小志(00:03:39): 各位晚上好，这里是小志杰，然后今天分享的主要是一个关于 context 的一个理解和思考，那么这就是 C 的一个结构图，在他们的一个技术报告上面。一个很简单的一个的一个架构，它的主要的核心架构是跟 DV 其实是一致的，只不过他们在输入的 len 上面多加了一个。参考的图片去通过 VE 进行一个 N，然后呢，同时把训练的图片和参考的图片一起送到这个。vision bench 里面里面，然后呢，在同时他们在这个 RP 上面多加了一个维度，叫做 T。T 呢，其实就是等于时间，我们训练的图片呢是 T 等于零。我们参考的图片呢 T 等于一，就是一个很简单的一个改进，但是呢，通过大量的。训练之后呢，得出的效果却是非常的有效，这边就是 T 等于0t 等于一，其实这一个是更多是参考的视频生成模型的一个操作。把这个图片去进行一个维度的扩展。

小志(00:05:09): 然后呢？ contest 的架构和 DV 有什么不同呢？就是在不训练参考的情况下去进行一个基于提示词去生成这个图像的一个训练。然后呢，第一呢，就是没有那个参考图片的情况下去生成，基于提示词去生成一个木有图片，然后这边我们可以简单的看到他们的一个训练和生成的一个过程。在第一张图训练这边，我们的提示词呢是把花改成这个紫色，这个是我们的提示词，然后呢，加上这个训练的一个。参考图片和我们的训练目标，一个紫色的花，首先他们这两张图是需要一个高度的一致性的在这一个。修改的一个 task 上面是需要一个高度的一致性的，然后呢，通过逐渐的把我们的 t0每一个时间。都去进行一个训练，然后呢，最后达到一个分配上的一个匹配，就是跟 D v1样的一个流匹配机制，然后训练成可以从下面我们生成的情况下，可以从一个纯噪声的一个情况下逐渐的把这个紫色花还原出来。那么这里就已经是比较直观的去展示了他的一个训练的过程和一个生成的过程。然后呢，通过这样的1个训练，其实是对齐了 t0和 T 一的一个位置信息，也就是我们前面看到的那个。t0就是所谓的多了一个维度，它的 ROPE 1个位置编码去多了一个信息，然后呢，我们 P 一那个位置编码的信息呢，就是一那个维度信息就是一。

小志(00:07:06): 然后呢，把两者的一个分布给连接起来，你要去进行一个训练，首先呢，这样训练呢，他们需要一个大小一致。然后呢，提示词呢，会使用到一个叫做 key 或者是 same 的一个提示词，去把这个任务训练到提示词里面。然后呢，当我们想要，当我们使用 content 的时候，我们想要一致性的话，我们就需要对齐位置信息。如果我们想要一个。生成一个不一致，比方说有不同的图像。The task. 比如说扩图或者是融合情况下，我们就可以用一个。不一致的信息去进行一个生成，当然他这个当你 t1和 t0它的一个大小，或者是说它们的一个位置信息不一致的时候，相似度就会降低。

小志(00:08:09): 如果我们既想要相似性，又想要，那么我们应该怎么办呢？如果从这边来看的话，我们当当我们的 t0把这个。画面这个拉长的话，我们的参考 t1虽然说它也能够大致上还原这个 t1的信息，但是呢，其实它并不是很严谨的对齐的。它好像这个花就会稍微的低了一点，对吧，他这个位置信息的话就不是很准确，那么最基础的使用就是。

小志(00:08:47): 我们根据我们 PE 去进行一个局部修改，把花变成一个紫色。然后呢，当我们想要一致性的扩图的话，就我们既想要 t0的信息。我们又想要保持它保持它的一致性。我们需要把这个图片加上一个 padding，或者是加上一个空白的区域，让他去对这一个区域进行一个针对性的生生成。然后结果呢，就大概是这样，我们能看到，在这个情况下，我们利用文字条件去增强它的一个位置信息的一致性。

小志(00:09:28): 能够得出两张是比较接近的一个图片，而扩图呢，主要是集中在修改它扩图的这一部分。在第三列。这里我们可以看到，其实它的一个实际上的一个对比，或者是实际上的一个。拼接的情况是一个 t0在1个上方，然后 T 一在下方，然后呢去进行一个这样的拼接，而不是我们。之前经常用到的那种水平拼接的方式，不过这边为了展示他们的不同，或者展示他们的一致性，我用一个水平拼接来展示一下。然后呢，这边的文字条件通常是用那个 keep 或者是 same 的一个文字条件去控制它的一致性。如果你想用。如果你想这个一个扩图，而它的一致性并不是你一个主要的目标的话呢，那可能就用一些别的一些提示词。来触发不同的修改类型来达到这个融合元素的一个效果，这边有不同的修改类型，是根据那个官方的一个测试集去进行一个。提取出来来展示的，它分分别是有这个 insertion 有分，这个 local 和还有 computer reference test editing ，还有 style reference。这几个不同的 task 呢？它对于一致性的要求其实是其实是不一样的。所以如果你想精准的去触发某些 task 某些训练的 task 的话，可以去参考那个官方的测试级使用到的。

小志(00:11:14): 基于这些不同的 task 使用到的一些提示词，比方说 remove repay 不同的一些提示词，然后呢，这边展示了一个。不同的 task，然后呢，它官方训练的这些不同的 task，首先它增强了模型的能力范围，然后呢会让模型泛化的更好，但是呢会有一些缺点，就是对用户的意图看。定不太准确，它会混淆不同的任务，然后降低一次性，就比方说右边的这一个，这个情况就是我们用一个。

小志(00:11:59): Image. 去进行一个 pattern，然后去生成两张不同的图片，它们有什么区别呢？它们主要区别就是文字条件上面第三张图我是删除了一些一致性的一个要求。就比方说 keep the far in the same position 类似的一个提示词，然后呢，它就会呢，因为它有训练。这个 instruct ion editing 的不同的 task ，它就会让判断不准确。我们想要的目标是什么，然后呢？生成一张和背景。和它自己生成的背景更融合一点的图片，然后呢，相对来说一致性就没有第二张图片这么高，但是它会比第二张图片更加的自然一些。

小志(00:12:52): 然后。分享就到这了，非常的短和快。接下来就是这个，然后有一些参考参考了不同的文章，还有这个官方的一个测试题，唉，就是这样，然后接下来呢，分享结束的快就到了这个分享。提问讨论环节，那么可以，如果有对这对今天的分享有有疑问的话，可以现在提出讨论。我想边路介绍。

小志(00:13:41): 我们都没有疑问吗？怎么训练怎么训练呢？首先要首先有几个方面，就是你要理解你想要训练的是什么？比方说你可能想训练换衣对吧，我们需要去准备相关的，然后去对这个训练的目标去进行一个设计，然后去进行一个。训练所以怎么训练这个太笼统了不，不太不太好回。

youkengi(00:14:19): 我我有个疑问，这个东西是不是原理上跟那个差异，炼丹法有点接近，就是它的训练，其实就是把这个。A 就是两个相似的训练集，把 a 变到 B，然后呢，就是为什么必须要他的那个就是按照他的格式来写呢，这个就相当于有点像 lola 的触发词。我觉得。是不是可以这么类比？不一样的。

小志(00:14:44): 不一样。我们可以，我可以再把这个屏幕共享一下。首先呢，它是一个这样的一个架构，他是首先差异炼丹或者是说这个。它的一个输入的情况就不一样，因为差异链呢？你是先把一张图给链到底膜上面，然后去对这张这个底膜去进行一些。加减的一些操作，这个呢是它同时把你的参考图和你的目标图去送到模型里面，也就是说。你的目标图它的所有的训练都是。和你的参考，它都是能够看到你的参考书。去进行一个修改。

youkengi(00:15:44): 他虽然方式不太一样，但是我觉得原理上很接近。不接近吗。

小志(00:15:48): 原理上也不接近，原理上也不接近，就可能最终它出来的效果你觉得比较接近，但是 contest 这种架构。它能够做到的准确性是要比差异炼丹要高很多，就是它，它的这个 FID 可以更低，这么说可能有些同学不是不是很明白，有些朋友不，不是很明白，就是他能够更好的还原原图。

youkengi(00:16:06): 对，这个是没错。

youkengi(00:16:19): 那这个是不是因为这个 contest 的参数就是参数量更强，所以才能做到这种还原呢？

小志(00:16:27): 我觉得我觉得是这个也可以，也有一部分吧，因为相信也不是很多人去尝试过在 context ，或者是说在 FIRST DV 上去进行一个差异炼丹，因为你要把一张图完全练到底膜上面，这个对于 B 来说其实还是比较难的，因为它的。参数规模会比较大。

youkengi(00:16:49): 一般性在那个就是比如说在 sdxl 上面练，其实也不会把底膜练到过里，其实就是把一个 lora 练到过里，然后而且只练一张图，然后做两张图的差异。这样会很就省时间很多，虽然效果就达不到，就这么好，但是在那个 SDSL 上面的差异基本上只能做到一些，就是概念类的这种差异变大。做不到，像这样的就是很精确的，比如说把 a 图里边的一个，比如说很精确的一个物体放到 B 图里面，这个这个在 sdxl 的上面是做不到的。

小志(00:17:32): 今天其实今天的分享是默认对，首先对这个对这个 stable diffusion 或者是对这个。有一定了解的同学去去听的，所以有同学觉得这个太学是听不懂是很正常的，因为它需要一些前提条件，比方说我省略了很多前提条件，比方说流匹配它是怎么样的，对吧？然后这个。first 他们原先的架构是怎么训练的对吧，其实这些都是被被省略掉，所以有有同学听不懂是很正常的，然后呢，我看应该专门出一些训练的。可以考虑，但是因为坦白说我训练还在摸索阶段，所以不太。不太好分享出来，因为我得到的一些结论也不一定是完全准确的，有时候一致性还不如 a 加加。

小志(00:18:45): 其实在我的分享里面也有提到怎么去提升这个一致性，比方说用这个例子，这个例子，对吧？其实它可以是做到很高的一致性，比方说中间中间的这一张图，但是呢，大部分的人情况去使用的情况呢，可能是像第三种这种情况，它提取了原图的一些元素，但是呢，它又不是在完全的还原这一个。

小志(00:19:15): 图一这种可能你可以通过一些位置的放摆放和这个提示词来提高它的一个一致性，然后呢在。稍等一下，中间的这张图，在扩图的情况下腿特别短，其实腿特别短。我们可以看一看。稍等一下。我们可以看看这张图，首先呢，它会 context 它训练呢，其实是对齐的位置信息，就是说你 t0的中间的这个像素是对齐这个你 t1中间这个像素的，也就是说在没有额外的训练的情况下。你用扩图的情况就是这个扩图的情况，他会去对齐你原图的一个信息。如果你的是一个衣服，假设说我们是一个衣服去生成一个模特，这个样子，你衣服的大小呢？它会。比较去遵从它原先的位置信息和大小，然后去进行一个生成，所以就造成了你的衣服的占比会比较大，然后呢，你的人的占比呢？可能就比较小。也是因为你原图参考的时候也没有人，所以就让 context 是去进行一个发挥，然后它就生成了一个特别特别短小的腿出来。

小志(00:21:14): 大概就是分享的这张图的样子。首先呢，你可以对比一下这个你的参考图跟你生成的图片，它的衣服的位置是不是对的特别准，这就是 context 它提供的一个。强而有力的一个一致性参考。但是呢，它又限制了去生成一个合适的人体比例。

小志(00:21:49): 然后 insertion 设计成 replace the area with reference 这样吗？可以这么说。可以这么说，我可以稍微分享一下我目前的一个训练级的一个情况，首先呢，我放出来的这个。训练集不对，我放出来的这个代码呢，是只提供了有限的单图训练的一个版本，我多图训练的版本还在我的。本地的一个 batch 上面，所以呢，多图训练版本是没有放出来的，但是可以看看我目前准备的训练机。

小志(00:23:03): 包括了特征和构图。可能对于一些特殊的用法，还是需要去进行特定的训练吧，我觉得就 conversation 到一个里面去进行一个生成，这是一个我觉得原原来它的官方的训练，也有考虑到这么一个使用方案。所以也是对于这个对其特征的一种运用吧。我目前的一个训练集大概是这样的一个灰度图，一个不是灰度一个灰色区域的图。然后呢，这个 mass 可以不用管，然后呢是一个衣服，然后呢是一个原图，然后我写的这个一个。评审呢，就是把这个灰色区域给替换成衣服，然后呢保持这个人还有它的脸，脸部特征，以至这种。

小志(00:24:10): 这种形式，然后刚刚也提到了，如果想要一次性，对吧，我们想要它有保持高度的一致性，或者说让他从参考图里面取得最大量的信息，我们就可以在参考图里面摆放不同的这个。组合合成的一个图片，然后让模型去配合提示词去进行一个生成。然后有很多朋友可能觉得 context 是不是图声图其实并不是 context 是文生图，它是从完全噪声里面去根据文字条件和。这个提供的参考图去一步一步把这个我们的最终目标图给生成出来，从这一个历史里面可以看到，它的一个训练和生成的一个范式是怎么样。

小志(00:25:13): 这张图是我参考了一篇论文里面的一个例子，去自己用，生成了两张图，然后还有这个自己家的一些噪声。所以他并不是这个图生图类似 IP，它会比 IP。有更多的信息吧，就常用的 ipad 通常是用这个图像编码器去提取图像特征，但是呢，这种图像特征呢是不完全的，或者是说和原图的相似度是不够的，因为它会更抽象一点。

youkengi(00:26:10): 我感觉他就是先看了答案，然后照着答案，按照你的要求去改答案。

小志(00:26:13): 声音有点声音有点小。可以大声一点。

youkengi(00:26:17): 我说我感觉这个参照图就是答案，他是看着答案，按照你的要求去改这个答案。

小志(00:26:24): 它是看着参考图加上文字条件，文字条件非常重要，它并不是完全的跟着这个参考图去进行一个生成的。那么。如果它完全跟着参考图去进行生成的话，它就变成了生成一朵白色的花出来了，对吧。但是呢，在我们有文字条件的情况下。它是能够比较好的去，或者是说比较差吧，也不也不也说不定，反正它是根据你的文字信息加上你的参考图进行进行一个生成。因为我们从这个官方的测试集里面就能够看到，它其实是有不同的一个修改的，一个 task 的，这些 task 呢，有些是需要比较对准的。的一个一致性的一个信息，但是呢，像风格转换，还有这个角色参考这些呢，它又并不需要说跟你的原图有，有一个很高的相似度，但是它有一个。比较抽象的一个相似结构上的一个相似。

小志(00:27:42): 其实 context 它也是基于像 is edit ic 罗还有各种各样的一个图像编辑基础上去进行一个训练的。

youkengi(00:27:42): 我有。

youkengi(00:27:59): 我有个疑问，他是怎么就是获取到参照图上面的图像信息的？

坏坏的薯片(00:28:07): 拼接在一起的吗？

小志(00:28:08): 可以再可以再大声一点，我这边比较吵我，我这边因为在训练，所以风扇比较吵。

youkengi(00:28:12): 我也大不了生小孩要睡觉了，我我说我有个疑问就是他是怎么获取到这个图像上的信息的，就参考图信息的。既然它是一个纹身图，那这个参考图的这个图像特征和信息。是以什么方式传输给他的？

小志(00:28:33): 因为他首先因为他训练的时候，他就把这个我们训练的生成的这个图片和我们的参考图进行一个拼接。但是呢，它不是一个水平上的拼接，它是一个通道上的拼接，所以你看到它这里是保留了原的原来的一个图片的一个 H，就是它的长宽。还有它的一个通道上，它加了一个它，它在它在位置信息上面加了一个通道，这个通道零呢，就是我们的生成的图片通道。一呢就是我们参考的图片。

小志(00:29:07): 然后这两张图片你可以理解成它是一个上下的一个拼接拼在一起，虽然它不是真的上下，它是通道的层面的一个评级。然后呢，它的一个实际情况呢，就类似于这样，他 t0和 t1是一起送到模型里面的，他在训练的时候呢，我们看右手边。这个训练的时候，我们看右手边在 T 完全是噪声的情况下，我们的 p1呢？完全是一个没有加噪的一个情况。

youkengi(00:29:44): 如果它是通过位置信息来传输，就就是通过那个训练集的调整，还有位置信息来做这个的，那刚刚就是 ppt 里面有一张图，就是它也获取到了这个紫色花几乎一样的特征。但是它的位置又改变了，那这个要怎么解释呢？

小志(00:30:05): 这就通过添加这个文字条件来进行一个解决，就比方说这个例子，这个例子，我们生成了三张图，我不对，我们生成了两张图。首先第一张图就是我们的一个参考图，然后第二张，第三张就是我们的不同的一个。

youkengi(00:30:13): 对，我说的就是这个。

小志(00:30:23): 生成出来的图片，然后呢，第二张的图片呢，是要比第三张的图片它更一致的，这是因为我在文字这个提示词上面去进行了一个描述，让它保持这个花的位置不变。还有一些其他的信息不变的情况下去生成它的下半部分。

小志(00:30:44): 提示词。

youkengi(00:30:45): 对我的我的疑问就是，为什么既然它是通过这个位置信息，那它位置应该是对应的这个像素像素的这个位置，那为什么它也能生成第三张图这样？

小志(00:30:57): 因为它因为它有，因为 context 它预训练的时候时候，它有这个不同的 pass，它并不是每一个 task 都会严格对齐你的参考图的。就比方说刚刚说到的这个风格的转换，它并不是严格参考你的这个参考图的，或者是说这个角色的参考，就可能你提供一个头像。然后呢，你的人呢？不一定是沿一个对要对应你的这个参考图的那个头像的大小和位置。

小志(00:31:34): 提示词，分辨率。提示是分辨率是是哪一个？能不能？在车上光打在车上不是？官方训练提示词其实其实可以参考我的一个接上的一个 report 里面去分析这个官方提供的测试集的一些视频的一个。频率，然后呢？首先官方的一个测试集，我稍微看一下在不在，对，这就是一个官方的测试集。它有不同的一个分类，比方说， style reference 我们看看这张图。首先官方它的一个提示词，我们看提示，先看提示词吧，它是大部分是一句话的，它并不会去训练特别复杂的一个提示词的关系和并并不会去训练在。一套提示词里面去进行一个多编辑的任务，因为这是很难训练的。然后我可以发一下我的那个，然后这个会议就已经快到时候了，快结束了。如果还有还有讨论的话，可能我会再开一个会议来进行一个讨论。这个。

小志(00:33:20): 换衣。场景它和 feel 模型训练有什么区别？首先在训练的好的模型上面，就是你训练了一个非常好的换音模型，和你的 context 训练了一个非常好的换音模型。根据我目前的测试来看，contest 是会取得一个更好的效果的 contest 它的一个还原的一个 FID 会更低，就是它的一些细节上面它会还原的更好。feel 是，就我的能力来说，是很难做到同样水平的一个细节的还原的。

坏坏的薯片(00:33:52): 我想。我想问一下为什么它会更好一些呢？

小志(00:34:03): 因为我觉得是因为训练范式的原因吧，就是其实是从一个完全噪声的情况下，根据你的一些辅助的方式去还原你的图片。但是呢，这个 context 呢，其实它是可以根据你已经有的一个图片去进行一个还原， content 是内置。

youkengi(00:34:32): 但是 contest 它在把这个图片信息传给这个去生成的时候，它还是转换成，唉，对，它还是转换成 L 的，他不是说直接把这个图片关来用。

小志(00:34:39): 还是转换成 later 还是转换成 later？但是它生成出来也是，也是，他就就对于我来说，我觉得它用 VAE 的一个 feature 是要不用这个。Image. feature 就是一些 image 的 feature 是要更好的，它能够提供的信息是更多。就我的理解不一不一定对。

youkengi(00:35:07): 知道，我个人感觉应该是参数量变大。

小志(00:35:12): 然后 content 是内置，为什么不是其实是另外一样东西，因为它是通过刚刚说的图像。编码器去把你的图像给提取一些图像特征，然后加到这个 test 的后面这个 test encode 的 feature 的后面，然后呢？去进行一个图像生成。

小志(00:35:39): contest 训练是一张图，对应多组 P 吗？ context 是一是两张图对应一个。就比方说用这个做例子，用这两个花做一个例子，这个提示词呢是把花改成紫色，我我们提供的呢，是一个白色的花和一个紫色的花。它的训练目标呢，是根据我们提供的这个提示词，把花改成紫色，和我们提供的白色的花生成一朵紫色的花，这么说不知道能不能理解。所以是一个提示词，对应两个图片。

小志(00:36:34): feel 用水平拼图的方式换一能打过 context 吗？我觉得如果两个都是训练的很好的情况下是打不过 context，就你可能已经有了一个用 few 的一个训练的一个模型。和原生的 context field 的模型做到的效果是要比你现在没有训练的一个 context 的情况要好的，但是如果你两边都是训练的很好的话。我觉得 context 它能够做到更低的 FID，就更好的还原度。

小志(00:37:14): 数据集数据集图片的精度对最终训练的效果有影响。当然有就是你，你的图片是怎么样子的，你的这个训练出来的模型就是怎么样？就如果你的训练图片可能你的 t1很好，因为你有原图，但是你的 t0不好，因为你是通过 context 或者是其他手段取得的 t0，它会在质量上有一些下降，那么你训练出来的模型就会带着那些下降的质量。

小志(00:37:45): 一起训练出来，所以你要训练高精度的一个高质量的一个模模型，最好还是用高质量图图像去进行一个训练。那么还有。对，继续。分辨率。

youkengi(00:38:01): 它这个对训练集的它这个对训练集的分辨率会不会有影响？比如说我的训练集是1024乘1024的一组图集，那我如果。

小志(00:38:11): 分辨率会有影响，我目前训练的 lola 是用512的一个分辨率，用1024会效果应该会更好，但是呢，需要进一步的测试，因为它会需需要更大的算力去进行训练。

youkengi(00:38:13): 对，我把它应用到。

小志(00:38:28): 用512训练也可以泛化到1024的一个分辨率。

youkengi(00:38:33): 就是他我的意思，他的泛化就假设现在用512来训练，那我把它泛化到1024的话，他会不会出现问题？他说他只有在512的情况下会较好。

小志(00:38:43): 它可以泛化，它可以泛化，因为我相信他们光要训练的时候也是用这种多阶段的训练，你可以，你可以尝试先训练512的模型，然后再在1024的这个分辨率上进行一个训练。进行一种这种后续训练，然后提高它的一个在1024上面的一个能力。5月二也可以。

youkengi(00:39:05): 我理解。有点像。有点像，先把概念练进去，再提高质量。有点像这样。

小志(00:39:18): 或者说我说一下我这个时间差不多了，如果还有这个想要讨论的话呢，我们可以开下一场会议，如果没有的话，那么今天就差不多到到这里了。还需要开下一场会议吗？没有微信群？明天继续吧，没有没有这么多。