**选题一、基于多模态大模型的视觉推理**

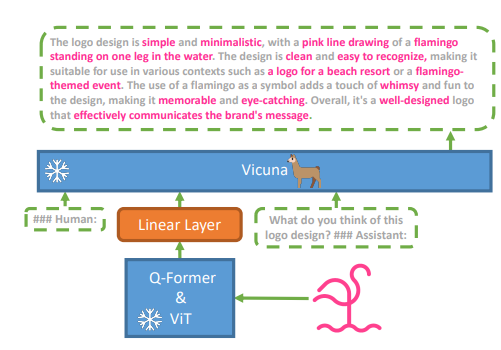
1. **作业要求**

**自由选择**开源的多模态大模型（Blip2, MiniGPT-4, LLaVA等）以及下游的视觉推理任务（VQA, image-caption），尝试对多模态模型做**简单的指令微调**后测试在下游任务中的性能。

**2. 可供选择的大模型**

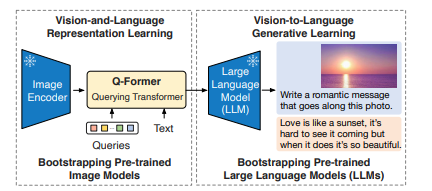
* MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models（https://arxiv.org/abs/2304.10592）（推荐）

<https://github.com/Vision-CAIR/MiniGPT-4/tree/main/minigpt4>



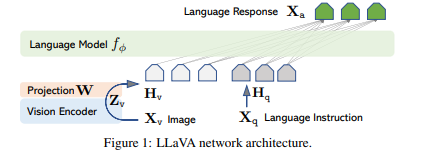
* BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models（<https://arxiv.org/abs/2301.12597>）（推荐）

<https://github.com/salesforce/LAVIS>



* Visual Instruction Tuning （Visual Instruction Tuning）

<https://github.com/haotian-liu/LLaVA>



可供选择的大模型包括但不限于上述三个模型，请同学们仔细阅读所选择模型的对应论文，并研究对应的项目GitHub。

**3. 可供选择的下游任务测试数据**

* **Image Caption**:

Flickr30K测试集 https://huggingface.co/datasets/nlphuji/flickr30k/tree/main

* **VQA**:

<https://drive.google.com/open?id=1_VvBqqxPW_5HQxE6alZ7_-SGwbEt2_zn>

OKVQA-testset <https://okvqa.allenai.org/download.html>

**……**

同学可以从中（当然也不局限于此）任选下游数据集进行测试。

**4．可供选择的指令微调数据**

COCO数据：<https://cocodataset.org/#download>

Flickr30K训练数据：<https://huggingface.co/datasets/nlphuji/flickr30k/tree/main>

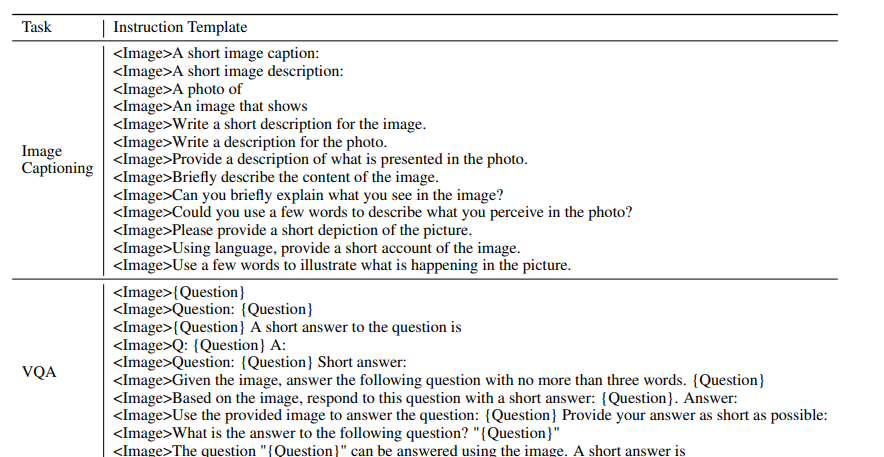
一些VQA数据：<https://visualqa.org/download.html>

具体选择哪些以及多大量级的指令微调数据不做任何限制，请同学们根据自己的目标下游任务选择合适的训练数据。

1. **指令微调方法**

* **数据预处理**

将训练数据按照指令模板构建成特定的格式（参考文章InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning <https://arxiv.org/pdf/2305.06500>的最后一页）：



* **微调参数**

在显存有限的情况下，可以只微调连接视觉模块和LLM的线性映射层（参考MiniGPT-4的文章和代码）；在显存比较充足的情况下，可以采用Lora的方式（LoRA: Low-Rank Adaptation of Large Language Models）微调LLM（参考MiniGPT-v2: Large Language Model As a Unified Interface for Vision-Language Multi-task Learning的文章和代码）。

**具体请参考你所选择的大模型所对应的论文和github库**