2024年2月22日 星期四
责任编辑 谢丞宁 美术编辑 金浩翔

输入“繁花”能演出爷叔看宝总的表情？专家解读：

Sora赋能予人，不能代替人

OpenAI近期发布的首个文生视频模型Sora大幅刷新行业多个指标，重新定义了AI文生视频在现阶段的技术极限，颠覆了生成式AI在视频领域的全球市场格局，被公众誉为给2024年开年后的AI生成世界扔了个“王炸”。

相对于一些公众惊呼的“炸裂！”，上海高校的计算机专家指出，从文字转视频模型角度而言，Sora并不是横空出世的第一个，其问世不能算技术革命，但一定是工程实践的成功。虽然被普遍认为会改变未来的影视传媒行业，但也有教授反问：要输入多少提示词，Sora才能演出《繁花》里爷叔看宝总的表情？

青年报记者 刘昕璐

Sora文字生成视频效果截图。 网络截图

Sora是一种模态翻译的模型

输入提示词：“一位时尚女性走在充满霓虹灯和城市标牌的东京街道上。她穿着黑色皮夹克、红色长裙和黑色靴子，拎着黑色钱包。她戴着太阳镜，涂着红色口红。她走路自信又随意。街道潮湿且反光，在彩色灯光的照射下形成镜面效果。许多行人走来走去。”Sora即根据这段简单的文本指令，生成了60秒视频，头发、服装等细节都没问题，并实现了连贯的场景转换，达到以假乱真的水平。

复旦大学计算机学院邱锡鹏教授认为，此次发布是OpenAI的一次“炫技”，也是从宣传上提升自身实力和口碑的方式，总体上让各界感到十分惊艳和震撼，特别是视频时长的巨大提升让人称道。

邱锡鹏告诉记者，视频生成对机器学习来说是非常难的任务，视频与图像不同，图像相当于一种空间关系，视频又增加了“一维”，时空关系使得视频生成会更复杂。而在过去，虽然业界也有所尝试，但能生成的尚为“几秒级”。

上海交通大学计算机科学与工程系教授赵海教授表示，在数字化世界里，图片是一种连续性介质，就像水一样。而汉语、英语等语言文字都是离散的，两个字符之间没有模态意义上的连续性。图片和文字还有“二维”和“一维”的差别。因此，类似ChatGPT的纯大语言模型无法直接生成图片，但通常会调用一个文图生成模型来实现图片生成功能，如OpenAI发布的Dall-E。

文图生成器经改造后，也能生成视频。赵海表示，从文图生成器过渡到文生视频大模型，不需要很大的技术突破，研发团队主要依靠的是大算力、大模型、大样本训练数据等“先天条件”。因而，Sora的问世不能算技术革命，而是工程实践的成功。

“我们可以把Sora理解为是一种模态翻译模型。实际上，我们说的多模态混合模型，好比是你先上传一张图片，然后针对上传的这张图片，你再来提问题，相当于你在机器的输入端同时输入了图片和文字，然后模型用文字来回答问题，这个是标准的多模态混合模型。而Sora其实已经是在把自然语言直接翻译成视频，其中的性质差别还是很大的。因此，文生视频本质上是在做模态的翻译，直接把自然语言翻译成视觉信号。”赵海说道。

他认为，在机器学习里，如果输出要发生改变，其实很多内容都要发生转化，数据集、训练方式甚至模型架构都要有所突破。

从5秒拉升到1分钟视频的背后

在Sora问世前，同类产品其实已经出现，只是在工程上特别是算力数据准备和其他某些方面有不完善的地方，最显著的就是视频可生成的时长。

邱锡鹏指出，从模型来看，此次1分钟已经非常长了，一般以普通视频一秒分成24帧，拼起来，信息量非常之大。特别困难的是在时序上的关系处理，如果图像只是看一个空间关系，那么时序就涉及到场景里物体运动规律等，因此，不能轻视了增加的又一维，背后要处理的各种关系变得更为复杂。而且，时间越拉长，这个关系处理过程就越复杂，再通过海量数据的驱动，故而可以让新模型根据输入指令自动去捕获其中的关系。

“在Sora前，只能根据提示词生成5~10秒的短视频，如今一口气把时长提到了1分钟，这确实带来的场景效能是不同等量级的。”赵海同样惊叹这拉长的时长。

赵海认为，1分钟时长的视频能做很多事情，可以覆盖很多电影相对完整的一个场景了。

换言之，其工业价值就提升了相当的等量级，原来可以说能代替一小部分人的工作，但是，现在能代替更多的人了，而且是高附加值的那些人了。

另外，从Sora此次发布的作品来看，多角度镜头中，人和物能保持前后一致性，不会因角度变换出现问题。在对物理规律的掌握方面，Sora也有不俗表现，比如在其生成的一段SUV行驶视频中，汽车影子与车身始终契合。

影视包含视频、文字、声音等，是全模态的，从现在的文生视频来看，现在还是无声的，但是给它配上有声插件等，完整度就将更为显著地提升。一旦开放，的确，可以用很低的门槛直接进入正常视频、影视的创造生产，甚至直播带货、上课都可以用到这样的模型。

将改变影视传媒等行业

Sora会怎样影响行业，改变世界？

赵海认为，文生视频大模型将首先改变影视传媒等行业的游戏规则，特别是以技术含量最高的科幻电影为例，目前拍摄一部科幻大片往往需要数年时间，大模型用于这个领域后，特效画面的制作时间有望缩短到几天，从而大幅降低科幻电影拍摄的时间成本和技术门槛。与此同时，视频摄制成本的降低，也将给传媒行业带来变革。今后，部分视频的摄制也许只要在电脑前输入一些提示词，省去了现场拍摄、后期剪辑等耗时费力的过程。

“确实会带来好处，提升效率”，邱锡鹏对此持相同观点。邱锡鹏说，动画制作，需要先做各种模型。面对“一只狗在雪地里玩”的需求场景，设计师要先见过一只狗，对狗的毛发，雪地里的细微颗粒，都需要去建模，可以说，传统的动画制作成本高，周期长。有了这样的文生视频大模型，给出提示词，就可以比较好地用电脑虚拟生成场景，并且做到风格多变。“目前，Sora并没有对外公开，从长期看应该会对行业产生影响，短期来说，还只是‘炫技’了一下。”邱锡鹏说道。

很难想象，用多少以及什么样的文字作为提示词，才能让Sora精确地输出“爷叔”如此登峰造极的表演效果？这是复旦大学新闻学院传播学教授邓建国对Sora问世后的一次反问。对于AI一直有所思考的他，从媒介学角度第一时间“马上评”，做出了一些试探性的分析。

邓建国认为，即使Sora可以高效和逼真地输出很多个某一类型的长达60秒或以上的视频片段，即使这些视频片段能让某些自媒体或普通用户更便利更廉价地创作，某些类型的视频片段的表演仍然只有像游本昌先生这样的专业老戏骨才能胜任；能将这些片段以符合观众接受心理的方式流畅编织起来、讲述一个如《繁花》一样精彩和卖座的故事的，目前也仍然只能是由专业的讲故事高手通过专业（同时也是昂贵）的设备才能实现。尽管人类创造力的高峰已经被Sora等人工智能技术重重包围，步步逼近，但最高处的红旗仍将猎猎招展，高高飘扬。

积极拥抱去了解和掌握技术

新闻从业者又应该如何适应“Sora时代”？邓建国认为，新闻报道追求真实，而Sora几乎全是虚拟，因此，和ChatGPT不同，Sora从本质上对新闻业应用面不广，甚至只有坏处没有好处，除非新闻业沦为“创意业”。在“视频记者”这四个字中，重要的不是“视频”而是“记者”。如果记者没有脚力、眼力、脑力和笔力，而仅仅满足于坐在空调房里进行网络内容搜索和拼凑，或者不断使用人工智能炮制内容，那么这些记者在任何时候都应该被人工智能替代。

“在各种‘虚拟现实’技术盛行的今天，新闻业应该更加坚守‘现实’本身。如果主动放弃自己的‘现实’阵地不加区别地拥抱虚拟现实，这是新闻业自毁长城的失败，而不是虚拟现实技术所向披靡的成功。”邓建国说道。

面对AI界的每次变革，自媒体上都会有大量“唱衰行业饭碗”的论调，邱锡鹏强调“大可不必”。他认为，总体上，面对新技术，普通公众并不应该以害怕或者悲观回避的心态来去看待，而是应该积极拥抱变革，毕竟这些AI技术的更新迭代最终是为了指向“普惠性”，因此，了解和掌握这些技术，变得尤为重要。

邱锡鹏说，过去的动画、特效创作成本很高，并且工种划分细致，即使你有一个想法，也还要找很多专业人士帮助你把想法变成现实。在未来，我们就可以利用Sora这样的工具去帮助自己达成，让每个人都可以利用新技术在更大程度上尽展其能，就像使用搜索引擎一样，相当于让更多人的既有能力有了一柄“放大器”。

从现在AI发展趋势来看，邱锡鹏认为，个体的知识性能力可以更多靠辅助技术轻松达成，对个体而言，未来更多的是要学习和培养综合能力，懂得如何去调动资源来为自己去完成一件“大事”的能力。当可以用一些外在的工具合理帮助我们更高效处理事务后，便可更多地把潜在能力、思维活力释放出来。