LLaMAGen - Autoregressive Model Beats Diffusion

Llama for Scalable Image Generation

Peize Sun[1] **Yi Jiang**[2†] **Shoufa Chen**[1] **Shilong Zhang**[1] **Bingyue Peng**[2]

Ping Luo[1*] **Zehuan Yuan**[2*]

[1]The University of Hong Kong [2]ByteDance

Codes and models: https://github.com/FoundationVision/LlamaGen

分享人：郝飞洋

12月29日

论文的背景

- 之前的自回归的图像生成模型：
- VQ-VAE
- VQ-GAN
- DALL-E
- Parti……（开源模型发展有限）
- Diffusion模型蓬勃发展：开源社区红红火火
- 但是Diffusion模型用的是完全不同的一套架构，不利于模型融合

图形生成模型的三大关键

- 1. 要设计一个好的图像压缩模块
- 2. 要设计一个便于往上堆规模的模型结构
- 3. 要有高质量的数据

设计哲学

- 减少 inductive bias，而且要和语言模型的架构一致——
- next-token prediction!

- 最近的另外一些工作:
- MaskGIT
- VAR
- ……

文章的贡献

- 1. Image tokenizer：具体参数先按下不表，我们的tokenizer比diffusion里面用的VAE还要好
- 2. 方便堆规模的图片生成模型：基于llama结构搞了从111M到3.1B参数的一系列模型，最大的模型在ImageNet 256*256的评测中比LDM和DiT效果好。
- 3. 高质量的训练数据：训练了一个775M参数的文生图模型，可以提供高质量的图文对
- 4. 用vLLM提高了生成速度326%-414%

图像生成的自回归模型概述

- 1. 使用image tokenizer对图像进行量化，将特征表示成codebook的索引的形式
- 2. 将上面的特征变成一维，在本文中使用的是光栅扫描顺序
- 3. 使用上面的一维token来训练transformer自回归模型
- 4. 在生成图像时先使用自回归模型生成image tokens，再用image tokenizer decoder来转成图片

Image Tokenizer

- 和VQGAN相同的结构——encoder-quantizer-decoder

- encoder：将图像的像素点x投影到特征空间f里面去
- quantizer：将特征空间里面的f映射到codebook里面距离最近的特征向量z，设z的下标索引是q；当decoding时根据索引q找到映射的z
- decoder：把z转化回图像的像素点x_hat

损失函数

- Straight-through gradient estimator

因为量化是一个不可导的操作，为了算出从decoder到encoder的梯度值，我们使用了一个估计的方法——直通估计straight-through gradient estimator。具体如下：

$$z = sg[z - f] + f$$

其中的$sg[\cdot]$ 是stop-gradient操作。

损失函数

- Codebook的学习

对于codebook的学习，损失函数为：

$$L_{VQ} = \|sg[f] - z\|_2^2 + \beta\|f - sg[z\|_2^2]$$

上面的第二项是commitment loss，可以推动从encoder提取出来的特征向量和codebook中的向量更加接近

损失函数

$$\mathcal{L}_{AE} = \ell_2(x, \hat{x}) + \mathcal{L}_P(x, \hat{x}) + \lambda_G \mathcal{L}_G(\hat{x})$$

第一项：逐像素的重建损失

第二项：感知损失（用预训练的模型来计算两张图片之间的差异）LPIPS - Learned Perceptual Image Patch Similarity

第三项：对抗损失（使用和image tokenizer同时训练的PatchGAN）

自回归图像生成——基本结构

- 模型结构很大一部分是在Llama的基础上做的，
- 用RMSNorm(Root Mean Square Layer Normalization)做pre-normalization
- 用了SwiGLU激活函数和旋转位置编码
- 为了保持和大语言模型的统一性，没有用AdaLN(Adaptive Layer Normalization)

自回归图像生成——给定类别

- 类别的embedding是从一组可学习的embedding中索引出来的。
- 生成图片时，从这个token embedding开始，然后用next-token-prediction的方法来生成图像token序列，到预先定义好的最大长度为止。

自回归的图像生成——给定文本

- 使用FLAN-T5 XL作为文本encoder来文字的信息注入自回归模型中，过了encoder之后的文字feature再过一层额外的MLP，然后作为起始的token embedding开始生成。

- 这只是暂时的方法，总有一天会有统一的语言和视觉词表!

站在前人的肩膀上……

- 上面说到的这些设计都很大程度上是站在前人的肩膀上：
- image tokenizer借鉴了VQ-GAN
- 图片生成借鉴了DiT和VQ-GAN
- 很多先进的设计（比如classifier-free guidance）在diffusion里面搞得热火朝天，但是自回归模型里面鲜有人做这样的工作，所以这个工作就把这些不错的设计带到自回归里面来。
- 文字领域的推理加速vllm