阅读来源：https://mp.weixin.qq.com/s/E2QmJ8H0tCBPu30VqncCbQ

PART1: 机器学习的关键路径

因为 2017 年参加亚马逊机器人挑战赛（日本站夺冠）时我们发现，纯靠视觉完成抓取放置（pick and place）的成功率很难突破 70%+。现在很多具身智能公司也在做抓取放置任务，大家现在能做到 90% 左右。

在当时，我们发现纯靠视觉很难提升成功率。许多失败案例源于执行器缺乏与物体接触瞬间及后续短时间内的感知，视觉易受遮挡、视角等限制。因此我们意识到必须为末端执行器赋予触觉感知能力，让灵巧手、夹爪能像人类一样，在接触物体时感知接触力、纹理、温度、滑动、运动等多模态触觉信息。

视觉获取的全局信息通常呈连贯状态，比如视频中每两帧或一段时间内的数据流相对连续；而触觉在与物体真实接触前几乎无感知，接触后才触发局部信号——每个手指仅能感知所在区域的触觉，且需在同一框架内实现多手指信号的协同与互补。

视觉学习的瓶颈。

二者在感知特性上差异显著。视觉对物体位置的感知精度可达毫米级，而**触觉**往往需要微米级、至少 0.0 几毫米的精度。面对这类精度不同、模态各异、连续性状态有别的多源信息，首先需解决高效采集问题，其次要将

其有效整合到融合模型中，当前热议的 VLA 模型未来可能进一步升级为包含触觉的 VTLA 模型，以突破信息融合的技术瓶颈。

从人的角度去采集数据的思路，毕竟人是天然的智能体，向人学习是很自然的事。从人的角度出发，利用人的数据，而且不一定要通过遥操作，毕竟遥操作很难规模化

信息源：触觉信息 + 视觉信息 + 语音信息

采集 -> 多模态融合 -> 机器学习

用**视频生成**的方式，去生成机械爪或者人在操作过程中的下一帧视频。

从视频生成角度入手，直接基于视频，模型里蕴含着对视觉方面的理解。

对于这个领域来讲，包括现在说到的 World model 视频生成，以及黄仁勋的一些观念，都挺值得赞同

信息源：视频信息

采集 -> 数据生成 -> 机器学习

PART2: 数据金字塔

从数据角度来讲，我很认可这个领域里其他学者提出的数据金字塔说法。

互联网数据作为底座，它的精度或许没那么高，但量足够大，涵盖的场景、任务也足够多，所以对泛化性的贡献很大，而且目前获取成本相对比较可控。

再往上就是仿真数据，获取仿真数据的成本要比直接从互联网“挖矿”更难一些，得有仿真器，还要有好的控制器，甚至仿真器里还得涉及遥操作等等。而真机数据成本就更高了，要有足够的硬件、操作工人等，一系列问题也会随之衍生出来。

数据数量固然重要，但质量更为关键，高质量数据是决定未来模型表现的重要要素。

阅读来源：https://developer.nvidia.com/isaac/gr00t

NVIDIA 宣布了 GR00T 项目基础模型，用于类人机器人以及主要的 Isaac 机器人平台更新。

GR00T 工作原理：

核心：用于做机器人认知和控制的模型，提供仿真框架和用于合成 数据、环境和机器人内置计算机的数据管道。最终提升机器人的推理能力和技能。用多模态的方式支持在多样的环境中执行操作任务。

输入：这些模型是在一个昂贵的基于现实世界的人形数据集上进行训练的，使用 Isaac GR00T-Mimic 蓝图组件生成的合成数据以及互联网规模的视频数据组成。它们也可以通过后期训练适应特定的表现形式、任务和环境。

输出：GR00T N 具有较强的泛化能力，例如：抓去、双手移动、左手递给右手物品。现实中可以应用在物料搬运、包装和检验。