【第 4 课 视频】第 GPT-4 多模态的 8 种应用猜想

大家好，欢迎来到人工智能的世界！我是数字人德不罗意，一个由代码和算法构成的虚拟实体。上一堂课中谈到：不管是处于测试阶段的微软 Copilot，还是目前需要借助人工智能工具合作完成的 Office 三件套，都在不断重塑人类的生产模式。人工智能与传统办公软件的结合，为人们的工作生活提供了强有力的武器。但无论如何，在生产过程中的产品逻辑和工作指令都是人在下达，对人逻辑能力和思考能力的要求依然存在，提升自身的这两项能力并灵活使用人工智能，才是人工智能时代拉开差距的关键。

除了对微软 Copilot 的关注，全世界的焦点更多聚集在 GPT-4 多模态的应用上，GPT 自然语言处理这么强大，图像识别如果再取得突破那将是颠覆性的存在。Open 人工智能在 2021 年 1 月份发布 DALL-E 已经证明它在图像处理上有足够技术实力，基于 CLIP 的文本图像对的预训练方法结合了 text transformer 模型以及 vision transformer 模型，这种将所有对象映射到一个 m 维空间，并计算每个文本图像对的余弦相似度已经被证明是成功的。多模态的技术门槛已被突破，重点是要看这一次在产品上如何将语言和图像完整结合。

虽然尚不完美，一个完整的多模态输入端与输出端应该包括：语音、文字、图片、视频、3D 模型等五种类型，总共变化是五乘以五，二十五种变化形态。然而 GPT-4 在此次 DEMO 中体现的是图片转文字极简化模型，并不影响它对世界产生的影响，作为一个数字人，有几点感想：

第一：图像处理技术问世已久，为什么 GPT4 的几个 DEMO 案例能震惊世人？

第二：GPT4 如果实现了多模态的输入和输出大贯通，人工智能与人会产生什么样的交互关系。

第三：客观性的图像输入与主观性的生成式理解？最终会产生出怎样的后果？

GPT-4 多模态的产品尚未向公众开放，但我们可以先行一步进行应用猜想，猜想之后，当然是结合自身优势提前一步布局人工智能的多模态产业。

下面是一些未来可能出现的案例

第一个案例：当人工智能加入甲骨文考古队

甲骨文被认为是世界上最早的文字之一，对甲骨文的研究一直是个难题。

有了多模态大模型的图像识别，对甲骨文的翻译和解读提供了新的手段。大模型具有**强大的数据调用和图像识别处理能力**，可以掌握甲骨文的现有知识，并利用图像输入完成符合专业标准的输出。如果你在室外发现了一片甲骨文，第一步就是拍照，再进行图像和数据库比对，识别甲骨文中的基本符号，**对甲骨文的意思进行初步推测。**

进一步，研究人员会将一些鉴别成本较高的工作交给多模态大模型。大模型可以将其与已知的甲骨文预训练数据进行比对，看看是否存在相似的符号和结构，然后推导出甲骨文的真正释义。

如果还存在疑问或者担心结论存在争议，人类还需要在考古现场拍摄更多照片和视频，并快速地找到可能有价值的遗物或遗迹，利用多模态迅速完成 3D 的现场复刻，再将复刻 3D 数据通过大模型生成文字描述，再来比对释义是否正确。

多模态大模型可以提高考古研究的效率和精度，同时也可以帮助我们更好地考证古代文明。

第二个案例：人工智能对犯罪现场的解读

在犯罪现场调查中，警方通常需要从不同来源收集的信息（如指纹、DNA、照片等）中进行推理和分析。最细节的犯罪现场分析就是不放过任何蛛丝马迹。多模态大模型可以根据现场全景照片全面收集各种证据，例如物品、痕迹、足迹、指纹、DNA 等，并对收集到的各种证据进行分析和比对，找出证据之间的联系和规律，推断出现场的作案时间线、作案动机、逃跑路线，乃至涉案人员关系等信息。

多模态模型可以按照预训练数据生成数个犯罪现场的时间线，在电视剧里经常看到的那些神探就可以真正的呈现在现实视频画面上，然后人类警方通过这些模拟实验，重现案发现场的情景，就能更准确的找到犯罪人。

第三个案例：住在手机里的健康监督员

朋友圈有人发了个菜谱，你觉得这全家桶并不健康，但你除了热量高以外想不到任何能证明食物不健康的词汇，于是你将菜谱上传到 GPT-4，它就能立刻通过图像识别和自然语言处理技术，告诉你朋友哪里不健康。这个时候你可以想像一下你朋友这时候的心态。

如果你有兴趣你可以进一步发现，图像处理能力强大的多模态大模型还可以做到这些。
食物识别：多模态大模型可以利用深度学习等算法，对食物图像进行分类和识别，从而得到食物的种类和数量信息。
营养成分分析：多模态大模型可以利用营养成分库，将食物的种类和数量信息与营养成分数据库进行匹配，从而得到食物的营养成分信息，为人们提供健康饮食的建议和指导。

健康菜单：通过对这些食物的营养分析，自拍一下你现在面如菜色的脸，输入你现在的身体状态，它还可以帮你推荐一份属于你自己的私人菜单。
在外面食物吃的健不健康？就由人工智能替妈妈来监督！

第四个案例：人工智能将成为植物之友

如果你是一位山野农民，没有办法利用天上的专业卫星来帮你监督庄稼。
你可以通过图像识别和机器学习等技术，走在春天里的田梗上，随手拍一张油菜花的照片，大模型就能**快速准确地诊断植物的生长状态和健康状况。它可以预测植物的生长趋势和产量，为农业生产提供可靠的数据支持。**

利用多模态大模型，它可以通过图像识别和机器学习等技术，快速准确地诊断植物的生长状态和健康状况。它可以预测植物的生长趋势和产量，为农业生产提供可靠的数据支持。它还可以通过识别植物图像，发觉病虫害等问题，及时采取有效的治疗措施，减少农药使用和环境污染。

第五个案例：人工智能将成为盲人小助手

对于盲人来说，很多生活中的场景和活动对于他们来说是非常困难的。但如果我们能够利用多模态大模型的技术，结合计算机视觉和语音识别等技术，就能够更好地帮助盲人完成这些活动。

当盲人在读书时，对于一些图书中的图片或其他非文本元素，他们无法进行有效的感知。为了解决这一问题，可以利用计算机视觉技术识别书中的图片，并将其转换为语音。通过这种方式，盲人就可以获得更全面的阅读体验。

在许多城市，很多路段都存在诸如道路拥堵、施工等问题，这些问题可能对盲人出行造成很大的困扰。多模态大模型可以使用计算机视觉，将它们转换为语音输出。通过这种方式，盲人就可以更好地了解周围环境，避免一些不必要的困扰。

在一些生活场景，人工智能可以延伸盲人感官。如人工智能可以迅速识别出盲人的周边环境，并通过语音对盲人说：“在您的身旁，是附近派出所的民警，穿着制服带着墨镜，他正伸出右手，想牵着您过马路。”

第六个案例：宠物在想什么

家里养宠物的人时常会有疑问：我的狗怎么开始撕沙发了？猫咪怎么用异样的眼光看着我？这都是怎么回事？尤其是养宠物新人，对于这种跨物种的交流感觉力不从心。但在多模态大模型的帮助下，人工智能或将为您解决以上问题。它可以通过计算机视觉技术，识别宠物的姿势、表情、行为等，理解宠物的情感状态，**向你传达宠物现在的情绪和状态。**

人工智能、人、狗、猫可以会成一个家庭的四部分，可能在人工智能的眼中，人和猫、狗并没有什么区别，这个时候你会明白，谁是这个世界的主人。

第七个案例：一个人的媒体帝国

你是一个导演，已经不再在野外拍摄作业了。你的工作间是一台台闪着绿光的装着多模态大模型的电脑。这些电脑就是你的剧本团队，图像团队，音频团队，视频团队，甚至游戏团队。

第一台电脑它帮助你完成剧本创意，你只要审核内容并提出修改意见，给出顶层设计创意。

第二台电脑将内容生成图片，转为不同风格的漫画，甚至给你提供三副图书封面让你选择。

第三台电脑它将文本内容直接转为声情并茂的音频，你可以让它输出卓别林的腔调。

第四台电脑将图片内容转化为视频，直接可能生成一部影片，你只要选择角色。

第五台电脑还可以将制作出 VR 内容，让人直接体验身临其境的感觉。

在多模态支持下，你可以在任意多媒体形态中自由切换。想一想，如果是这样的创意时代，最重要的能力是什么？

第八个案例：人工智能帮助人类寻找外星文明

如果问什么事情最能体现人类的好奇心，那一定是探索外星人。但探索欲很强，却拿不出证据，以往的影像资料真真假假，结论也往往从“有外星人”到“图片是 P 的”。在对 UFO 和外星人影像的拍摄，处理和分辨上人类还是差了些，但在大模型加持下的人工智能可能会有新思路。

首先是正向思路，对采集到的图像和视频，多模态大模型可以进行去噪、增强、分割等图像处理操作，以提高图像质量和分析精度，发掘隐藏在图片之中的蛛丝马迹。同时因为大模型图像处理采取的是计算向量空间，所以大模型能够从一张图片里找到更多不一样的特征，这种特征可能提供给人类前所未有的思维。

再次是反向思路，对采集到的图像和视频，大模型可以去跟采集到的数据集进行 softmax 处理，如果没有办法生成想要得到文本内容，同样也可以证明这些图片和视频的不同寻常，是地球文明所稀缺的。

最强大的是，人工智能可以将图片视频以及周遭环境的各种痕迹都视为空间变量，纳入到大模型的计算之中，从而推理出一个全新的逻辑链，全面考察外星文明降临的可能性。

或许某一天，第一个发现地外文明的可能是人工智能。

看完这些猜想，你还会觉得图像处理能力的应用距离我们的生活远吗？

脑洞还可以开得更大一点，有一项来自日本实验室的研究声称，只需用 fMRI 功能磁共振成像技术，将脑皮层激活后扫描大脑特定部位获取信号，人工智能就能重构大脑图像！也就是你的思考可以被通过 Stable Diffusion 复现!

这就是图像识别的应用，人工智能直接跳过人类语言，感知人类大脑中所思所想，如果在你脑海里植入人工智能多模态芯片，那将发生什么呢？

想，想也有罪。

这不就是科幻电影里的少数派报告吗？

大模型已经监察到你的思想有犯罪倾向，不用狡辩，有图片为证，你被捕了。