OpenAI首款文生视频大模型对物理世界的重建能力惊艳世界

Sora“超级涌现力”将把AI引向何方

■吴飞

美国人工智能研究公司OpenAI最新发布的文生视频模型Sora，能够在接受人类输入的文本提示词后，生成一段长达60秒的视频，实现了内容合成从文本到图像、再到视频的领域跨越。

这一次次带来震撼的技术背后，都遵循着同一个原理：对合成内容中的最小单元进行有意义的关联组合。比如，在保持连贯的上下文语境中，对若干个单词进行有意义组合，从而连缀成一个会意句子；在保持合理的空间布局下，对众多图像小块进行有意义组合，拼合为一幅精彩图像；在保持一致的连续时空内，对一系列时空子块进行有意义组合，从而拼接成一段动感视频。

现实生活中，我们每个人都在通过有价值的内容组合来进行交流、设计和创作。唐代诗人卢延让对“吟安一个字，捻断数茎须”的感叹，讲的就是诗人从百千个候选字词中反复对比、精心挑选出一个合适的单词，从而写就一篇传世之作。南宋诗人陆游所说的“文章本天成，妙手偶得之”，惊叹的就是让词汇恰如其分地出现在了其应该出现的位置，形成语意连贯、文气贯通的天然佳作。

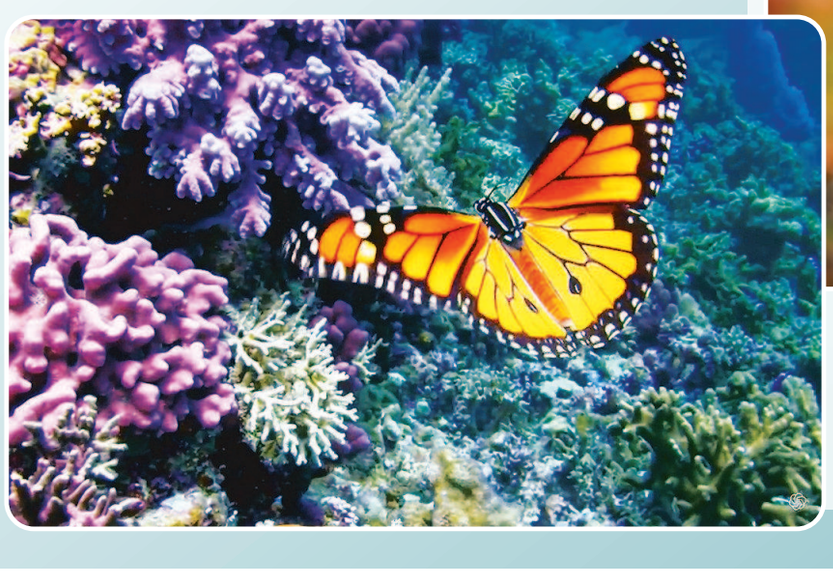
那么，从ChatGPT到Sora，人工智能（AI）大模型何以合成出有意义、有价值的内容？Sora所呈现出的“超级涌现力”将把AI引向何方？

最近，OpenAI发布的文生视频大模型Sora牢牢占据着科技圈头条。它的技术配方、其所带来的行业影响，以及“眼见不再为实”的全新风险，成为全球关注的话题。

堪比大片水准的Sora视频演示引发业界极大震撼，其所展现出来的能力几乎可用“碾压”来形容。人们不禁要问：从ChatGPT到Sora，人工智能（AI）大模型是如何实现迭代进化的？本报特邀浙江大学上海高等研究院常务副院长、浙江大学人工智能研究所所长吴飞教授为读者释疑解惑。

▶一段合成视频中，两名冲浪者在一座具有历史感的大厅里乘风破浪。

▲一位女性的秋日特写人像，细节模拟精致到位。



◀Sora模拟视频中，在海中飞舞的蝴蝶犹如实景拍摄。

共生即关联
从文本构建意义的网络

2017年，谷歌公司发表了一篇题为《注意力就是你所需的一切》的论文，提出了一种以自注意力机制为核心的神经网络架构Transformer。

只要给定足够多的句子，Transformer就可学习句子中单词与单词之间的共生关联关系。比如，“项庄舞剑，意在沛公”这样的句子在若干篇文章中出现，那么Transformer就会认为“项庄”“舞剑”“沛公”等单词之间存在共生关系，于是就在它们之间建立关联，这种关系被称为“注意力”。

可以想象，在对海量语料数据库进行学习的基础上，人工智能算法就可以建立起一个巨大无比的单词共生关联网络图。此后，每当人们给定一个单词，算法就可按照要求，从单词共生关联网络图中找到下一个与之关联关系最密切的单词，作为给定单词的后续单词——就这样一个个接缀合成出句子，最终达到自然语言合成的目的。因此，OpenAI公司CEO山姆·阿尔特曼曾说：“预测下一个单词是通用人工智能（AGI）能力的关键。”

那么，Transformer模型是如何被训练的？一般采用的是“完形填空”的方法，即如果模型所填单词与被移除单词不一致，说明模型尚未形成填空能力，于是可根据其产生的错误来不断调整模型参数，直至模型完美完成填空任务。在人工智能领域，这种“填空训练”的过程被称为“自监督学习”，即模型算法自己准备用来训练模型参数的“数据燃料”，自行按照预定目标进行学习。

为了让Transformer从预测下一个单词到具备“说人话、做人事”的能力，研究者提出了一种被称为“提示学习”的方法。在提示学习中，人类设计所谓的“提示样例”，来教人工智能模型学习如何更好地说话。

比如，“我很喜欢这部电影，因为电影呈现的剧情很精彩”“猫比大象要小，因此大象比猫更大”就是典型的提示样例。一旦设计提示样例后，算法将样例中后半句某个关键单词“移除”，然后让模型去预测被移除的单词。如此不断学习，模型就得以知晓在给出前半句后，如何更自然地合成后半句话。

为了进一步提高模型合成语言的性能，Transformer还引入了人类反馈中强化学习（RLHF）的技术，将在交流中人类对模型合成内容的反馈作为一种监督信息输入给模型，对模型参数进行微调，以提高语言模型回答的真实性和流畅性。

在“数据是燃料、模型是引擎、算力是加速器”的深度学习框架下，以Transformer为核心打造的ChatGPT涌现出统计关联能力，洞悉海量数据中单词-单词、句子-句子等之间的关联性，体现了语言合成能力。

在大数据、大模型和大算力的工程性结合下，ChatGPT的训练使用了45TB的数据、近万亿个单词，约相当于1351万本牛津词典所包含的单词数量。经折算，训练ChatGPT所耗费的算力，大概相当于用每秒运算千万亿次的算力对模型训练3640天。

GPT的出现为探索AGI的实现提供了一种方式，被誉为“AI的iPhone时刻”。英国《自然》杂志列出的2023年度十大人物中，首次将ChatGPT这位“非人类”列入榜单。

重建物理世界
并非简单“鹦鹉学舌”

人工智能程序一旦捕获了单词与单词之间的共生关联，就可利用这种关联来合成句子。那么，如果将图像切分为空间子块，或者将视频切分为时空子块，人工智能模型去学习这些子块在空间维度中的布局分布、在时间维度上的连续变化等信息，同时学习子块之间运动、颜色、光照、遮挡等复杂视觉特征，就可能重建、合成新的视频序列。

目前，合成视频需要先提供文本提示词，然后通过文本单词和时空子块之间的关联来合成新的视频。但因文本单词与视觉信息分属于不同类型，故而存在异构鸿沟困难，这是首先需要解决的难题。其次，还要克服由视频图像分辨率过大而带来的维度灾难，以及其所引发的操作上的挑战。

为应对这些挑战，Sora先将文本单词和视觉子块映射到同构低维隐性空间，在这一低维隐性空间中引入扩散模型，对视觉信息反复迭代，千锤百炼地挖掘文本单词、空间子块和时空子块之间的关联关系。

这种方式好比先通过“车同轨、书同文”，将文本、视觉等异构信息投影到同构空间，然后再通过“先破坏（添加噪音）”“再重建（去除噪音）”的迭代手段，来洞悉视频中各种不同单元在时间和空间中的关联关系，从而甄别和学习纹理、运动、光照、遮挡、交互等复杂视觉物理规律。

这就好比鲁班学艺，不断将大桥拆散再拼装，从这个反复过程中知晓它们的跨结构、支座系统、桥墩、桥台和墩台之间的组合关系，从而练就重建大桥的能力。因此，Sora合成视频的过程并非是简单随机的“鹦鹉学舌”，而是对物理世界的重建。

由此可见，尽管Sora并未使用与过往不同的新技术，几乎所有技术都是已经公开的，但其所用的视频生成方式对算力要求极高，而这种对算力和资金消耗极大的方式，大幅提升了同行跟进的门槛。同时，Sora利用GPT系统对提示词进行了润色与丰富，从而拉开了与之前文本生成视频模型之间的差距，形成了对手短期内难以跟进的优势。

Sora涌现力
自然世界“昨日重现”

Sora这次带来了多重惊喜：其一是具备合成1分钟超长视频能力。此前的文本生成视频大模型无法真正突破合成10秒自然连贯视频的瓶颈；其二是Sora视频是对自然世界中不同对象行为方式的“昨日重现”，比如能有效模拟人物、动物或物品被遮挡或离开/回到视线的场景，因此有媒体认为Sora是数据驱动下对物理世界进行模拟的引擎。

Sora对长时间视频合成的能力，来自Transformer能够处理长时间信息中最小单元之间的自注意力机制。例如，同样是基于Transformer的GPT4允许处理3万多个tokens（机器模型输入的基本单位），而谷歌最近发布的多模态通用模型Gemini 1.5 Pro就把稳定处理上下文的上限扩大至100万个tokens。

Sora之所以能对物理世界规律进行模拟，一个可能的原因在于大数据驱动下，人工智能模型体现出一种学习能力，即Sora通过观察和学习海量视频数据后，洞察了视频中时空子块单元之间所应保持的物理规律。

其实，人类也是基于对自然界斗转星移、节气变迁和昼夜交替，以及微观物质世界物质合成与生命演化的观测，推导出各种物理规律。虽然Sora很难像人类一样，将物理世界中诸如牛顿定律、湍流方程和量子学定理等，以数学方程罗列于人工模型中，但Sora能记住时空子块单元之间应遵守的模式，进而利用这些模式约束时空子块的组合。

理查德·费曼在《物理学讲义》中曾提及，在生物学、人类学或经济学等复杂系统中，很少有一种简洁的数学理论能与数学物理学理论中的数值精确度相媲美，其原因在于“其过于复杂，而我们的思维有限”，这被称为“费曼极限”。

数据驱动的机器学习由于其函数逼近能力，擅长从微观上发掘复杂系统的模式，以统计方法拟合高维复杂系统，被誉为神经网络模型的“涌现能力”。涌现性是一种结构效应，是组成成分按照系统结构方式相互作用、相互补充、相互制约而激发出的特征。

机器学习模型展现出的涌现能力具有重要的科学意义。因为，如果涌现能力是永无尽头的，那么只要模型足够大，类人人工智能的出现就是必然。当然，神经网络的涌现性目前仍然是一个开放的问题。

Sora的涌现力或许可以这样认为：在亿万个非线性映射函数组合之下，人工智能模型对最小时空子块单元进行各种意想不到的组合，合成出先前从未有过的内容。而这正是这一轮人工智能在数据、模型、算力“三驾马车”推动下飞速发展的必然结果。

▲Sora模拟生成的一只戴着贝雷帽、穿着黑色高领毛衣的柴犬

▶“SORA”云彩图像

人类始终是人工智能的『总开关』

■吴飞

在目前所看到的由Sora生成的视频中，仍存在一些违背物理规律的不真实内容，比如漂浮在空中的椅子，篮球穿过球框边缘线圈，行人突然消失等。这或许可以认为，Sora对物理规律的理解并非基于抽象后的普遍概念（如规则或方程等），而是“依葫芦画瓢”般的死记硬背。

造成这些问题，可能是Sora记住了训练视频数据中包含了由人工生成的这类情景不合理的视频，或者Sora在对时空子块单元进行组合时作出了统计意义上的“错误决策”。

虽然还存在这样那样的不足，但Sora使普通人都可通过自然语言这一简单明了的方式合成先前没有的视频，就已经使其成为改变影视创作、广告、设计等领域游戏规则的利器。

同时，我们也应该注意到，这次OpenAI公布的Sora合成视频所对应的提示词写得很精彩，颇具生动细节。这也说明，要想生成耳目一新的视频，必须具备独特的想象力和出色的语言能力。

因此，善于提出问题、设计场景和利用工具，也是我们每个人在从信息化时代迈向数智化时代需要不断学习和加强的能力。而这也是当下和未来教育所要特别关注的问题。

据说，Sora这一名字来自日本语单词“空”，寓意“无限之创造潜力”。如何培养学生形成如笛卡尔所言的“对人类思想字母表”进行创新组合的能力，为人类进步产生增量知识，正是当今教育所面临的挑战。

1605年，培根在《学术的进展》这一著作中，雄心勃勃地绘制了“人类知识全貌”，并在其中划分出了学校学习和阅读校勘两个领域，用以表示教育的两种不同手段。培根认为，阅读校勘以质疑和批判来创造新的知识，而学校学习则以知识传授为主，从而使人类的知识积累和传承更为有效。

然而，目前ChatGPT和Sora等人工智能（AI）系统可对人类全量知识进行整合，这显然对以知识传授为主的教育理念提出了巨大挑战。

事实上，早在1936年，爱因斯坦在纪念美国高等教育300年的会议上发表的一篇名为“关于教育”的演讲中就提到，“教育首要的目标永远应该是独立思考和判断的总体能力的培养，而不是获取特定的知识”。

当教育培养的目标不是“获取特定知识”，而是形成创造性解决问题的能力，那么站在AI时代的门槛上，我们就应该用进化的观点去看待这个过程，最大限度地发展种种可能性。

毋须质疑，未来将是人与AI共同进化的时代。人类和Sora、ChatGPT、AlphaFold等人造物将如影随形、协作共进、相得益彰。但是不论怎样，人类始终是人工智能高度、广度和深度的总开关和决定者，也是人和人造物的协调者。

数学家和哲学家诺伯特·维纳在1950年出版了一本极具洞察力和先见之明的著作《人有人的用处：控制论与社会》，目的就是希望人在技术世界的环绕中更有尊严与人性，而不是相反。这提醒我们不应陷入“人机相斗”和“人机相害”的臆想中，同时警惕将人工智能等同于人类大脑这种不切实际的想法，以及类似“人工智能奴役人类”的杞人忧天，利用好人工智能这一人类帮手，在人机协同中创造更加美好的未来。

Sora也能够生成动画视频，图为一个怪物家族的卡通视频截图，它采用扁平化的设计风格，包括毛茸茸的棕色怪物、带天线的黑色怪物、斑点绿色怪物和小小的圆点怪物等。（本版图片均为OpenAI官网视频截图）

Sora时代，如何守护信息安全？

■本报记者 孙欣祺

OpenAI首个视频生成模型Sora一经发布，便引起全球高度关注。但与此同时，人工智能（AI）技术的高速发展，也进一步加深了业内人士对其安全性的担忧。

美国加利福尼亚大学伯克利分校信息学院哈尼·法里德教授指出，“文本生成视频技术的快速进步，使我们愈发接近真伪难辨的境地。如果这项技术与AI驱动的声音克隆技术结合，那么人们可以深度伪造他人从未说过的话、做过的事。”

从网上流传的片段不难发现，在刻画复杂场景时，Sora生成的视频中仍存在一些不符合客观规律的错误。这说明，这些伪造视频到目前为止仍然肉眼可辨。但白帽黑客公司“社会认证安全”联合创始人雷切尔·托巴克表示，“Sora绝对有能力产出可以骗过普通人的视频，因为许多人尚未意识到篡改视频和修改图片一样简单”。

对于信息真实性方面的安全隐患，OpenAI发言人表示，在Sora正式公之于众前，公司将采取多项重要的安保措施。比如，OpenAI组织行业专家开展安全演习，同时公司也在开发工具，辅助检测误导性内容。此外，图像生成模型DALL-E 3配套的安全工具也将应用于Sora。

在政府层面，Sora的推出势必将加速信息安全监管措施的出台。日前，欧盟各国政府已批准《欧盟人工智能法案》，欧洲议会预计将于4月签署这项法案。如一切顺利，这套法规将于2026年生效。

据悉，该法案将禁止使用存在“不可接受风险”的AI系统，例如使用生物识别数据推断公众敏感特征的系统；高风险应用（如招聘和执法中使用的系统）必须满足一定条件，比如开发人员必须证明其模型对用户安全透明，符合隐私法规，且不存在歧视；对于风险较低的AI工具，开发者仍需在用户与AI生成的内容进行交互时明确告知。该法案适用于在欧盟地区运行的AI模型，任何违反规定的公司都可能面临金额高达其年度全球利润7%的罚款。

为执行该法案，欧盟委员会将成立一个AI办公室，负责监督通用AI模型，并由独立专家提供咨询。该办公室将研究相关方法，以评估这些模型的能力并监测其风险。不过，德国利希超级计算中心AI研究员杰尼亚·吉采夫提出质疑称，即使像OpenAI这样的公司遵守法规并提交数据，但公共机构未必有足够的资源充分审查提交的内容。