# **Sora带来了什么 还缺什么**

原创 Eagle 当下瞬间进入历史 *2024年03月09日 15:12* *新加坡*

**

Sora因为高保真的长视频演示，发布以后迅速掀起关注风暴。

OpenAI公布的技术细节很少，大家只能根据他们发布的两三页技术报告做各种猜测，另一可靠的依据就是他们列出的32篇引用论文，但究竟哪些技术被采用，以及实现中的各种技巧（trick）都很关键，甚至通常会公布的一些参数，比如上下文长度（context length）都没有透露。这也是最近OpenAI受到众多批评的原因，Elon Mask利用起诉的方式表达了强烈不满，Demis Hassabis最近也评论说，OpenAI是一家贡献创新想法少，但依靠工程和规模（scaling）成功的公司。

**Sora实现了什么**

在吸引眼球的图像和视频展示背后，Sora有哪些突破？

主要是出现了涌现现象。所谓涌现就是在训练中没有显式优化的能力在结果中呈现出来。Sora的涌现主要包括视频中的3D效果和一些直觉物理（intuitive physics）能力的习得。这些能力是建立物理世界模型和模拟世界的核心能力。直觉物理被认为是婴儿阶段在掌握语言前通过视觉学习的关键方法，也可能是掌握常识的重要方法。因为这些能力的出现，Sora被OpenAI定位成世界模拟器，并认为是通向AGI的重要一步。

Sora可能是视觉生成带来涌现的早例，所以也有评论认为Sora是视觉的GPT2时刻。

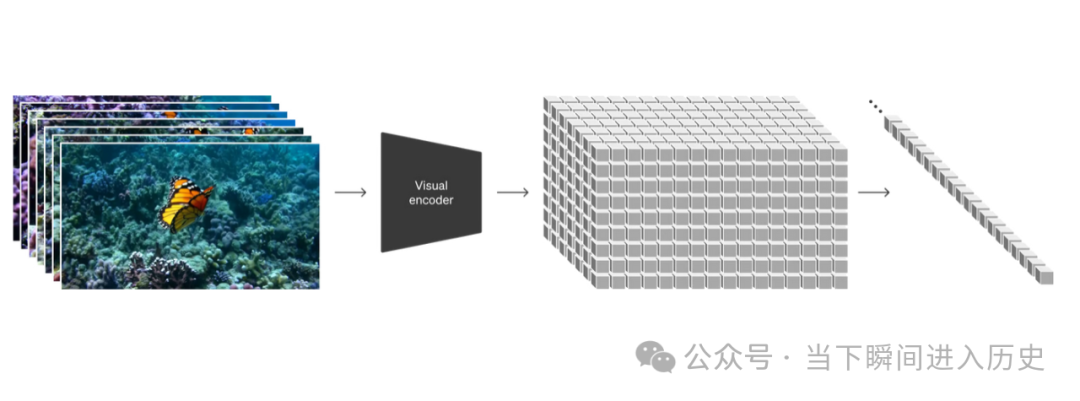
视觉的能力对智能的贡献是什么，好像还没有共识。Ilya曾经认为语言就足够带来AGI。Ilya已经沉默很久了，相信他是在思考更深远的问题，也希望早日看到他重新回到他理应可以贡献的地方。

视觉对于解决物理世界的问题肯定是必不可少的，对于理解物理世界是必须的。但似乎我们人类的智能并不依赖我们生成图像/视频的能力，这和语言直接依赖于找到下个词汇的能力是天然不同的，Ilya坚持走decoder only的GPT路线原因也在于此。

Sora今天还有大量的失败模式（failure mode），顾险峰有一篇精彩的分析，指出了可能是现有大模型/统计模型的根本限制，直觉感觉也可能是今天大模型方法遇到的样本困境（sample efficiency），也许自动驾驶类的应用还需要对失败模式有更好的理解才能真正到来。

Sora视觉的失败模式和语言模型的失败模式（hallucination）有一个显著的不同，似乎人类更容易判别视觉的失败模式，但对语言的幻觉很难识别，这对于一些容易人机协同的场景（man in the loop）会有不同的效果。

**Sora采用的核心技术**

Sora没有透露实现细节，只是粗略的提到基于Diffusion Transformer（DiT）技术，沿用了DiT在图像生成的技术路线，将它延伸到视频领域。DiT的一作Bill Peebles也是Sora团队技术领导。

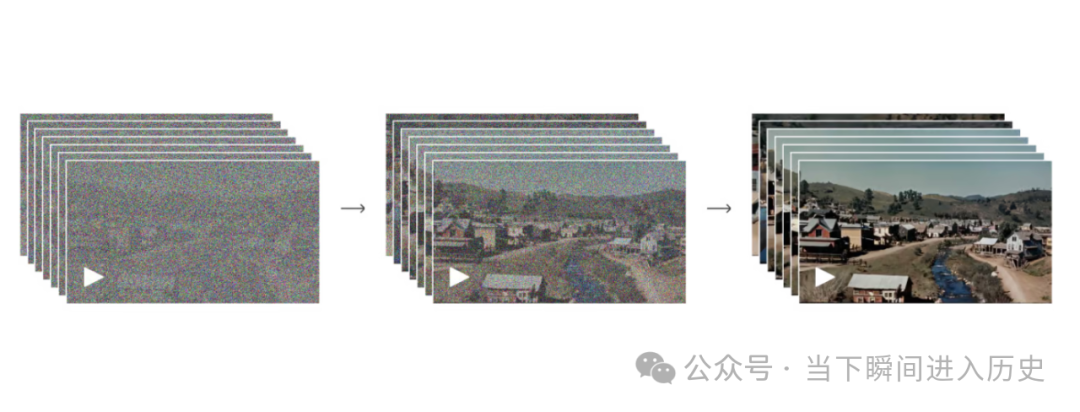
**向量化一切（Tokenize everything）**

其中关键的一步是将视频数据转换成Tokens，以便利用强大的Transformer处理，这个过程即Tokenize，Oriol说过Tokenize一切。

这个方法最早可能出现于Vision Transformer（ViT），开创了如何将图片切片（patchify）然后向量化（embeding）输入Transformer，完成了不同模态信号统一架构处理的奠基性工作。在ViT成功后，Ilya和Oriol曾在推特互动，认为CNN应该可以退出历史舞台了。

Sora视频中物体在长时间序列上的一致性也正是由Transformer从时空Token序列中学习到的关系。由于视频数据是超高维的，这种长上下文关系的获取至少在工程上是惊人的突破。Transformer有解锁长上下文的潜在能力，但需要的算力是和上下文长度的平方成正比的，OpenAI也宣示了凶悍的算力能力。其中优化技巧肯定很多，利用预压缩到隐空间（latent space）是能看到的，也许还有在计算时间相关的时候利用低分辨率图像计算。这里上下文长度参数就很关键，最近Gemini1.5和Claude3是公开的支持百万级别上下文的模型。

Sora在预处理数据的时候保留了数据原有的格式（长度，分辨率等），他们认为这是他们能够获得高质量结果的原因之一，变长视觉数据处理一定也使用了很多技巧。

**视频扩散模型（Diffision model）**

Sora的视频生成过程还是利用的扩散模型。这和LLM语言模型的方法是不一样的，LLM是用回归生成下一个Token的方法来实现的。

在回答扩散模型会发挥什么作用的时候，Ilya思考了一下后，认为扩散模型只是提供了更好的并行训练方法，他显然有点犹豫。

会不会有一天计算和生成都统一成Transformer？

扩散模型在图像生成领域这两年大获成功，但在DiT之前，主干的计算大多用CNN，DiT用Transformer替换CNN部分似乎是自然的事，尤其看Stable Diffusion的架构，都会疑问他们为什么不用Transformer？

这些年明显的进步包括Stable Diffusion把扩散过程从原始像素空间转移到隐空间（Latent space），大大降低数据维度，提升训练和生成效率；CLIP利用具有标注的图像数据训练，将语言和图像信息关联，可以利用语言作为条件控制图像生成；Dalle 3将原始的语言标注利用LLM扩展，形成详细语言描述，强化语言细节和图像细节的对应，进一步提升图像生成能力，也是利用合成数据的经典案例。Sora综合了这些技术，利用scalling大力出奇迹，在视频领域成功。

**Sora还缺什么**

如果目标是AGI，Sora的突破似乎不大。今天大模型缺失最多的还是所谓系统二的能力，Ilya认为今天人类在百毫秒级别能完成的任务，现有的大模型应该都可以解决，但对于人类长时间思考的问题，今天的大模型还缺两三个关键环节。

模型推理思考的深度正比于今天模型架构的层级，大约200级别，对于复杂的推理显然是不够的。DeepMind认为Alpha Go用到的强化学习方法，比如蒙特卡洛搜索（MCTS）等是一个值得期待的方向，但如何结合到大模型架构中，值得期待，希望几家能多开放一些。

对于今天根植于统计模型中的可靠性问题，包括语言幻觉和图像违背物理定律，似乎也只有经验主义的思路，能否有更深入的理解和解决方案还不清楚。

大家对AGI的到来时间已经明显乐观，多数认为是10年内的事，对齐AGI也还是开放问题。