文生成图任务

杨国兴　　2020000124

任务介绍

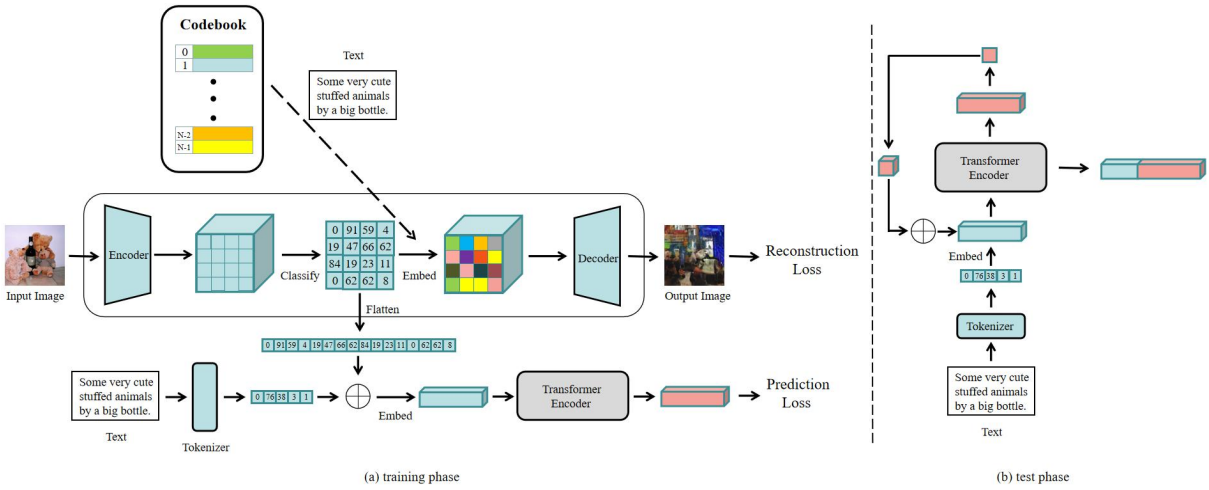
文生成图任务的目标是根据一段文本描述生成一张符合文本描述的真实图片。过去的大多数模型都使用 GAN 模型作为基础，针对某个特定域的数据集进行训练。受 OpenAI 最新发布模型 DALL-E 的启发，我们也采用了 VQ-VAE + Transformer 的架构对该多模态任务进行建模，并在 MS-COCO 这一具有域多样性的数据集上进行了训练。

数据集介绍

我们采用具有域多样性的 MS-COCO 数据集作为我们本次训练的数据集。我们采用的是 2014spilt，其中包括了 8.2w 的训练数据以及 4w 的测试数据。我们对每张图片进行了随机剪裁并把图片分辨率统一为 256*256。

模型介绍

我们的模型与 DALL-E 的方法相似，均采用 VQ-VAE + Transformer 的网络架构，但是不同的是，为了增强图片的质量，我们采用了 VQ-GAN 代替 VQ-VAE。我们的模型结构如下图所示。



我们的模型是一个 two-stage 的网络结构。网络的 first-stage 由 Encoder 与 Decoder 构成，但是与普通的 VAE 不同的是，其采用的 VQ-VAE 的思想，对中间的 feature map 的每个元素用一个预定义的 code book 进行了量化，用量化后的 feature map 作为 Decoder 的输入。值得注意的是，这个词表也是参与训练的。VQ-GAN 与 VQ-VAE 不同的地方在于，其在网络结构外额外加入了一个 Discriminator 判别生成图片的真伪，以辅助 Decoder 能够合成更清晰的细节。

在训练好 first-stage 的 VQ-GAN 以后，我们以图文对的形式进行 second-stage 的 transformer 的训练。在用 Encoder 以及 code book 对一个输入图片进行量化编码后，我们能够得到一个表达图片信息的 token 序列，在对文本也进行 token 化后，我们便能够得到输入图文对的两个 token 序列。在我们训练 transformer 的过程中，我们将这两个 token 序列视作同等的 token，并不对其做区分，而是简单的把它们拼接起来让 Transformer 去学习如何区分他。可以看到，我们的 transformer 其实只是个 decoder-only 的结构，在训练时，每个位置的预测过程仅仅能够看到之前位置的输入，而在测试时，图片部分的预测值是一个一个生成的。 在得到预测的 token 序列后，我们便能够用 Decoder 将其恢复为一张生成的图片。

生成结果展示

A half eaten meal sitting on a plate

A plate of finger foods next to a blue and raspberry topped cake

A plate on a wooden table full of bread

A bean and corn mixture, rice, and broccoli on a plate

A surfer is riding on a small wave

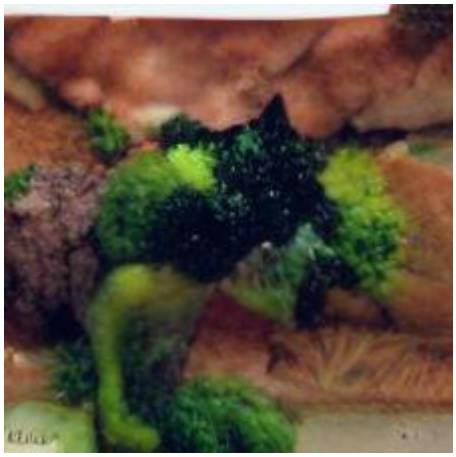
a plate that has some cut up vegetables on it Carrots and small green beans on a white plate

a close up of a slice of pizza on a plate

A group of zebras walking away from trees

A man standing at the beach shoreline and watching the sunset

The white bridge stretches out over the horizon as the transport train travels below it



Cooked broccoli sitting on a plate with salmon

The fire hydrant in the green grass is red

可以看到，我们的模型对文本的语义内容以及基本理解并且将之与视觉信息对应起来了。比如其理解了 Zebra 是斑马并能够将其与黑白相间的视觉特征对应起来。但是同时也能看到，因为数据集本身域多样性的原因，GAN 还是无法太好地适应这种数据集，因此在合成的图片上出现了明显的扭曲和失真。