周报-——2022.5.12

一、完成工作

这两周主要看了三篇论文——**CLIP、DALL-E、StyleGAN**（详细的论文笔记可以看https://github.com/pang-discussion/gq）

1. CLIP

   这是一篇关于**文本——图像预训练**的论文，发表于2021年初，这篇论文的最大贡献是证明了可以通过对比学习（自监督）的得到同时含有图像特征和文本特征的**多模态特征**。这是一篇关于预训练的论文，类似于BERT，它是与下游任务无关的，可以利用它在下游做一些多模态的任务、比如CLIPDraw、StyleCLIP（这两个都是做文本——图像生成的），另外在一些目标检测和视频检索的下游任务中，CLIP也可以起到很好地作用。

   从这篇文章中学到的东西：

   - 对比学习方法，这种方法在2020年的时候很火，现在最出名的何凯明大神的MoCo就是一篇对比学习的论文。对比学习天然就可以用在多模态当中（学习一个特征，这个特征可以做到视角不变性，可以参考论文CMC）
   - CLIP已经被OpenAI封装成了clip包，更方便做下游任务（目前，只尝试了使用它做分类任务，后面会看看styleCLIP那篇论文的代码，看一看在其他方法的应用），另外有些论文采用了CLIP的改编作为一个损失函数，后面也可以看看主要的思想
   - prompt template（提示模板），这是一个在下游任务中常用的方法。在imagenet分类中，通过"A photo of a {label}"代替"{label}"，这样可以解决歧义问题，另外对于确定的数据集可以采用更详细的template

2. DALL-E

   这是一个关于**文本——图像生成**的论文，是一个类似于GPT的生成式模型，发表于2021年初。它的目标是将文本token和图像token当成一个数据序列，然后采用Transformer自回归的进行训练。DALL-E主要由四部分组成——**dVAE、BPE Encoder、稀疏Transformer、CLIP**。它采用dVAE对图片进行降维，并且将图片编码为token。采用BPE Encoder对文本进行编码成token，然后将两个token拼接在一起（**拼接方法没有看懂，下一步会通过代码详细了解其中的思路**），作为Transformer的输入。最后通过自回归生成图片，用CLIP来判断哪个图片与给定文本最接近。

   从这篇论文中学到的东西：

   - VAE和Transformer的原理，太长时间没看了，忘记了VAE和Transformer的原理，通过这篇论文，我再一次去了解了VAE、DQVAE和Transformer。下一步会去阅读这两个东西的源码，并且复现一下，加深一下印象
   - 对图片输入输出的处理方法

3. StyleGAN

   这篇论文是关于**图像生成**的论文，发表于2019年末，它的思想来源于风格迁移，它的目标是生成更多样化并且更高分辨率的图片，另外它要求生成图片的特征应该是解纠缠的（可以通过不同的分辨率层控制生成图片不同的特征），这篇论文可以算是DeepFake方向的论文，后面有很多人物图像生成和图像编辑都是在它的基础上展开的。这篇论文的源代码是TensorFlow1.*写的，阅读起来很困难（大致将网络的框架和损失函数部分的代码逐行研究了一下，另外找了一个pytorch的代码正在阅读和复现）。这篇论文是在ProGAN的基础上写的，省略了判别器和损失函数，因此只能通过代码去了解了这部分（后面有空也会看看ProGAN这篇论文）。

   从这篇论文学到了：

   - StyleGAN论文的基本框架，渐进式的控制生成图片的不同特征
   - AdaIN归一化方法，自适应实例归一化在风格迁移的论文当中很常见（在SimSwap这篇论文中就采用了这个归一化方法）。这个归一化的思想来源于风格迁移，认为**图像的风格来自于图**

像的均值和方差，因此可以通过改变图像的均值和方差得到不同风格的图片
    - 训练技巧：样式混合（可以用来阻止相邻特征耦合）和截断（解决低密度区域的生成质量问题）

二、后面安排

- 复现一下VAE和Transformer
- 继续研究和复现DALL-E和StyleGAN的源代码
- 看一看StyleGANv2和StyleGANv3这两篇论文，了解一下这些论文的改进思路
- 重点看StyleCLIP这篇论文，研究透这篇论文的代码，了解一下这种预训练模型在下游任务的具体做法。

，