Label-to-Image Generation

Team6
Yu-Chung Cheng (anthonycheng@vt.edu)

Agenda

- Problem
- Label Encoder
- Position Encoder
- Pipeline
- MaskGiT
- Label-to-Image generation

Reference

- Pre-Trained Model:
  - Hugging Face
    https://huggingface.co/llvictorll/Maskgit-pytorch
- Model and Example code:
  - Victor Besnier and Mickael Chen
    https://github.com/valeoai/MaskGIT-pytorch
- My Repo
  - https://github.com/sunnyanthony/MaskGIT-pytorch

Problem

- Training and inference processes in image generation are time-consuming.
- Our goal is to utilize MaskGIT as our pre-trained model to facilitate few-shot learning with new data.
- MaskGIT is recognized for its rapid image generation capabilities."

Label Encoder

- We have 1000 classes in ImageNet.
- For example, class number 7 is represented as "chicken".
- Since we don’t use text for prompting, we can use torch.nn.embedding to encode our labels.

Position Encoder

- Position Embedding:
  - "**Attention Is All You Need**" paper for initial concept of position embedding.
- MaskGiT
  - Utilizes **Vision Transformer (ViT)** as its backbone architecture.
- ViT
  - Sinusoidal position embedding is not used. Instead, ViT employs image patches for position embedding, diverging from the traditional sinusoidal approach.

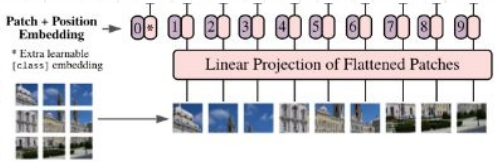


Image Source: Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby: **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.** ICLR 2021

Pipeline in MaskGiT

The model adopts a two-stage design inspired by VQVAE, known for its autoregressive transformer design in image generation.

- First Stage:
  - Retains the use of VQ-GAN (Vector Quantized Generative Adversarial Network).
  - This stage focuses on creating a compressed but meaningful representation of the image data.
- Second Stage:
  - Utilizes MVTM (Masked Visual Transformer Model) instead of the traditional VQVAE approach.
  - MVTM enhances the model's ability to generate high-quality images by learning from the representations formed in the first stage.

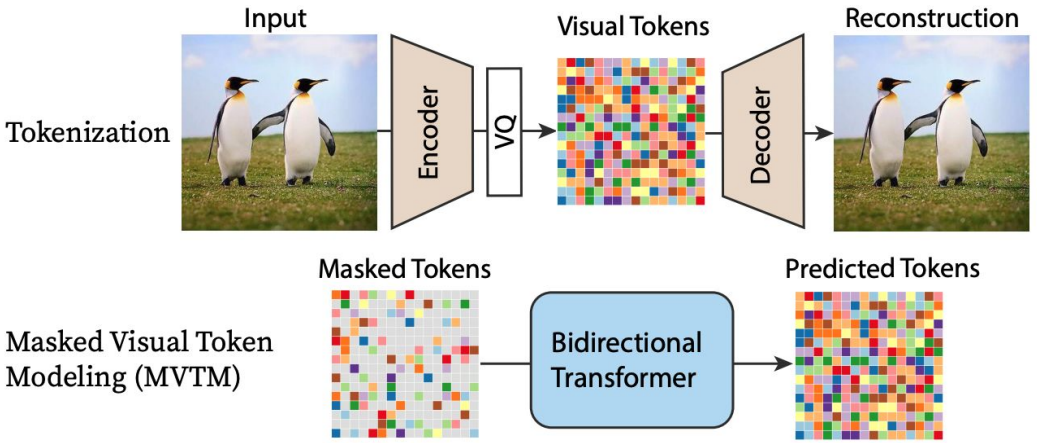


Image Source: Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby: **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.** ICLR 2021

DataSet

- ImageNet-512: A variant of the standard ImageNet dataset adapted for 512x512 resolution.
- CIFAR-10: It is a widely used dataset in computer vision. It consists of 60,000 32x32 color images in 10 different classes, with 6,000 images per class.
- Training Approach:
  - Direct training on ImageNet-512 was not conducted by our team.
  - Instead, we utilize a model that was pre-trained on ImageNet-512 and train with **CIFAR 10**.
  - This pre-trained model provides a robust foundation for further applications or research.

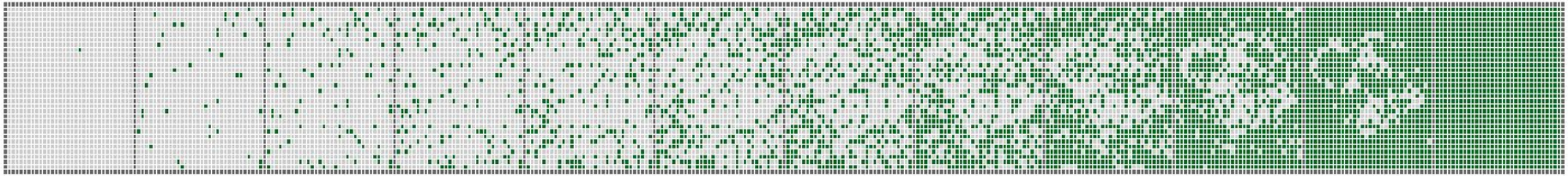
Experiments

- Ustlize the code from https://github.com/valeoai/MaskGIT-pytorch and its ipython
- Train with specific images and a new label
  - Make MaskGIT Transformer as backbone
    - Copy parameters from MaskGIT
    - Random the remain uncopied parameters
- The original dataset (ImageNet) has 1000 classes and we use the 1001 to generate the image
  - It should be like prompting to make a generated image as well even if the label does not exist.
  - Use **CIFAR 10**'s label 5 (dog) as 1001 label for our new model

New Label without training

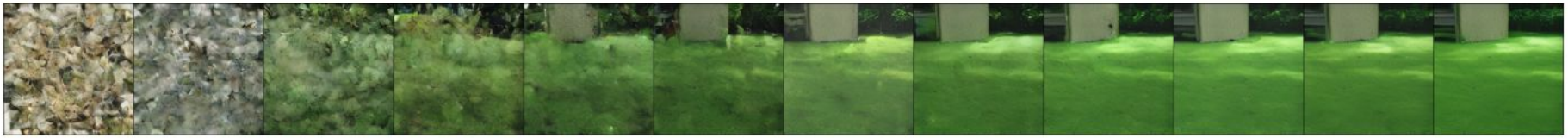
Generated Image

Masked Data for inference

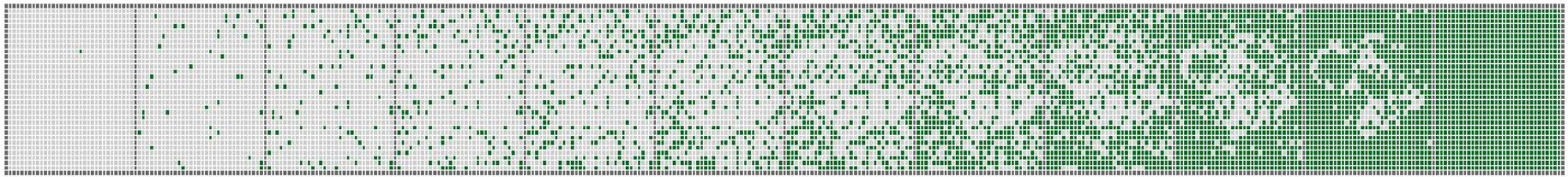


New Label after training

Generated Image with new label



Masked Data for inference



Why?

(hypothesis) The image resolution is too small cause the network can't recognize the distribution of the class.

Another generated example

Learned

- Gaining proficiency in PyTorch for writing and tracking the code.
- Know different way in Gen AI

Q & A