DALL-E 2: Text-to-Image Generation

Mustafa Hajij
June 20, 2025

What is DALL-E 2?

- A text-to-image model that generates images from natural language prompts.
- Built on CLIP and a diffusion-based Decoder with a Unet.
- Produces diverse, high-quality images aligned with text descriptions.

Objective
Transform text into visually coherent images using advanced deep learning.

CLIP: Contrastive Language-Image Pre-training

- **Role**: Encodes text and images into a shared latent space.
- **Structure**:
  - Text encoder: 512-dim, 8 heads, 256 seq len.
  - Visual encoder: 512-dim, 256x256 images, 32x32 patches.
- **Training**: Contrastive loss to align text-image pairs.
- **Usage**: Provides embeddings for conditioning the decoder.

Key Feature
Enables semantic understanding of text prompts for image generation.

Unet: The Core Generator

- **Design**: U-shaped network with encoder-decoder and skip connections.
- **Purpose**: Generates images from noise in the diffusion process.
- **Configuration**:
  - `dim = 128`: Base feature dimension.
  - `image_embed_dim = 512`: Matches CLIP's image embeddings.
  - `dim_mults = (1, 2, 4, 8)`: Scales features across layers.

Strength
Balances local details and global context in image synthesis.

Decoder: From Text to Image

- **Components**: Integrates `Unet` and `CLIP`.
- **Process**: Diffusion over `timesteps = 100` steps.
- **Conditioning**: Text and image embeddings guide the generation.
- **Output**: Images conditioned on CLIP embeddings.

Mechanism
Iteratively refines noise into images guided by text semantics.

Training Process of DALL-E 2

- **Overview**: Two-stage process involving a prior and a decoder, both using diffusion models.
- **Stage 1: Prior Training**
  - **Goal**: Map text embeddings $c_t = \text{CLIP}_{\text{text}}(t)$ to image embeddings $c_i = \text{CLIP}_{\text{image}}(x)$.
  - **Method**: Diffusion model in embedding space.
- **Stage 2: Decoder Training**
  - **Goal**: Generate images x from image embeddings c_i.
  - **Method**: Unet-based diffusion model in image space.

Distinction from Imagen
DALL-E 2 uses a separate diffusion prior to generate image embeddings from text, whereas Imagen directly conditions the diffusion process on text embeddings.

Prior Training Loss: Explained

- **Loss Function**:
$$\mathcal{L} = \mathbb{E}\left[\|\epsilon - \epsilon_\phi(c_{i,t}, t, c_t)\|^2\right]$$
    - $c_{i,t}$: Noisy image embedding at time step t.
    - ϵ: Actual noise added to the image embedding.
    - $\epsilon_\phi(c_{i,t}, t, c_t)$: Noise predicted by the prior model ϕ, given $c_{i,t}$, time t, and text embedding c_t.
- **Idea Behind It**:
    - Measures the difference between the actual noise and the model's prediction.
    - Trains the model to denoise embeddings by learning the noise distribution.
- **Why It's Key**:
    - Enables the prior to generate image embeddings that align with text prompts.
    - Captures the semantic relationship between text and images in the embedding space.
    - Foundation for the decoder to produce coherent images.

Decoder Training: Detailed Loss Explanation

- **Loss Function**:

$$\mathcal{L} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(x_t, t, c_i)\|^2\right]$$

  - x_t: Noisy image at time step t in the diffusion process.
  - ϵ: True noise sampled from a Gaussian distribution, added to the clean image.
  - $\epsilon_\theta(x_t, t, c_i)$: Noise predicted by the Unet model θ, conditioned on x_t, time t, and image embedding c_i.
- **Mathematical Intuition**:
  - Mean squared error (MSE) between actual and predicted noise.
  - $\mathbb{E}$ averages over noise samples and timesteps, ensuring robust learning.
  - Minimizing this loss trains the Unet to reverse the diffusion process.
- **Diffusion Process Connection**:
  - x_t is derived from x_0 (clean image) via $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon$, where α_t controls noise level.
  - Unet learns to predict ϵ to recover x_0 step-by-step.
- **Conditioning Role**:
  - c_i (from prior) embeds text semantics, aligning the denoising with the prompt.

Key Outcome
The Unet effectively denoises images, producing results consistent with the text input.

Decoder Training Loss: Explained

- **Loss Function**:

$$\mathcal{L} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(x_t, t, c_i)\|^2\right]$$

  - x_t: Noisy image at time step t.
  - ϵ: Actual noise added to the image.
  - $\epsilon_\theta(x_t, t, c_i)$: Noise predicted by the Unet model θ, given x_t, time t, and image embedding c_i.
- **Purpose**:
  - Trains the Unet to accurately predict the noise in the image at each diffusion step.
  - Ensures effective denoising to generate a coherent image that aligns with the image embedding c_i.
- **Conditioning**:
  - c_i: Image embedding generated by the prior from text embedding c_t, guiding the image generation.

Key Insight
The loss ensures the generated image is semantically consistent with the text prompt via the image embedding.

Inference: Text to Image

- **Input**: Text prompt processed by CLIP.
- **Steps**:
  1. Generate text embedding via CLIP.
  2. Use prior to generate image embedding from text embedding.
  3. Initialize random noise (e.g., 256x256).
  4. Denoise iteratively using the decoder to produce the image.
- **Result**: Image reflecting the prompt's meaning.

Flexibility
Stochastic diffusion allows multiple outputs per prompt.

Applications and Challenges

- **Applications**:
  - Art and design creation.
  - Visual prototyping and storytelling.
  - Synthetic data generation.
- **Challenges**:
  - High computational cost.
  - Limited handling of complex prompts.
  - Data bias risks.

Next Steps
Optimize efficiency and improve robustness.