DALL·E 2

Hierarchical Text-Conditional Image Generation with CLIP Latents (2022) 452

GAN:

$z \rightarrow$ Generator $\rightarrow x \searrow$

$x' \rightarrow$ discriminator $\rightarrow$ 0/1

Pro: 真实性高，需要的数据量不大

Con: 训练不太稳定，训练两个网络有平衡的问题，创造性不好

AE (auto-encoder):

$x \rightarrow$ Encoder $\rightarrow z$ (bottleneck) $\rightarrow$ Decoder $\rightarrow x'$ compare x and x'

DAE (Denoising auto-encoder):

$x \rightarrow x_c$ (corrupted x) $\rightarrow$ Encoder $\rightarrow z \rightarrow$ Decoder $\rightarrow x'$ compare x and x'

更加稳健，不容易过拟合；图像冗余性较高，即使污染后还能抓住本质

VAE (Variational auto-encoder):

$x \rightarrow$ Encoder $\rightarrow \mu, \sigma$ (distribution) $\rightarrow z \rightarrow$ Decoder $\rightarrow x'$

$q(z|x)$ $z = \mu + \sigma \cdot \epsilon$ prior $p(x|z)$

改为学习一个分布，假定为高斯分布则可以用μ, σ表示，从中sample得到z

VQVAE (Vector Quantised VAE):

$x \rightarrow$ Encoder $\rightarrow f \rightarrow$ codebook $K \times D \rightarrow f_n \rightarrow$ Decoder $\rightarrow x'$

codebook有K个长度为D的聚类中心，当图片经过encoder得到特征图f，对比特征图与codebook中的向量，找出最接近的聚类中心，使用这个聚类中心对应的特征f_n

DALLE

Text $\xrightarrow{\text{BPE}}$ $f_t \searrow$

concatenated vector $\rightarrow$ GPT

Image $\xrightarrow{\text{VQVAE}}$ $f_q \nearrow$

Diffusion models

$x_0 \quad x_1 \quad \dots \quad x_{t-1} \quad x_t \quad \dots \quad x_T$

forward diffusion $\rightarrow$ add noise

reverse diffusion $\leftarrow$ U-Net, attention… time embedding in U-Net 提醒模型学到第几步了，从粗略特征学到细微特征

DDPM → improved DDPM → Diffusion beats GAN → GLIDE → DALLE2

DDPM贡献:

1. $x_t \rightarrow x_{t-1}$

   $x_t \rightarrow \epsilon = x_{t-1} - x_t$ U-Net only predicts the residual (similar to ResNet)

2. 对于一个分布，固定方差学习均值，效果就很好了

Improved DDPM改动：

1. 学了方差，效果更好
2. 添加噪声的schedule，从线性改为余弦
3. 实验发现scale效果好

Diffusion beats GAN:

1. 更大更复杂的模型
2. 使用classifier guidance引导模型的采样和生成：在训练模型的同时，训练一个classifier（可以在ImageNet上加了噪声训练），reverse diffusion $x_t \rightarrow x_{t-1}$ 过程中，将x_t扔给classifier看分类得对不对，算出交叉熵目标函数得到梯度∇g，用梯度帮助模型的采样和生成

牺牲了一部分多样性换取了真实性，在IS, FID上超过了GAN

一些其他guidance:

$p(x_{t-1}|x_t) = ||\epsilon - f_\theta(x_t, t, y)||$ — y是引导

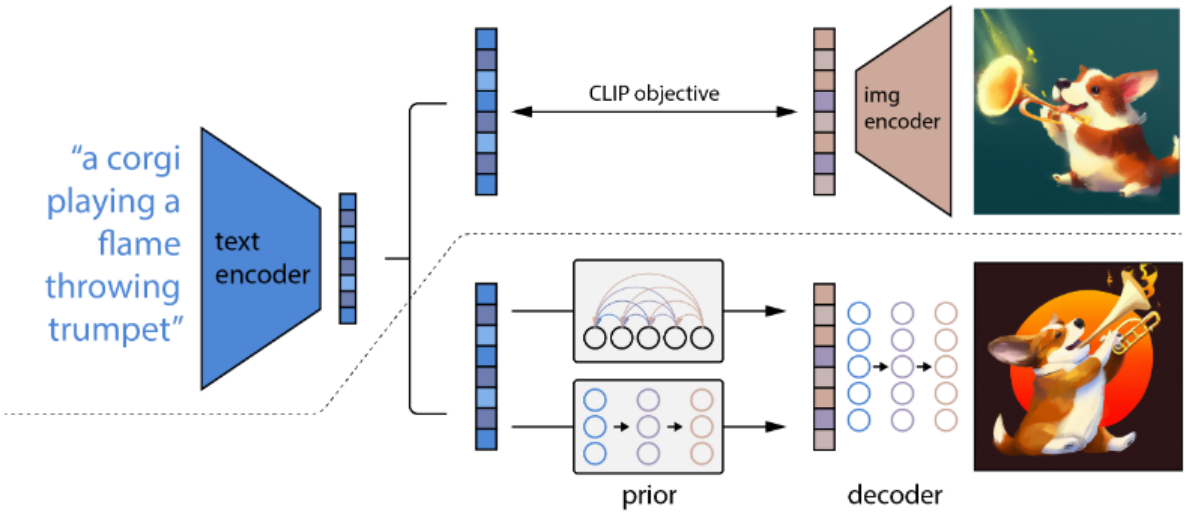
简单的classifier换成CLIP模型，可以把文本和图像联系起来，不仅用梯度，也用文本控制模型的采样和生成；图像除了像素层面的引导，可以用特征层面的引导，风格方面的引导；文本可以用已经训练得很好的大语言模型引导

GLIDE:

classifier-free guidance: $f_\theta(x_t, t, y) - f_\theta(x_t, t, 0)$ 训练中生成有条件时的输出&无条件时的输出，得到一个方向可以从无条件时的输出到有条件时的输出，这样在无条件

时可以推测出有条件时的结果

DALLE 2



CLIP: learn representations of images

text → text encoder ⟺ image encoder ← image

Training dataset: pairs (x, y) of images x and their correspoding captions y

$$P(x|y) = P(x, z_i|y) = P(x|z_i, y)P(z_i|y)$$

decoder: 给定y和z_i生成x　　prior: 给定文本y生成图像embedding的z_i

prior (diffusion model): generates a CLIP image embedding

decoder (experiment with autoregressive ❌ & diffusion model ✔️): generates an image conditioned on the image embedding

CLIP guidance & classifier-free guidance

transformer in decoder