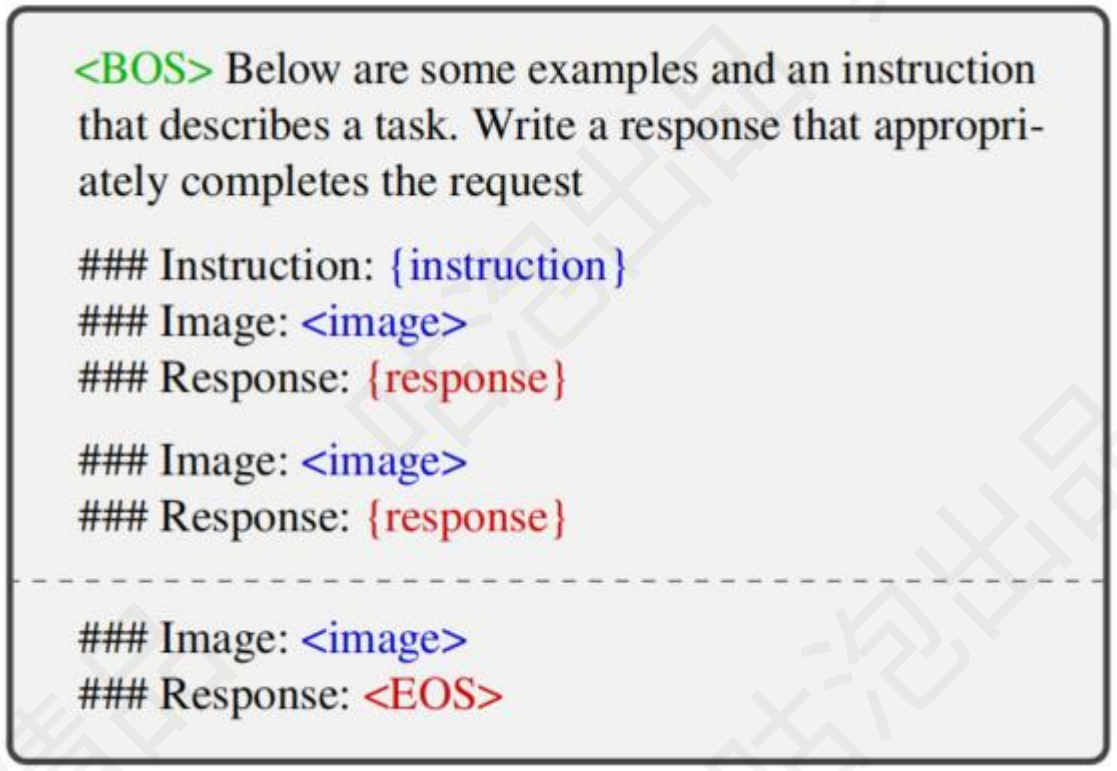
咕泡科技
GUPAO TECH

大模型实践-
多模态机器人

多模态上下文学习（M-ICL）



与传统的从丰富的数据中学习内隐模态的监督学习范式不同，ICL的关键是从类比中学习。具体而言，在ICL设置中，LLM从几个例子和可选指令中学习，并推断出新的问题，从而以少量的方式解决复杂和看不见的任务。

ICL通常以无训练的方式实现，因此可以在推理阶段灵活地集成到不同的框架中。与ICL密切相关的一项技术是指令调整经验表明它可以增强ICL的能力。

在MLLM的背景下，ICL已扩展到更多模态，从而产生了多模态ICL（M-ICL）。在推理时，可以通过向原始样本添加一个演示集，即一组上下文中的样本来实现M-ICL。

多模态上下文学习（M-ICL）

<BOS> Below are some examples and an instruction that describes a task. Write a response that appropriately completes the request

Instruction: {instruction}
Image: <image>
Response: {response}

Image: <image>
Response: {response}

Image: <image>
Response: <EOS>

多模态的应用而言，M-ICL主要用于两种场景：

- 解决各种视觉推理任务
- 教LLM使用外部工具

前者通常包括从几个特定任务的例子中学习，并概括为一个新的但相似的问题。根据说明和演示中提供的信息，LLM可以了解任务在做什么以及输出模板是什么，并最终生成预期的答案。相比之下，工具使用的示例通常是纯文本的，而且更细粒度。它们通常包括一系列步骤，这些步骤可以按顺序执行以完成任务。

多模态思想链（M-CoT）

通过融合特征或通过将视觉输入转换为文本描述

- Learnable Interface
这种方法包括采用可学习的界面将视觉嵌入映射到单词嵌入空间。然后可以将映射的嵌入作为prompt，将其发送给具有其他语言的LLM，以引发M-CoT推理。

- 专家模型
引入专家模型将视觉输入转换为文本描述是一种替代的模态桥接方式。

如何构建链条：

- 基于填充的模态
- 基于预测的模态

具体而言，基于填充的模态需要在周围上下文（前一步和后一步）之间推导步骤，以填补逻辑空白。相反，基于预测的模态需要在给定条件（如指令和先前的推理历史）的情况下扩展推理链。这两种类型的模态有一个共同的要求，即生成的步骤应该是一致和正确的。

LLM辅助视觉推理（LAVR）

LLM as a Controller

在这种情况下，LLM充当中央控制器：

将复杂任务分解为更简单的子任务/步骤

将这些任务分配给适当的工具/模块

LLM as a Decision Maker

在这种情况下，复杂的任务以多轮方式解决，通常以迭代的方式。决策者通常履行以下职责：

总结当前上下文和历史信息，并决定当前步骤中可用的信息是否足以回答问题或完成任务；

整理和总结答案，以方便用户的方式呈现。

LLM as a Semantics Refiner

LLM作为一种语义精炼器，研究者主要利用其丰富的语言学和语义学知识。具体而言，LLM通常被指示将信息整合到一致流畅的自然语言句子中，或根据不同的具体需求生成文本。

咕泡科技
GUPAO TECH

下次课预告

多模态大模型实践-VLA

To do list:

1. 阅读大模型书第七章
2. 扒本次课代码