元宇宙（5）：交互方式 1——用户交互

临菲信息技术港

自然的互动是提高 Metaverse（元宇宙）沉浸感的一个基本条件。它既可以重现朋友和名人的面孔来实现真实的互动，还能将用户的幻象放置到熟悉的和著名的地方。这种情况下，**暂时的离解、集中和增强是互动时的三个重要因素**。因此人的情感的控制、好奇和内在动机等情绪都被使用在了用户互动上。

此时，360 度的视野就是空间识别的兼听域。为了提高视频处理效率，需要大量的图像和失真校正。并且，为了减少晕动症和疲劳，则需要视觉和身体感觉的碰撞和替代的感觉。同时还需要多模态的感官感知，处理语言、手势和对话流。

语言交互

日常对话是一种通过语音识别传递用户意图的基本方法。换句话说，语言被用在各个地方，因为它在隐含的意义上简明地描述了复杂的情况。**在 Metaverse 里也同样需要创建一个环境，在这个环境中通过语言、抽象、问答和翻译来理解当前的情况。**

近来，有研究人员提出了 ParlAI，这是一个使用多任务训练、数据收集、人类评估和在线强化学习来训练和测试对话模型的综合框架。ParlAI 可以做到在同一界面上执行各种对话数据集的任务（例如，SQuAD，bAbI task，MCTest，WikiQA，QACN，QADailyMail，CBT，bAbI dialog，Ubuntu，OpenSubtitles，VQA）。

语言在强化学习领域被用来定义目标和抽象出人类可理解的任务，这也是最有效的一个方式。例如，AQM（Questioner's Mind）代理可以在面向任务的对话系统中最大限度地获取信息；知识图谱 A2C（KG-A2C）是一种可扩展的探索性方法，用于使用语言行为和动态知识图谱推断基于模板的工作空间中的游戏状态。当然，**在聚集了各种语言的人的 Metaverse 环境中，翻译也是一种必不可少的方法**。也有研究人员提出对大量未配对的语言和少量的语言对进行联合训练，这样来提高神经机器翻译的性能。

多模态交互

其实，人们的交流不仅是对话，还有基于多模态信息的方式（如面部表情、手势和语气）。处理每个模态的方法却很难处理多种复杂的情绪，所以多模态互动显得非常重要且必需。一般来说，多模态比单模态含有更多的信息，这对于理解当下的环境情况是有利的。**这就像社交媒体帖子中的文字和图片是在语义上有交叉的，图文往往并不具有相同的意义，而是具有更复杂的意义**。此时，多模态学习就更能突出优势。

举个例子，在变形金刚上映后，人们就开始进行研究怎样通过共同学习视觉和语言来减少使用预训练模型。有研究人员提出了一个统一的视觉语言字典训练（VLP）模型，其核心就是使用一个共享的多层变形器对视觉语言生成进行微调。

多任务交互

由于 Metaverse 要处理网络世界中的许多繁杂的情况，那么同时处理多个任务的模型会在处理复杂性问题的情况下尤为有用。对于这样的模型，**知识蒸馏**被用来做一个小模型，除了执行本职功能外还可以处理其他模式类型，例如，视觉 QA（VQA）。

显而易见，多任务比单任务要复杂，因为多任务模型要在有限的表达空间里平衡各种任务。人们就用端到端的方法来有效地执行各种任务。谷歌开发的 Translatotron 正是通过将语音语调输入翻译成其他语言的语音输出，它的优势很明显可以反映实际说话人语音语调，再以相同形式进行回应。与级联模型相比，端到端模型的优点是在整个过程中可以利用到大部分的输入而不会损失数据。

嵌入式交互

Metaverse 与其他一般互动的不同之处在于具象互动的比例相对较高，例如，嵌入式 QA（EQA）和视觉语言导航（VLN）。**这里的嵌入式交互其实所需的技术和 EQA 和 VLN 类似，但区别出现在主动交互还是被动交互一说**。虽说前面提到的 VQA 的目的是回答关于给定图像的文本问题，而 EQA 执行的任务就是分析通过主动探索物化的代理获得传感器信息。例如，为了回答一个关于远处汽车颜色的问题，代理会主动移动、识别，并根据汽车位置和路径的先验知识做出反应。

这样，人们可以以非语言的形式，通过指着一个物体而不是语言来交流信息。**当用户通过手指指向一个特定的位置时，它就可以成为一个预定的指令**。具体的指令则是以多模态交互的方式进行的，包括动作和语言。因此，有研究人员提出了嵌入式多模态交互（EMIL），这是一个反映活体启发机制的神经认知模型（例如，时间尺度的隐性适应）。

本部分元宇宙的内容介绍已完结，请关注下一期《元宇宙（5）：交互方式 2——元宇宙的实现》。