2152118 史君宝 脑认知 小作业

跨模态语言大模型　综述

1.引言：

近年来，ChatGPT（生成式语言大模型），凭借其出色的性能，引发了业界的广泛关注，成为了颠覆性的“通用人工智能”技术里程碑。ChatGPT 的主要核心是基于大规模无标签数据的预训练大语言模型，在指令微调、基于人类反馈的强化学习等众多技术加持下进行训练，并展现出了卓越的任务通用泛化能力。

经过广大用户的使用，相信人们已经认识到了 ChatGPT 的一些局限性了。虽然 ChatGPT 在纯文本自然语言处理任务上表现十分出色，但它无法直接实现对复杂多模态物理世界的认知以及通过交互对其产生影响。因此，为通用认知大模型引入多种模态的信息处理能力，无疑是通用人工智能技术发展的必然趋势。

纵观目前的大模型的技术演进历程，可以看到多种模态的信息处理能力正逐步融入预训练大模型体系中，随着 ChatGPT 等认知大模型的出现，研究焦点从面向特定任务的多模态感知，逐渐转变为更高层次的跨模态通用认知，整体技术演进呈现出如下图所示的三个范式转变：

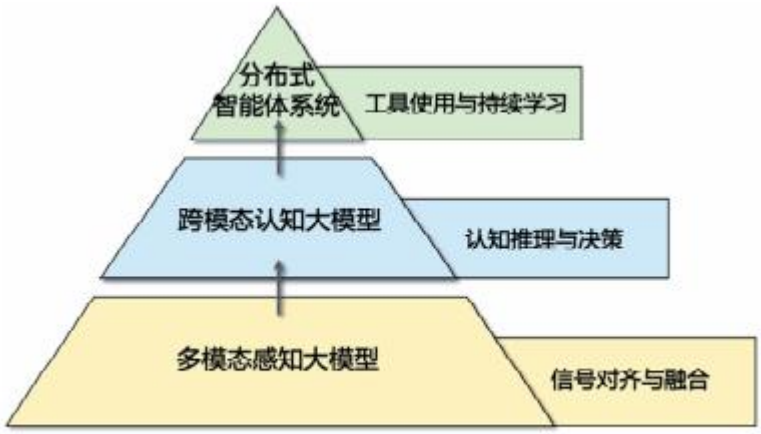



图 1　跨模态语言大模型三种范式概念关系图

2. 跨模态语言大模型发展历程：

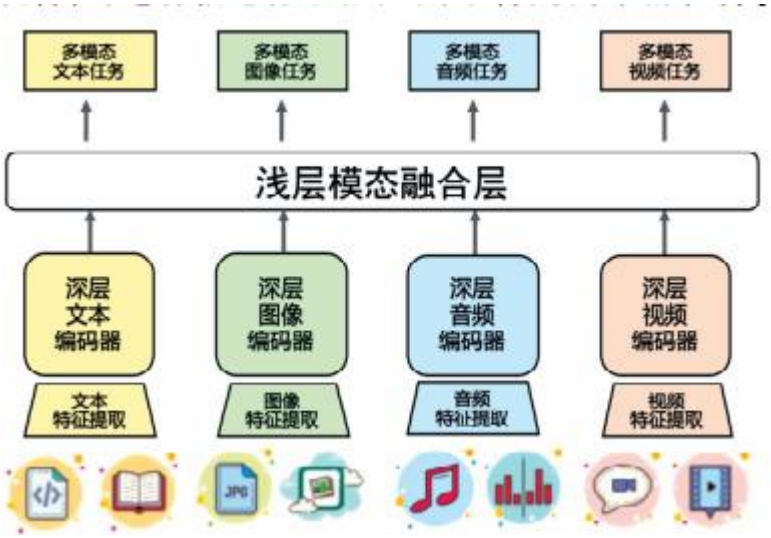
首先是多模态感知大模型，此范式的研究焦点是特定任务的多模态数据感知和分析。大模型从视听文等不同模态的数据中，独立平等地提取各模态通道的信号，然后再进行对齐和融合。

其次是跨模态认知大模型。语言大模型的出现促使研究重点从多模态感知向跨模态认知范式转变。这里的“跨模态”不同于各通道相对独立平等的“多模态”概念，其涵义是指各种模态信息被内生性的同时处理，不同模态信息在统一编码框架中自由交叉混合。

最后是以认知大模型为核心的分布式智能体系统。在此范式中，多模态感知能力和其它专业技能与黑箱认知大模型解耦，整个架构将语言大模型作为核心控制器，以自然语言或者形式化语言为接口实现大模型与外部多模态感知模型以及专业工具的信息交换，形成“１＋Ｎ”的分布式智能体系统。而构建出不断进化的分布式智能体系统。

3. 多模态感知大模型

多模态感知大模型的范式如下图所示：




它通过对每个模态的信号进行初步独立处理，提取各模态关键特征，进行信息融合后应用于特定的下游任务。依据多模态特征提取和融合交互方式的不同，通常可以分为双编码器、融合编码器、统一骨干网络三种主流架构。

3.1.双编码器架构

双编码器将不同模态的数据分别编码，然后仅使用简单的相似度匹配将模态表征映射到同一特征空间。一种典型代表是 CLIP 模型。通过在大量的“图像—文本对”数据上进行预训练，模型实现了图像与文本之间的联合理解和推理。CLIP 的文本编码器通常采用 Transformer，图像编码器可以选择（Vision Transformer）或 CNN。经过预训练后的 CLIP 模型在各种任务中，如图像分类、图像描述等，都表现出卓越的性能和泛化能力，例如 CLIP 在 ImageNet 上取得了 76.2%的零样本分类准确率。

3.2. 融合编码器

在模态融合阶段，CLIP 仅使用了简单的相似度匹配方法，无法充分综合多模态推理所需的不同模态信息。为解决这个问题，融合编码器架构在编码过程中就进行模态特征融合，以提取更深层次的跨模态特征。

融合编码器早期通常依赖于在特定任务上预训练的模型来提取多模态特征。随着 Transformer 在各个单模态任务上的广泛应用，它也逐渐成为融合编码器中通用特征提取架构的主流技术。

3.3. 统一骨干网络

尽管基于双编码器和融合编码器架构的模型在特定多模态处理任务上表现优异，但由于它们往往是针对单一任务进行设计、训练或优化，在解决多个不同的下游任务时，都需要进行额外的训练数据、模型架构、目标函数等调整，效率较低，性能调优复杂。因此，统一骨干网络应运而生，所有模态的数据在同一骨干网络中处理，以进一步增强不同模态间的交互和融合。同时，统一架构也利于处理来自不同下游任务的多种输出。

统一的生成式训练目标简化了模型设计和训练流程，同时提升了模型的泛化

能力和性能，在各种多模态任务上取得出色的表现。这些基于序列到序列的生成式统一骨干网络架构缩小了多模态任务和自然语言处理任务之间的鸿沟，为跨模态认知大模型的发展奠定了良好的基础。

4. 跨模态认知大模型

多模态感知范式下，虽然已经出现了多种融合方法和统一架构，但各个模态都被视为独立感知通道，总体被平等对待，以优化特定任务为主。

随着文本语言大模型展现出复杂推理和决策规划等通用认知能力，逐渐出现了以语言为核心的认知型跨模态语言大模型的研究趋势。跨模态认知范式下，各通道的感知从属于通用认知目标，语言大模型成为认知处理核心，借助统一的计算框架，在统一的语义空间内实现对所有模态信息的全面融合，进行全方位的理解、推理决策以及语义生成。

4.1. 基本框架：

跨模态认知大模型范式的基本框架如下：

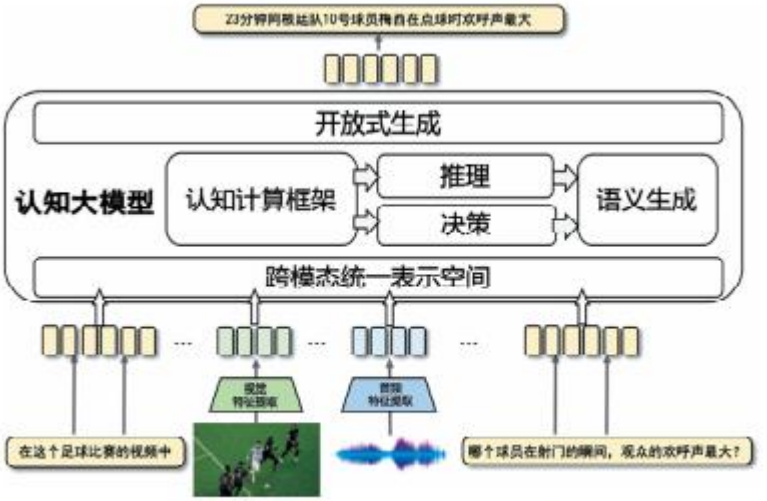



图 3　跨模态认知大模型示意图

跨模态认知大模型范式的主要思想是以语言为中心，将文本语言大模型用作处理各种模态编码的通用协议接口。这种策略可以理解为将不同模态的信息作为“外语”输入到语言大模型中，以实现联合建模。其优势在于能够将各种跨模态下游任务的预测统一转化为开放式文本生成，从而充分发挥语言大模型在小样本上下文学习和思维链推理方面的能力，实现处理复杂非特定任务的“通用智能”。

这种方法与认知科学的双系统理论相符，其中，各模态的编码器可以视为系统一，负责快速的模态感知，而语言大模型可以视为系统二，它对感知到的多模态信息进行深度融合，并通过内部的认知系统推理预测输出结果。

4.2. 跨模态指令微调

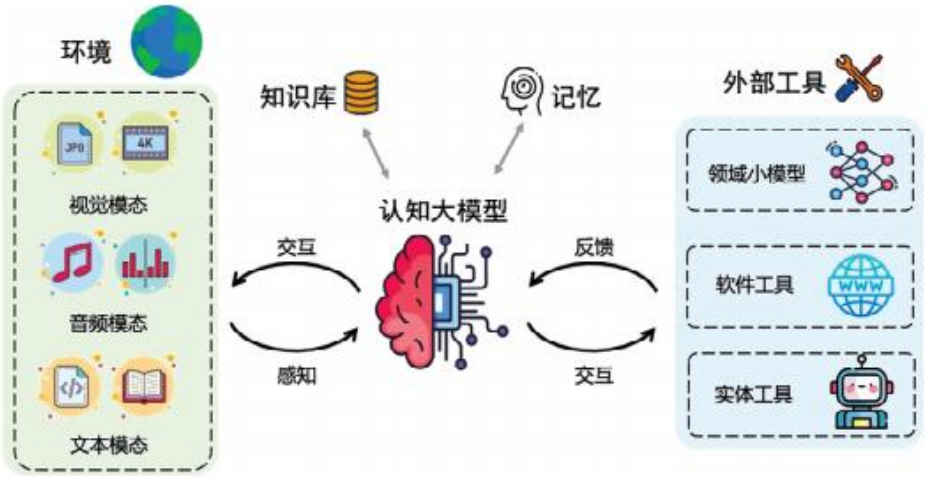
在自然语言处理领域，指令微调是一种对语言大模型进行微调的技术，其目标是使模型能够根据各种自然语言指令执行多样化的任务。这种方法首先通过自然语言指令详细描述任务，然后在这些与人类意图高度一致的指令数据上对语言大模型进行微调。这样做可以显著增强模型的泛化能力，并激发出如复杂决策和

思维链等显式认知能力。尽管现有的跨模态认知模型已经展示出了一定的认知和推理能力，但目前还缺乏系统的跨模态指令微调数据集。

5. 分布式智能体系统

端到端的跨模态认知大模型将不同模态的感知模块和语言大模型进行强耦合，虽然实现了较强的多模态感知和跨模态推理能力，但是仍然存在两方面的问题：一是扩展性较差，模型训练完成后不能动态加入处理其他模态数据的能力，二是不具有根据历史经验进行长期进化的能力。

为了解决上述问题，最近越来越多的研究开始将大模型的感知能力和认知能力解耦，以认知大模型作为中心控制器，通过灵活模块化的方式调用其他多模态感知模型或工具，构建起了分布式“感知—认知”智能体系统。这些系统通过与外部环境的交互实现持续学习，从而构建出一个能持续进化的分布式智能体系统，这种系统不需要进行大规模参数调整，就能实现多模态感知、认知决策和长期进化。




5.1. 外部工具使用

跨模态分布式智能体系统主要由几个核心部分组成：首先是作为控制核心的认知模型控制器，通常采用文本语言大模型；其次是各种模态的模块化工具，这些外部“工具”包括各个模态的预训练模型以及各类感知或执行工具；最后，系统还会与能够提供反馈的外部环境进行互动，这些“环境”可能包括待处理的图像、视频、音频，或者包含各种信息的可交互环境等。

5.2. 记忆增强的持续学习

除了能够学习使用外部工具，人类作为具备通用智能的智能体，拥有持续与现实多模态世界互动来适应新环境和掌握新技能的能力，这使得人类具备了终身学习的可能性。然而，现行的文本语言大模型只能依赖输入的上下文进行规划和推理，它们没有记忆和更新机制，不能从以往的经验中学习和累积知识，这限制了它们的持续学习和能力提升。

因此，为了建立一个具有持续学习能力的智能体系统，需要在大型模型系统中加入记忆、反思和长期规划的功能。GITM 和 VOYAGER 模型在文本语言大模型的基础上，结合了外部多模态工具的使用，进一步引入了记忆模块和互联网知识

库，成功构建了能进行持续学习的多模态智能体。这种智能体不仅可以设定并追求长期目标，还能通过与环境的持续交互，积累经验并进行自我反思，以实现技能库的不断扩展，这种模式代表了一种全新的终身学习方式。

总结：

多模态感知大模型主要针对不同模态的信号进行感知、对齐和融合，形成了双编码器、融合编码器、统一骨干网络等各种典型模型架构，模型的多模态信号处理能力得到了显著提升。

跨模态认知大模型使用以文本语言大模型为基座，在统一的语义空间内融合了各个模态的信息，实现了跨模态的语义理解和推理，在此基础上多模态指令微调则进一步提高了模型的认知和推理能力。

分布式智能体系统将大模型的感知和认知能力解耦，以认知大模型为核心，通过外部工具调用，加入记忆、反思机制，初步实现类人的外部工具使用能力和根据外部反馈持续进化的能力。

虽然跨模态大语言模型得到了快速发展，但是整个研究领域仍然处于初级阶段，面临着诸多挑战和机遇。