Imagen

本文是 google 提出的，是继 DALLE2 的后续工作，整体框架比 DALLE2 简单的多。

"A brain riding a rocketship heading towards the moon."

"A dragon fruit wearing karate belt in the snow."

"A small cactus wearing a straw hat and neon sunglasses in the Sahara desert."

Contribution

可以直接用文本模型（ T5，这个模型冻住不更新）抽取文本的特征，利用**这个特征指导扩散模型生成对应文本的图像。**结合了强大的（text-only）语言模型和 conditional diffusion model 来做生成，可以生成高质量的图像。使用 **dynamic thresholding** 来改进 diffusion sampling。从而生成更真实和细节丰富的图片。改进了 U-Net，提出 Efficient U-Net，可以更省内存。

Method

模型架构非常简单：

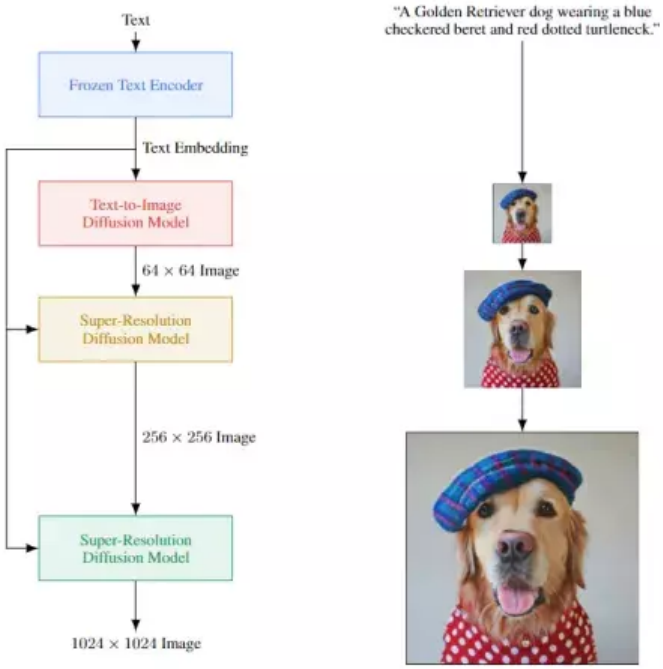


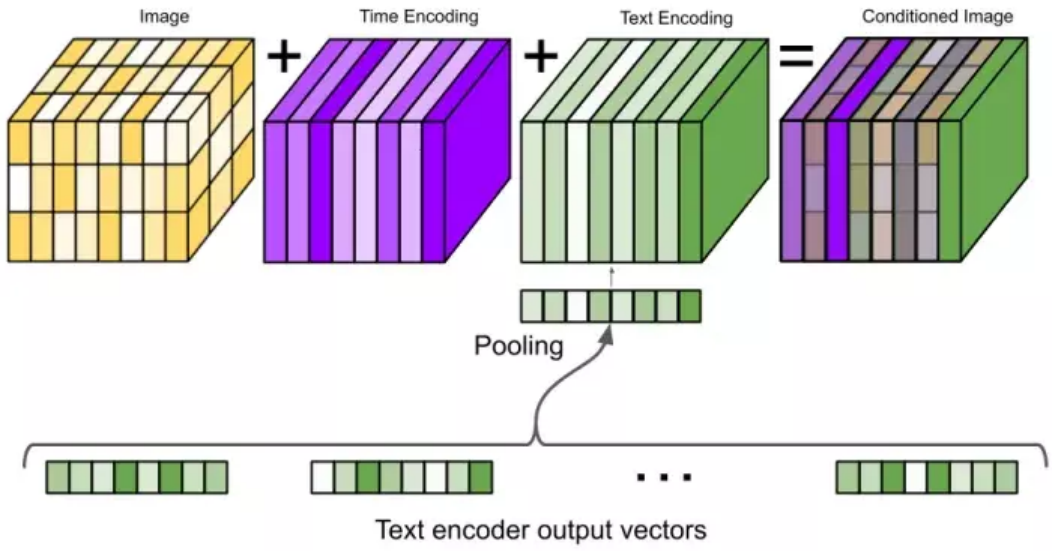
Figure A.4: Visualization of Imagen. Imagen uses a frozen text encoder to encode the input text into text embeddings. A conditional diffusion model maps the text embedding into a 64×64 image. Imagen further utilizes text-conditional super-resolution diffusion models to upsample the image, first $64 \times 64 \rightarrow 256 \times 256$, and then $256 \times 256 \rightarrow 1024 \times 1024$.

Imagen 流程如下：

首先，把 prompt 输入到 **frozen text encoder** 中，得到 **text embedding**（这个表达已经蕴含了所有文本信息）把 **text embedding 输入到生成模型中**，其实就是给模型信息，让他基于这个信息去生成图像。**第一步先成低分辨率的图像，然后再串联 2 个 super-resolution 网络，这两个网络的输入是前面的低质量图像和 text embedding。最终就可以输出高质量的图像。**

更深入的解析：

把 NLP 中很强大的语言模型拿过来用，而不是像使用 CLIP 那样在 image-text pair 训练的 text encoder。（另外说一下，DALLE2 的做法是用一个 prior 网络把 text encoder 输出的 embedding 转为 image encoder 的输出，多了一个模型来做图文 embedding 的转换～）。直觉上，语言模型训练的数据量远远大于 image-text pair ，并且其模型大小也远远大于当前的 image-text 模型，显然语言模型对于文本的理解能力更强，理解了文本才能生成高质量的图像。把 text encoder 抽到的文本信息做一个 pooling 之后，作为一个 embedding 加在原来的图像上从而实现 condition 操作。（**当然，也有直接做 cross-attention，实际论文不是这么操作的，经过 ablation 之后发现其实直接用 concat 到后面然后做 cross attention 的效果会好**）



classifier-free guidance：训练的时候随机扔掉 condition 信息（设置为 NULL），这样做就可以实现 classifier-guidance 的效果。（也就是无需外部的 classifier 也能做 guidance。这里的 classifier 不一定是分类器）。**这里主要的作用在于可以利用这个调整 text 对于图像的引导程度**。threshold：在 sampling 的时候，预测的 x 范围应该在 [-1,1] 之间，然而实验发现，如果 guidance weights 过大，会很容易让预测的 x 超出范围，因此作者提出了两种解决方案。（后者好一些）static thresholding — 直接暴力截断到一定的范围内。dynamic threshold—— 直接除以一个尺度因子

1. First, the caption is input into a **text encoder**. This encoder converts the textual caption to a numerical representation that **encapsulate(封装) the semantic information within the text.**
2. Next, an image-generation model creates an image by starting with noise, or "TV static", and slowly transforming it into an output image. To guide this process, the image-generation model receives the text encoding as an input, which has the effect of **telling the model what is in the caption** so it can create a corresponding image. The output is a small image that reflects visually the caption we input to the text encoder.
3. The small image is then passed into a **super-resolution model**, which grows the image to a higher resolution. This model also takes the text encoding as input, which helps the model decide how to behave as it "fills in the gaps" of missing information that necessarily arise from quadrupling the size of our image. The result is a **medium sized image** of what we want.
4. Finally, this medium sized image is then passed into yet **another super-resolution model**, which operates near-identically to the previous one, except this time it

takes our *medium* sized image and grows it to a **high-resolution image.** The result is 1024 x 1024 pixel image that visually reflects the semantics within our caption.

- How Imagen Actually Works