BeforeTrain

Vision Encoder和Projector的准备

使用预训练InternViT-300M-448px-V2_5作为一个baseline，然后寻找一个1.5B左右的Qwen LLM，准备Projector进行两次实验。

冻结LLM，Projector不动，VIT一冻一不动，NTP作loss。

暂时先这样，github上找image-to-text的专业数据集。

takara-ai/image_captions

google-research-datasets/conceptual_captions

AntZet/men_women_children_wearing_clothes

CaptionEmporium/coyo-hd-11m-llavanext

jackyhate/text-to-image-2M

OpenFace-CQUPT/HumanCaption-HQ-311K

allenai/pixmo-cap

TheFusion21/PokemonCards

Attila1011/img_caption_EN_AppleFlair_Blip

AlexandrosChariton/cool_images_urban_and_nature

daniel3303/GroundCap

进一步简化实验：找使用InternViT-300M-448px-V2_5作encoder的vl baseline模型——InternVL2.5-4B，使用llama factory的freeze finetuning。

InternVL2.5-4B效果

AntZet/men_women_children_wearing_clothes 数据集的描述：

文本: this is an image of a little boy with glasses posing in front of a blue background wearing a spongebob t - shirt and black sweatpantss

简短描述: boy wearing glasses and a t - shirt with sponge

详细描述: this is an image of a little boy with glasses posing in front of a blue background wearing a spongebob t - shirt and black sweatpantss

InternVL的回答：

User: <image>

Please describe the image shortly.

Assistant: The image shows a child wearing glasses and casual clothing. They have a t-shirt with SpongeBob SquarePants characters printed on it. Black pants also feature a few SpongeBob designs, and the child is wearing black shoes with white soles. The background is a solid blue color.

User: <image>

Please describe the image in detail.

Assistant: The image features a young boy wearing a t-shirt and black pants with various cartoon character patches. He is standing with one hand on his hip against a solid light blue background. On the t-shirt, there's a character resembling SpongeBob SquarePants in the center. His pants have patches with different characters, including Patrick Star and others. He is wearing glasses, has short blonde hair, and is wearing black sneakers with white soles. The overall appearance is playful and coordinated with a cartoon theme.

InternVL在ft前的回答都比数据集的原始描述更优秀，所以需要更换数据集，排除 AntZet/men_women_children_wearing_clothes

TheFusion21/PokemonCards数据集的描述：

索引: 732

文本: A Stage 1 Pokemon Card of type Grass with the title Dartrix and 80 HP of rarity Rare Shiny evolved from Rowlet from the set Shiny Vault and the flavor text: A bit of a dandy, it spends its free time preening its wings. Its preoccupation with any dirt on its plumage can leave it unable to battle. It has the attack Sharp Blade Quill with the cost Colorless, the energy cost 1 with the description: This attack does 20 damage to 1 of your opponent's Pokemon. (Don't apply Weakness and Resistance for Benched Pokemon.) It has the attack Leaf Blade with the cost Grass, Colorless, Colorless, the energy cost 3 and the damage of 50+ with the description: Flip a coin. If heads, this attack does 20 more damage. It has weakness against Fire 2.

InternVL的回答：

User: <image>

Please describe the image shortly.

Assistant: The image is a Pokémon card for the species "Darix," a Stage 1 Pokémon. It evolves from "Rovulet" and has 80 HP. Darix is depicted as a bird with a blue head, white body, and red legs and beak. The card includes the moves "Sharp Blade Quill" and "Leaf Blade," with descriptions and stat information like the weight and height of the Pokémon.

User: <image>

Please describe the image in detail.

Assistant: The image depicts a Pokémon card for the species Darrix. It is labeled "Stage 1" and has an HP value of 80. The card text notes that it evolves from a Rovulet and contains the number 723.

The main illustration shows Darrix, which resembles a bird with a blue head, white body, red feet, and a tuft of red and green on its head. It has striking yellow eyes and is giving a quill-like wing a thumbs-up.

The "Sharp Blade Quill" attack does 20 damage to one opponent's Pokémon. It specifies that this attack should not be affected by Weakness or Resistance for Benched Pokémon.

The "Leaf Blade" attack describes flipping a coin; if heads, it increases the damage by 20.

Darrix is noted as a Blade Quill Pokémon and has height and weight details (Height: 2'04"; Weight: 35.3 lbs.).

The illustration is credited to Shigenori Negishi, and the card indicates that it is from the 'SV3' or 'SV94' set of Pokémon games.

开始冻结小实验

数据：philschmid/amazon-product-descriptions-vlm，loss: NTP

Llama factory的冻结操作：

#对于要冻结的参数

param.requires_grad_(False) # 冻结

#对于要训练的参数

if cast_trainable_params_to_fp32:

    param.data = param.data.to(torch.float32) # 只转换数据类型

requires_grad 保持默认的True(没用上，要转换所有的，保持一致)

数据处理：

首先是对话模板，其次是label和原始的input_ids的转化。

input_ids 的tensors 长度的最大化对齐，以及image_values的问题。

Llama factory使用APOLLO，没用上

结论：训练5轮后，模型能够对询问商品名字的特定问题，作出模板式的回答：The product name is :。

但是不能正确输出商品的名字。和Ray讨论后，认为数据的文本信息不是自然的长文本，同时商品的名字不是标准的品牌的信息。因此需要更换数据集。

继续寻找数据：pushparajanrahul/manfaclens_dataset(10条 哈哈），DBQ/Bottega.Veneta.Product.prices.United.States，DBQ/Burberry.Product.prices.United.Arab.Emirates，DBQ/Gucci.Product.prices.Australia，

用DBQ的Brand系列。

决定用这个了DBQ/Matches.Fashion.Product.prices.France，44,840条。无语，图片链接无效

！！！真正开始的实验！！！

用了这个DBQ/My.Theresa.Product.prices.France，大概有9万多条数据。

由于数据集没有caption，所以使用Qwen-VL-72B生成包含Brand信息的长文本，构造训练的文本。

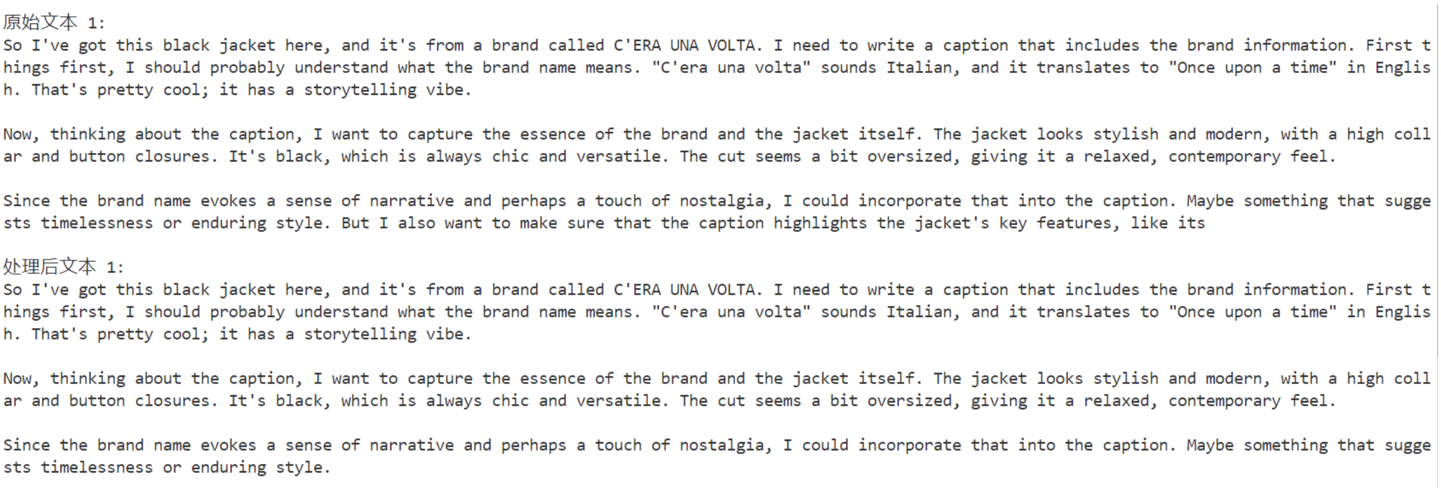
取2万条数据后，发现有5000多条异常文本，有循环输出的，乱码的。剔除异常文本，训练（设置和上面的实验一样）后发现微调后的模型能传递和文本内容差不多的信息，但还是无法传递Brand的信息或者传递出错误的Brand信息。

因此，更换生成文本的模型为Qwen/QVQ，并设置生成模型的参数，

```
    temperature=0.1, # 降低温度使输出更加确定性
   top_p=0.2,      # 降低采样范围，使输出更加保守
   max_tokens=200, # 控制回答长度
   presence_penalty=0.0, # 不鼓励模型谈论新主题
   frequency_penalty=0.0, # 不惩罚频繁词汇
```

检查了80/500的数据，发现brand信息全包含在文本里了，但是大小写有些会变化，同时，由于设置了最大tokens数量，所以导致文本内容戛然而止，解决方法：将最后一个“.”后面的文本删除。

处理结果示例：



```
原始文本 1:
So I've got this black jacket here, and it's from a brand called C'ERA UNA VOLTA. I need to write a caption that includes the brand information. First t
hings first, I should probably understand what the brand name means. "C'era una volta" sounds Italian, and it translates to "Once upon a time" in Englis
h. That's pretty cool; it has a storytelling vibe.

Now, thinking about the caption, I want to capture the essence of the brand and the jacket itself. The jacket looks stylish and modern, with a high coll
ar and button closures. It's black, which is always chic and versatile. The cut seems a bit oversized, giving it a relaxed, contemporary feel.

Since the brand name evokes a sense of narrative and perhaps a touch of nostalgia, I could incorporate that into the caption. Maybe something that sugge
sts timelessness or enduring style. But I also want to make sure that the caption highlights the jacket's key features, like its

处理后文本 1:
So I've got this black jacket here, and it's from a brand called C'ERA UNA VOLTA. I need to write a caption that includes the brand information. First t
hings first, I should probably understand what the brand name means. "C'era una volta" sounds Italian, and it translates to "Once upon a time" in Englis
h. That's pretty cool; it has a storytelling vibe.

Now, thinking about the caption, I want to capture the essence of the brand and the jacket itself. The jacket looks stylish and modern, with a high coll
ar and button closures. It's black, which is always chic and versatile. The cut seems a bit oversized, giving it a relaxed, contemporary feel.

Since the brand name evokes a sense of narrative and perhaps a touch of nostalgia, I could incorporate that into the caption. Maybe something that sugge
sts timelessness or enduring style.
```

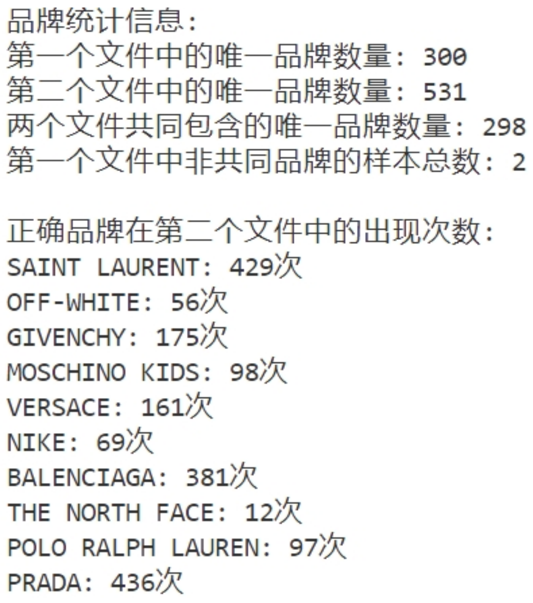
计划在拿到新的2万条数据后，只冻结LLM进行训练。

实验细节

第一次实验冻结LLM，训练20轮，每隔5次保存权重。累计梯度更新，batch_size = 24.

经过测试，发现第5次后产生过拟合的现象，第五轮在测试集上有26%的准确率（未训练的模型准确率是11%），之后的准确率更低。

分析发现数据集中每种品牌的样本分布不均，模型对于样本数量少的品牌训练不够。



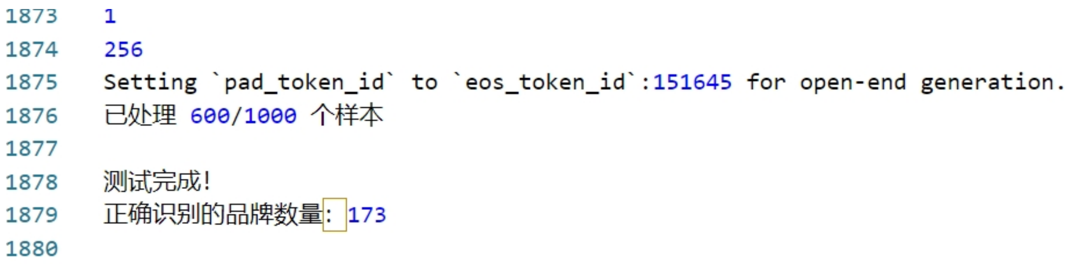
```
品牌统计信息:
第一个文件中的唯一品牌数量: 300
第二个文件中的唯一品牌数量: 531
两个文件共同包含的唯一品牌数量: 298
第一个文件中非共同品牌的样本总数: 2

正确品牌在第二个文件中的出现次数:
SAINT LAURENT: 429次
OFF-WHITE: 56次
GIVENCHY: 175次
MOSCHINO KIDS: 98次
VERSACE: 161次
NIKE: 69次
BALENCIAGA: 381次
THE NORTH FACE: 12次
POLO RALPH LAUREN: 97次
PRADA: 436次
```

第二次筛选出品牌样本数量大于100的数据，重新进行训练，冻结LLM。

此时的训练集从20000条减少到12000多条，测试集从1000条减少到600多条。

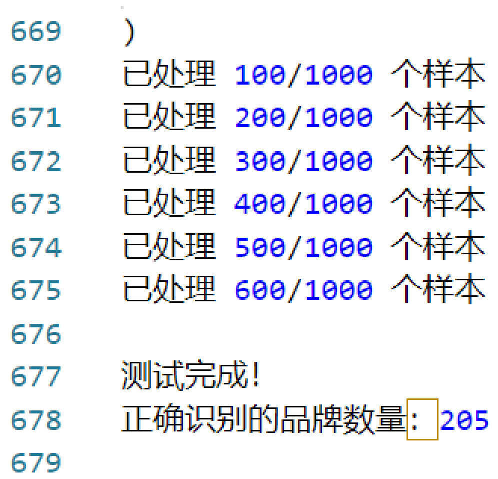
第五轮的测试结果如下，平均损失：0.6077，正确率达到28.3%：



```
1
256
Setting `pad_token_id` to `eos_token_id`:151645 for open-end generation.
已处理 600/1000 个样本

测试完成!
正确识别的品牌数量: 173
```

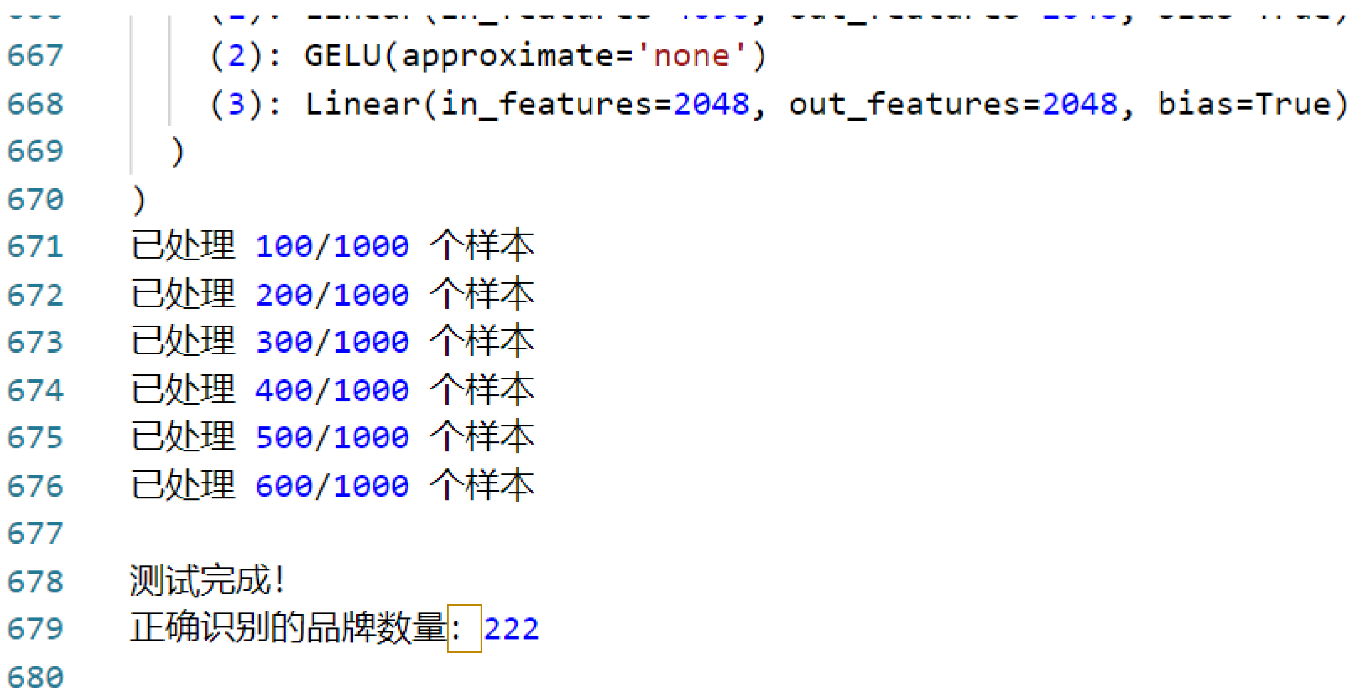
第十轮的测试结果如下，平均损失：0.3977，正确率达到34.1%：



```
)
已处理 100/1000 个样本
已处理 200/1000 个样本
已处理 300/1000 个样本
已处理 400/1000 个样本
已处理 500/1000 个样本
已处理 600/1000 个样本

测试完成!
正确识别的品牌数量: 205
```

第十五轮的测试结果如下，平均损失：0.1205，正确率达到37%：



```
      (2): GELU(approximate='none')
      (3): Linear(in_features=2048, out_features=2048, bias=True)
    )
)
已处理 100/1000 个样本
已处理 200/1000 个样本
已处理 300/1000 个样本
已处理 400/1000 个样本
已处理 500/1000 个样本
已处理 600/1000 个样本

测试完成!
正确识别的品牌数量: 222
```

第二十轮的测试结果如下，平均损失：0.0524，正确率达到36%：

```
    )
)
已处理 100/1000 个样本
已处理 200/1000 个样本
已处理 300/1000 个样本
已处理 400/1000 个样本
已处理 500/1000 个样本
已处理 600/1000 个样本

测试完成!
正确识别的品牌数量：216

```