Build

processor

qwen2vl的不要动，主要改tokenizer，加<image_token>等特殊新token，把对话模板改一下，然后根据图片token数填充<image_token>等特殊新token。

processor最后要返回图片的两个参数，和input_ids，attention_mask

模型

代码里把类写好（config类和模型类），主要是处理Llama模块，加一个self.visual，forward里改一下数据处理inputs_embeds = inputs_embeds.masked_scatter(image_mask, image_embeds)

记得写一个initial_config.json用于初始化最初的模型，只用一次。

Pretrain

llava的数据就行，不管是trainer还是手动最后都收敛的一样

预训练阶段使用了“过滤后”的 CC3M 数据

都是描述任务（caption）

```
{'Describe the image concisely.',
 'Give a brief description of the image.',
 'Give a short and clear explanation of the subsequent image.',
 "Present a compact description of the photo's key features.",
 'Provide a brief description of the given image.',
 'Render a clear and concise summary of the photo.',
 'Share a concise interpretation of the image provided.',
 'Summarize the visual content of the image.',
 'What is in the photo?',
 'What is this?',
 'Write a terse but informative summary of the picture.'}
```

SFT

过滤一下数据，有的轮数太多了

QA任务，选择题，caption，视觉定位

训练数据不仅有图像文本多模态数据，同样也**有文本单模态的数据参与训练**

Data	Size	Response formatting prompts
LLaVA [28]	158K	–
ShareGPT [38]	40K	–
VQAv2 [12]	83K	Answer the question using a single word or phrase.
GQA [14]	72K	
OKVQA [33]	9K	
OCRVQA [34]	80K	
A-OKVQA [37]	50K	Answer with the option's letter from the given choices directly.
TextCaps [39]	22K	Provide a one-sentence caption for the provided image.
RefCOCO [17, 32]	30K	*Note: randomly choose between the two formats* Provide a short description for this region.
VG [18]	86K	Provide the bounding box coordinate of the region this sentence describes.
Total	665K	

COT

```
<SUMMARY/>`问题是xxx。`</SUMMARY>
<CAPTION>`图中xxx`</CAPTION>
<REASONING>
1.
2.
3.
...
</REASONING>
<CONCLUSION>xx<CONCLUSION>
```

RAG

基于Milvus

Embedding模型用**BGE-M3**

先用模型把文本和对应的图像做成langchain的document

然后利用Milvus生成数据库vectorstore

retriever查询即可

Quantization

根据激活值来做的，需要校准数据，训练数据的0.1-0.5%