Text-to-Image Generation

Google Muse

- Large-scale paired image-text data can be collected from the web
- Pre-trained LLMs can be leveraged

Text to image generation: State of the art

- Dall-E 2
- Imagen
- Parti
- MUSE

What is MUSE?

- Not diffusion or auto-regression
- Extremely fast
- High CLIP score
- Application
  - One shot
  - Mask free

Overview of MUSE Architecture

- Mostly transformer based (for text and image portions)
  - CNNs also used (in VQGAN)
- Image tokens are in the quantized latent space of a CNN-VQGAN
- Trained with masking loss
- 2 models, 256×256, and 512×512

Pretrained LLM

- T5 XXL pretrained and frozen text model
  - 4.6B parameters
- Text prompt is converted into a sequence of 4096D vectors
- Embedding is projected down to a lower dimension and fed into the base MUSE model
  - Cross-attention

Vector Quantized Latent Space

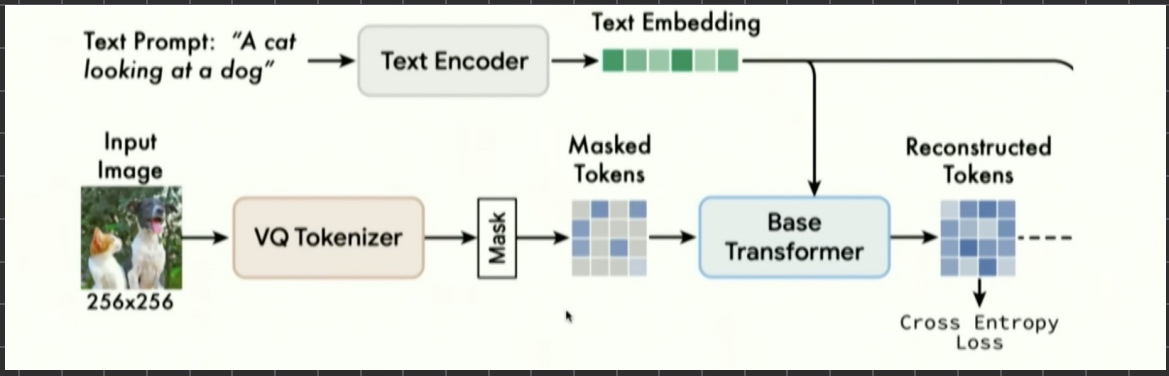
- Use tokens generated from VQGAN
- Quantized tokens are amenable to cross-entropy loss
- Entirely convolutional encoder/decoder structure
- Downsampling ratio of 16 which generates latents of size 16×16 from 256×256

Variable Ratio Masking

VQ Tokenizer → Mask → Masked Tokens

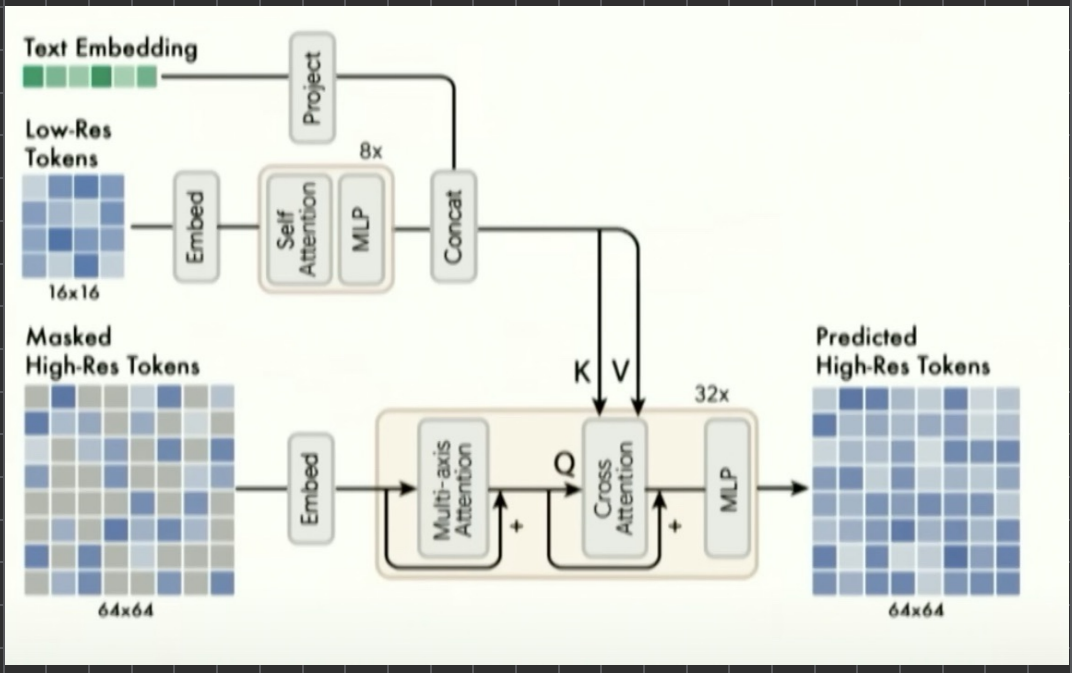
- Sample from a distribution of masking tokens with a mean of .64
- Variable masking enables model to handle different masks at inference time

Base Model



- Cross attention from text tokens to image tokens
- Self attention from image tokens to image tokens

Super Resolution Model



Classifier Free Guidance at different Time

- Heuristic tradeoff of diversity and quality
- Higher guidance scale makes sharper images but foggy background

Negative Prompting

- E.g. "no trees", "no lemurs", "not blurry"

Iterative parallel decoding

- Base Model uses 24 steps; super res uses 8
- Significantly lower than the 50-1000 required for diffusion

Qualitative Evals

- Coordinality
- Style
- Composition
- Text Rendering
- Usage of full text prompt

Note:

The rest of the lecture is subjective and quantitative analysis of MUSE and competitors. Very interesting, but no new concepts.