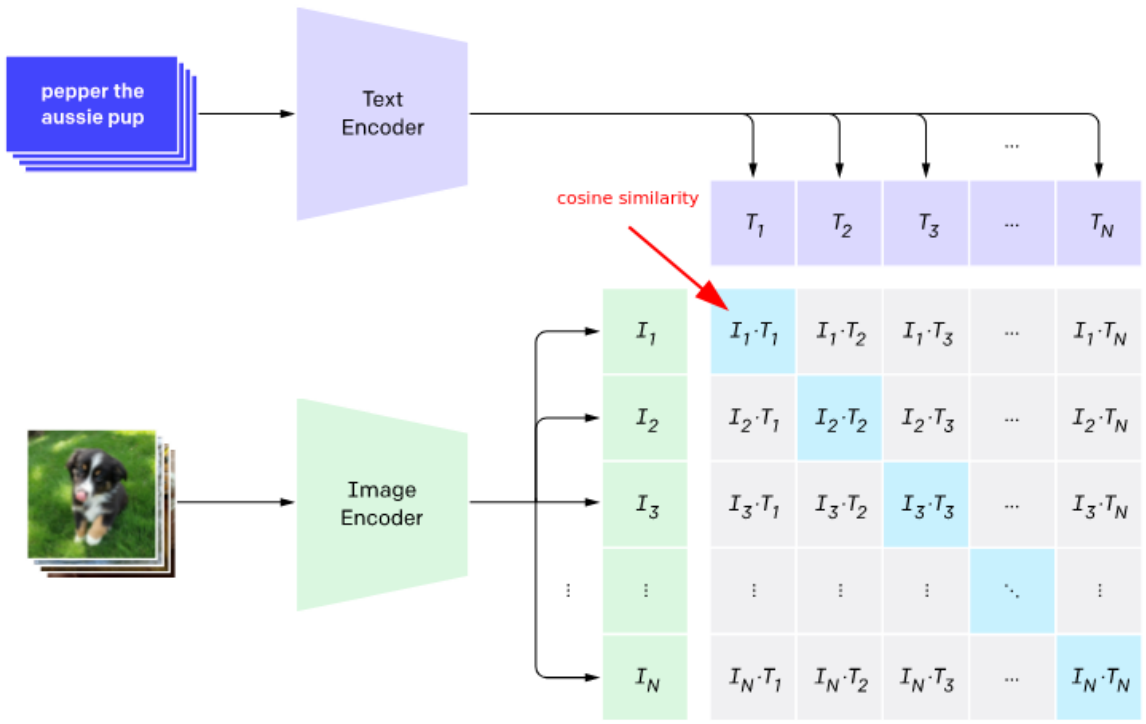
Watch video "OpenAI's DALL-E 2 and DALL-E 1 Explained"

DALL-E 1 uses discrete variational autoencoder (dVAE), next token prediction, and CLIP model re-ranking, while DALL-E 2 uses CLIP embedding directly, and decodes images via diffusion similar to GLIDE.

OpenAI's CLIP

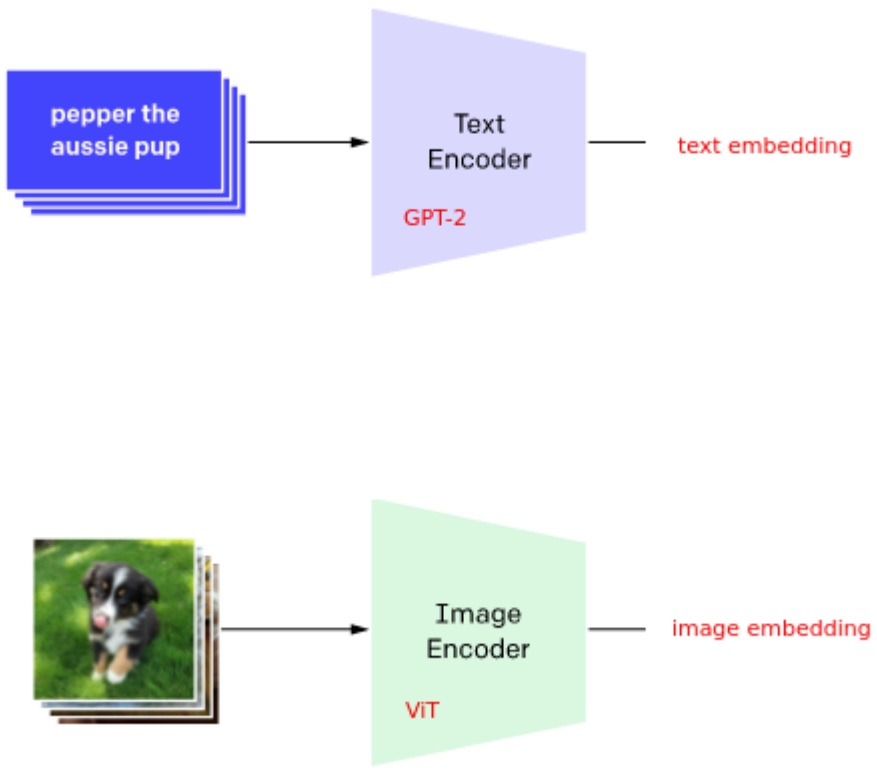
- CLIP: Connecting Text and Images (Jan 2021): encodes image, and text to similar embeddings
- dataset was proprietary WebImageText (WIT is not Wikipedia-based Image Text Dataset (WIT)) 400M of various images with a caption text from the internet
- now open-source image-text datasets like LAION-400M available, open source CLIP models as well
- trained with contrastive learning, maximizing cosine similarity of corresponding image and text
- CLIP's output image embeddings contain both style and semantics
- zero-shot classification, but fails on abstract or systematic tasks like counting

1. Contrastive pre-training



CLIP Architecture

- text and image have separate transformer encoders
- visual encoder is ViT (vision transformer)
- text encoder is GPT-2 transformer
- the fixed-length text embedding is extracted from [EOS] token position,
- text token embeddings and image patch embeddings also available
- trained on 256 GPUs for 2 weeks



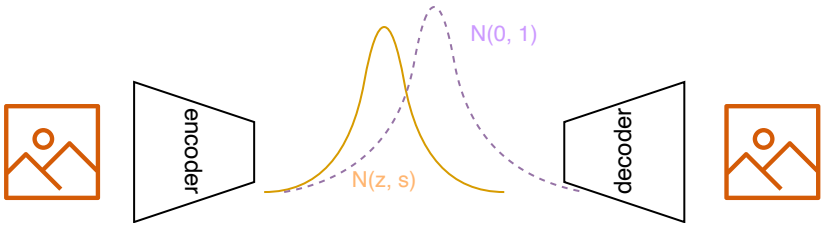
CLIP architecture

CLIP Applications

- DALL-E 1 uses
  - discrete variational autoencoder (dVAE), next token prediction,
  - and CLIP model for re-ranking,
- DALL-E 2
  - uses CLIP embedding directly,
  - and decodes images via diffusion similar to GLIDE.
- zero-shot image classification:
  - create for each class a text -> embedding
  - cosine similarity between image and text embeddings
- image-text classification
  - sum up the two output class token embeddings zero-shot similar
  - or the two output class token embeddings fed in to a shallow MLP classification head
  - or the two output sequences fed into a transformer with classification head

Variational Auto-encoder (VAE) Models

- model the image distribution via lower bound on maximum likelihood
- encode each image as a gaussian distribution on the latent space
- random sampling from latents not differentiable
  - => re-parametrization trick $z = \sigma * r + \mu$ where r is random vector
- loss is to reconstruct (L2) the image and latents to have normal distribution (KL)
- sample, or interpolate from the latent normal distribution and generate images - may find disentangled representations



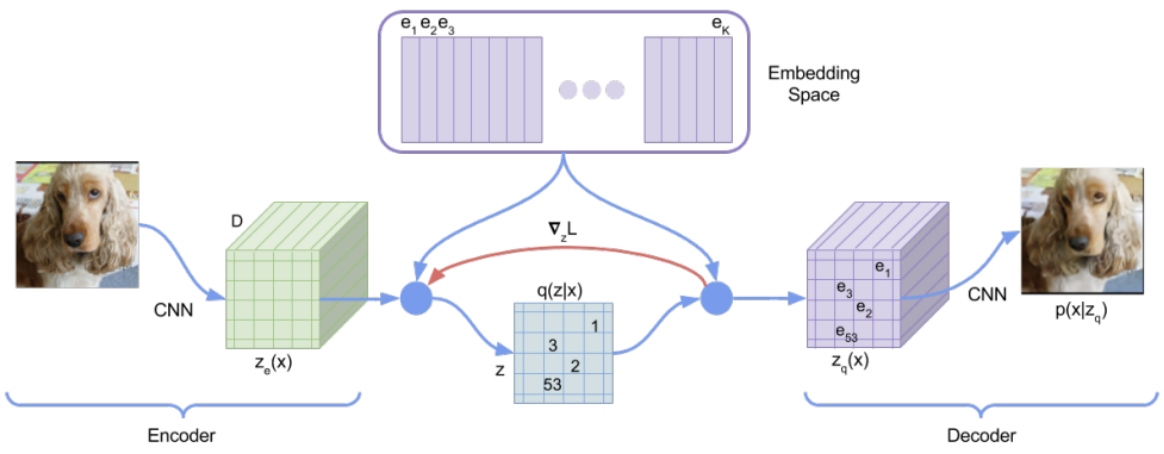
variational autoencoder

Quantization

- related to tokenization in that it outputs finite number of items from a dictionary
- is used in Wav2vec and DALL-E 1 and VQ-VAE
- replaces the input vector with the closest vector from a finite dictionary of vectors called codebook
- during training, backward pass uses Gumbal softmax over the codebook to propagate gradient
- product quantization: concatenation of several quantizations then linear transformation

Discreet Variational Auto-Encoder (dVAE)

- introduced in VQ-VAE 1 and VQ-VAE-2 (dVAE, up-scaling)
- image encoder maps to latent 32x32 grid of embeddings
- vector quantization maps to 8k code words (visual codebook)
- decoder maps from quantized grid to the image
- copy gradients from decoder input z to the encoder output



Discreet Variational Auto-Encoder VQ-VAE 1

OpenAI's DALL-E 1

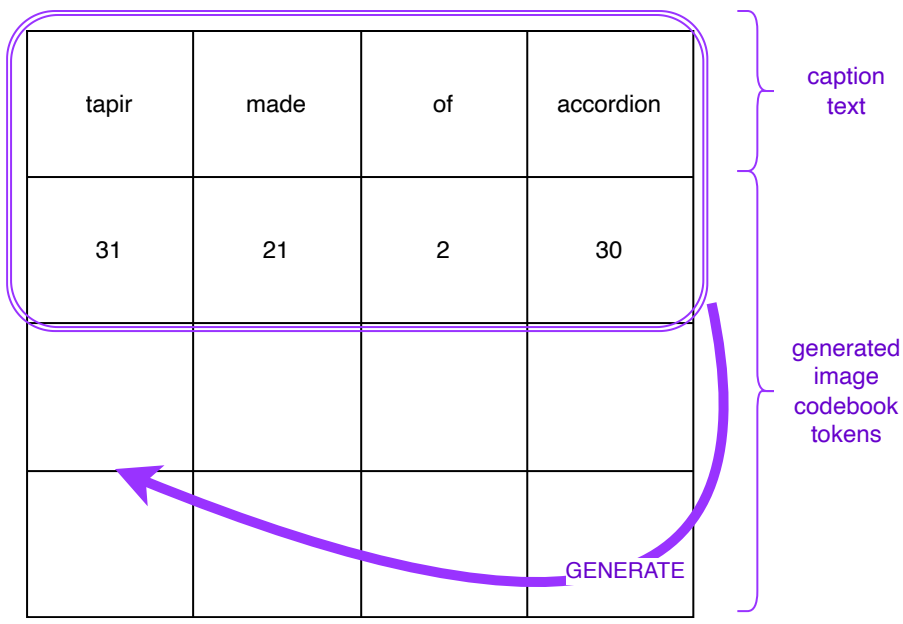
- OpenAI introduced DALL-E 1 text-to-image generator in introduced in paper and code.
- generates 256×256 images from text via dVAE inspired by VQ-VAE-2.
- autoregressive-ly generates image tokens from textual tokens on a discrete latent space.

DALL-E 1 Training:

1. train encoder and decoder image of image into 32x32 grid of 8k possible code word tokens (dVAE)
2. concatenate encoded text tokens with image tokens into single array
3. train to predict next image token from the preceding tokens (autoregressive transformer)
4. discard the image encoder, keep only image decoder and next token predictor

DALL-E 1 Prediction:

1. encode input text to text tokens
2. iteratively predict next image token from the learned codebook
3. decode the image tokens using dVAE decoder
4. select the best image using CLIP model ranker



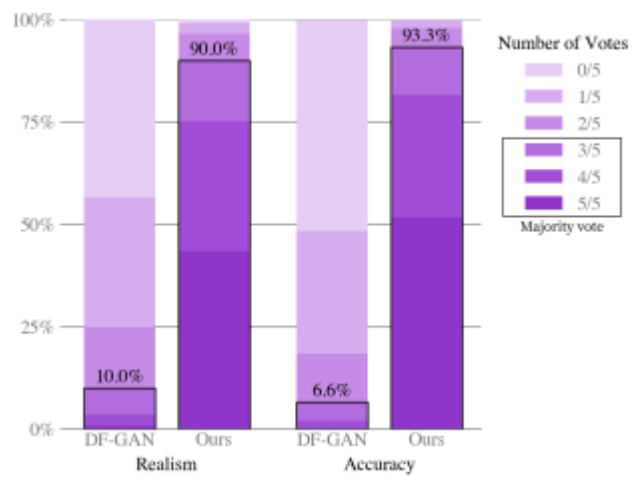
DALL-E-1 generates tokens

DALL-E 1 Discreet Variational Auto-Encoder (dVAE)

- instead of copying gradients annealing (categorical reparameterization with gumbel-softmax)
- promote codebook utilization using higher KL-divergence weight
- decoder is conv2d, decoder block (4x relu + conv), upsample (tile bigger array), repeat

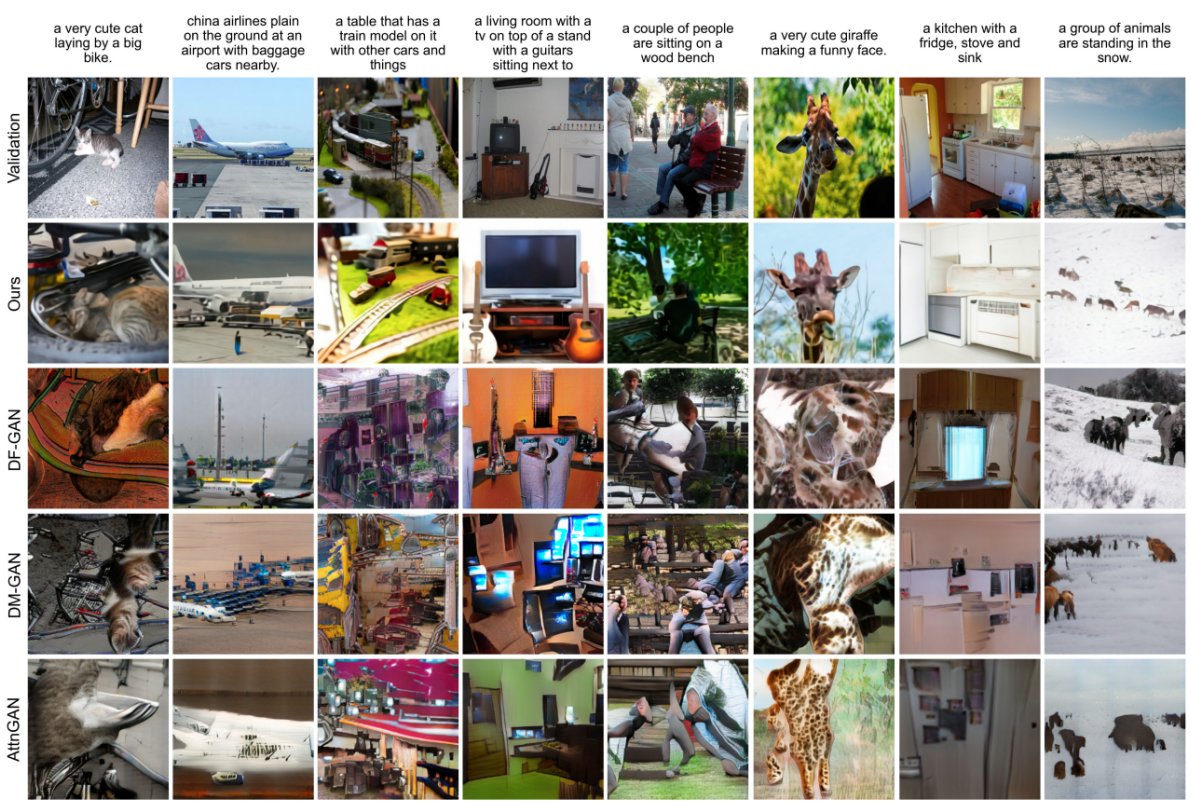
DALL-E 1 Results

- competitive in zero-shot fashion, preferred 90% time by humans
- Human evaluation which is preferred DALL-E vs DF-GAN, zero-shot



DALL-E 1 results

DALL-E 1 Examples



dall-e 1 examples

- Training task is to predict the added noise with mean-squared error loss.
- Similar to normalizing flow models like OpenAI's Glow which are additionally single step and invertible.
- Diffusion model can be formulated as an ODE solution, where de-noising step represents time dimension step. The **training image data form a manifold**. **Adding noise to the images expands** the manifold volume. **The expansion direction and step size of the expansion define the ODE**. The ODE's solution is the probability density function. We link gradient of the density function to the L2 loss of denoising function. The step size is scaled with a function dependent on the noise level.

diffusion model - progressive denoising examples steps (Denoising Diffusion Probabilistic Models)

OpenAI's GLIDE

- Diffusion text-to-image (256 × 256) generator introduced in paper.
- GLIDE outperforms on human preference DALL-E 1.
- CLIP guided diffusion
  - task: "predict the added noise given that the image has this caption"
  - training task is prediction of the noise and guidance towards the CLIP text embedding
  - training loss has additional term of gradient of dot-product with the CLIP text embedding
  - CLIP encoders are trained on noised images to stay in distribution
- text-conditional diffusion model
  - GLIDE diffusion model is a transformer (ADM model)
  - text is embedded via another transformer
  - text embeddings are appended to the diffusion model sequence in each layer

OpenAI's DALL-E 2

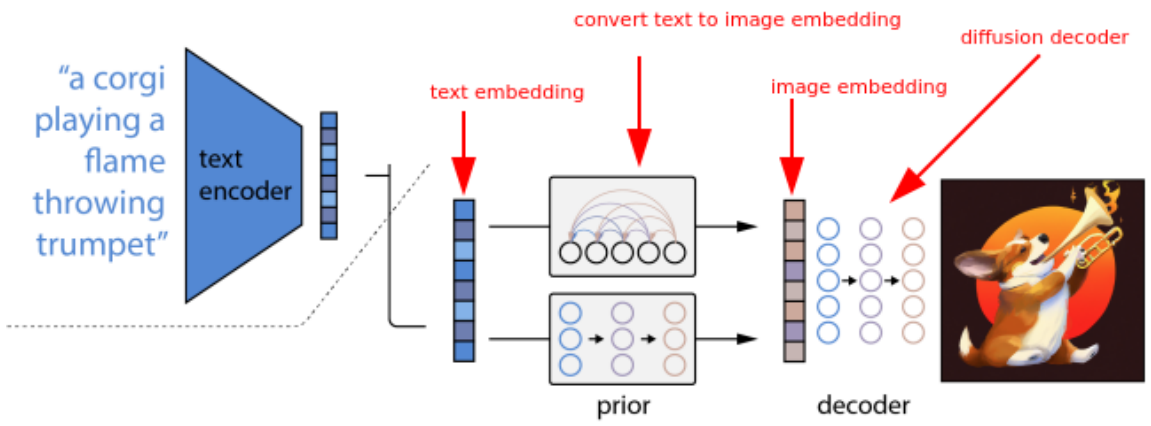
- OpenAI introduced DaLL-E-2 in the paper
- model name is unCLIP while DALL-E 2 is seems to be a marketing name
- generates 1024 x 1024 images from text using diffusion models.
- generates more diverse and higher resolution images than GLIDE.

DALL-E 2 Training

1. generate a CLIP model text embedding for text caption
2. "prior" network generates CLIP image embedding from text embedding
3. diffusion decoder generates image from the image embedding

- Can vary images while preserving style and semantics in the embeddings
- Authors found diffusion models more efficient and higher quality compared to autoregressive

DALL-E 2 Image Generation



DALL-E 2 decoder

DALL-E 2 "Prior" Network

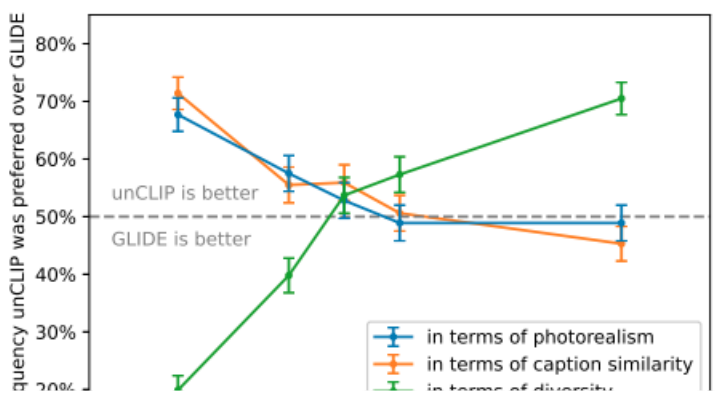
- Prior decoder generates CLIP image embedding from text
- tested autoregressive and diffusion prior generation with similar results
- autoregressive prior uses quantization to discrete codes
- diffusion prior is more compute efficient
    - Gaussian diffusion model conditioned on the caption text

DALL-E 2 Decoder

- diffusion decoder similar to GLIDE
- additionally condition also on CLIP image embedding
    - projected as 4 extra tokens
    - in addition to the text present in the original GLIDE

DALL-E 2 Evaluation Results

DALL-E 2 competitive photo-realism while more diverse images than GLIDE



dall-e 2 human eval results preference

DALL-E 2 Examples

Comparison:

DALL-E 2 vs DALL-E 1 vs GLIDE

Sample (“A teddybear on a skateboard in Times Square.”):

samples from DALL-E “A teddybear on a skateboard in Times Square.”

Created on 13 Apr 2022. Updated on: 22 Apr 2022.

Thank You!

Vaclav Kosar

Let's connect! I may unlock opportunities or help you climb over obstacles.

email address Subscribe

You'll love also...

Vaclav Kosar

Explore this area for additional insights to learn and apply tomorrow.

OpenAI's Image-Text Model CLIP

Encode image, and text into similar embedding vectors for multimodality.

Encoder-Only vs Decoder-Only vs Encoder-Decoder Transformer

Wrap your head around the main Transformer variants in 5 minutes.

OpenAI's Glow - Flow-Based Model Teardown

Interpretable latent representations by composing non-linear invertible functions and maximizing the exact log-likelihood.

Multimodal Image-text Classification

Understand the top deep learning image and text classification models CMA-CLIP, CLIP, CoCa, and MMBT used in e-commerce.

Wav2vec: Semi-supervised and Unsupervised Speech Recognition

Word2vec for audio quantizes phonemes, transforms, GAN trains on text and audio from Facebook AI.

ELECTRA - How to Train BERT 4x Cheaper

Reducing training flops 4x by GAN-like discriminative task compared to RoBERTa-500K transformer model.

Manipulate Item Attributes via Disentangled Representation

Using attribute-specific embedding subspaces for image manipulation retrieval, outfit completion, conditional similarity retrieval.

Vaclav Kosar