Baseline BLIP2 com algumas fotos / perguntas

Do e-mail

caption: the model wears the jeans and a white t - shirt

prompt: Question: What is she wearing? Answer:
A white tee shirt and black jeans

prompt: Question: What is the color of the shirt? Answer:
Grey

prompt: Question: What is the shirt color? Answer:
Grey

prompt: Question: What is the color of the jeans? Answer:
Grey

prompt: Question: Describe the shirt? Answer:
The shirt is a cotton-blend shirt with a button-down collar, short sleeves, and a short hem.

prompt: Question: What are the shirt details? Answer:
The shirt details are the buttons and the

prompt: Question: What are the jeans details? Answer:
The jeans are made of a stretchy denim

incluida no treino com SaffalPoosh/deepFashion-with-masks

caption: the back of a woman wearing a white tank top and shorts

prompt: Question: What is she wearing? Answer:
A white crochet tank top and black shorts

prompt: Question: What is the color of the shirt? Answer:
White

prompt: Question: What is the shirt color? Answer:
White

prompt: Question: What is the color of the shorts? Answer:
White

prompt: Question: What are the shirt details? Answer:
The shirt details are the sleeves, the neckline, the back, the collar, the hem, the sleeves, the neckline, the collar,

prompt: Question: Describe the shirt? Answer:
It's a white knit tank top with a cutout back

prompt: Question: What are the shorts details? Answer:
The shorts are made of a soft, stretchy fabric that is perfect for summer

Testar perda de capacidade em outro contexto

caption: two cartoon monsters sitting around a campfire

prompt: Question: What is the dinosaur holding? Answer:
A torch

prompt: this is a cartoon of
two monsters sitting around a campfire

prompt: they look like they are
having a conversation

Testar perda de capacidade em outro contexto

```
caption: three people sitting on the beach with surfboards

prompt:  Question: Why are they standing around? Answer:
Because they are waiting for the waves to come in
```

Testar perda de capacidade em outro contexto

```
caption: a train is pulling into a station with people on the
platform

prompt:  Question: Is this a modern train? Answer
```

Foto do test set para caption

```
caption: the back of a man wearing a black and white floral print t
- shirt

prompt:  Question: What is she wearing? Answer:
A black and white floral print t - shirt

prompt:  Question: What is the color of the shirt? Answer:
```

Black

prompt: Question: What is the color of the pants? Answer:
Black

prompt: Question: Describe the shirt? Answer:
The shirt is black and white with a floral print on the front and back

prompt: Question: What are the shirt details? Answer:
The shirt is made of a polyester/cotton blend and features a floral print on the front and back

prompt: Question: What are the pants details? Answer:
The pants are made of a stretchy fabric that is designed to fit the body and the legs

Foto do test set para caption

caption: a man sitting on a stool wearing red pants

prompt: Question: What is she wearing? Answer:
A pair of red pants and a grey sweater

prompt: Question: What is the color of the shirt? Answer:
Red

prompt: Question: What is the shirt color? Answer:
Red

prompt: Question: What is the color of the pants? Answer:
Red

prompt: Question: Describe the shirt? Answer:
The shirt is a grey sweater with a red and black striped pattern

prompt: Question: What are the shirt details? Answer:
The shirt details are the collar, cuffs, and hem.

```
prompt:  Question: What are the pants details? Answer:
     The pants are made of 100% cotton and have a relaxed fit.
```

Resumo dos experimentos:

Experimento 1: Realizar finetune em image caption em um modelo generativo e testar aplicar o resultado em VQA

Dataset: "SaffalPoosh/deepFashion-with-masks"
Modelo base: BLIP 2
Código: Retrabalhado no experimento 2 junto com um novo dataset. Checar esse experimento
Modelo endpoint: Retrabalhado no experimento 2 junto com um novo dataset. Checar esse experimento

Resultado: Não satisfatório
Motivo do resultado: O modelo perdeu a capacidade de responder perguntas (VQA).
Observações:

```
outputs = model(input_ids=input_ids,
                    pixel_values=pixel_values,
                    labels=input_ids)
```

Segundo o tutorial do hugging faces o modelo recebe input_ids = labels. Não seria a label cortada ? (ex: 4 palavras iniciais)

Próximos passos:

- Revisar treino do modelo. Tentar fazer o input ser parte do label cortada (ex: 4 palavras).
- Revisar dataset
- Gerar um dataset de VQA:
 Opção 1: Utilizar um transformer para gerar perguntas através do texto
Opção 2: Gerar perguntas através de um dataset de classificação
- Avaliar se há outros modelos generativos com mais documentação/informações sobre o finetune. Avaliar se um desses modelos tem menos problemas em perder capacidade de VQA
- Avaliar viabilidade de treinar o modelo sem adaptadores (peft e LoRA) e se isso impacta em menor perda de capacidade
- Avaliar possibilidade de mesclar um dataset de VQA com um de caption e ver se eles mantem a capacidade

- Procurar informações sobre finetune em modelos multimodais e transferência de contexto com finetune em 1 task e uso em outra.

- Utilizar um modelo multilinguagem ou um transformer para traduzir prompts e respostas para português

Experimento 2: Continuação do experimento 1
Finetune em caption e aplicar a VQA

Dataset: Deepfashion Multimodal DeepFashion-MultiModal - Google Drive (Dataset 2)
Modelo base: BLIP 2
Código: Versão final ->" V2_Caption_Fine-tune BLIP2 on an image captioning dataset PEFT.ipynb"
Modelo endpoint:
- First epoch:
https://huggingface.co/leoreigoto/Data2_V2_BLIP2_Finetune_Caption_First_Epoch
- Last checkpoint:
https://huggingface.co/leoreigoto/Data2_V2_Blip2_Finetune_Caption

Resultado: Não satisfatório

Testes da baseline por época no arquivo do código

Observações:

Resultado está alucinando informações não reais na imagem. Descrição de cores está sempre como "cor sólida". Dataset tem descrições boas de cores? Precisa ser checado

Passa a descrever mais detalhes de roupas (apesar da alucinação), mas perde a capacidade de VQA (em contexto de roupas sempre retorna descrição). Treinar mesclando com dataset de VQA?

Cortar input cortado não é válido. Ele precisa ter mesmo tamanho do label.
Também testado colocar texto parcial em :

```
encoding = self.processor(images=image, padding="max_length",
return_tensors="pt
```

```
encoding = self.processor(images=image,text= caption[:4]
padding="max_length", return_tensors="pt
```

Sem sucesso (mantido sem texto parcial).

Outros modelos generativos também não tem documentações / exemplos de finetune. Não foram encontradas muitas informações sobre treinar modelos multimodais. Não foram encontradas menções a gerar um novo contexto em uma task com um dataset e aplicar esse novo contexto em outra task (finetune em Caption e uso em VQA).

Não foi possível alocar esse modelo em memoria para treino sem utilizar peft e LoRA. Na documentação menciona múltiplas GPUS para treina-lo.

Próximos passos:

- Revisar dataset
- Gerar um dataset de VQA:
 Opção 1: Utilizar um transformer para gerar perguntas através do texto
Opção 2: Gerar perguntas através de um dataset de classificação
- Avaliar possibilidade de mesclar um dataset de VQA com um de caption e ver se eles mantem a capacidade
- Testar gerar descrição com o modelo gerado nesse experimento e utilizar um transformer de Q & A para responder perguntas baseado na descrição gerada
- Utilizar um modelo multilinguagem ou um transformer para traduzir prompts e respostas para português
- Retestar experimento 2 com learning rate menor (parece estar dando overfit)

Experimento 3: Continuação do experimento 2, utilizar um transformer de Q & A para responder perguntas baseado na descrição gerada

Dataset: Não aplicável
Modelo base: timpal0l/mdeberta-v3-base-squad2
Código: Não aplicável
Versão final: modelo base (não foi feito finetune)
Modelo endpoint: modelo base (não foi feito finetune, apenas testado o conceito)

Resultado: Melhoria em relação ao experimento 2.

Testes da baseline por época no arquivo do código

Observações: Pode ter potencial com uma melhoria no modelo 2. É necessário mais descrições de cores e afins.

Retestar experimento 2 com learning rate menor (parece estar dando overfit)

Próximos passos:

- Revisar dataset
- Gerar um dataset de VQA:
 Opção 1: Utilizar um transformer para gerar perguntas através do texto
Opção 2: Gerar perguntas através de um dataset de classificação
- Avaliar possibilidade de mesclar um dataset de VQA com um de caption e ver se eles mantem a capacidade
- Utilizar um modelo multilinguagem ou um transformer para traduzir prompts e respostas para português
- Retestar experimento 2 com learning rate menor (parece estar dando overfit)

Experimento 4: Geração de dataset de VQA e finetune com VQA

Dataset: Gerado a partir de In-shop Clothes Retrieval Benchmark - Google Drive
Arquivos em excel tratados em : https://drive.google.com/drive/folders/19PBosKXr-VlVWzjL72mb_yOdxdVmfy5Q?usp=sharing (dataset3)
VQA é gerado no código
Modelo base: BLIP 2
Código: Versão final -> V3_Data3_finetune_BLIP2_on_a_VQA_dataset_PEFT.ipynb
Modelo endpoint: leoreigoto/Data3_V3_BLIP2_VQA · Hugging Face

Observações:
Prompt para vqa : "Question: <pergunta> Answer:"

Foram feito testes com label "<resposta>" e não deram certo. Também tentei realizar um padding para gerar inputs e prompts com

```
encoding = self.processor(images=image,text="Question: <pergunta> Answer:", padding="max_length", return_tensors="pt")
```

Mas não deram certo.

O resultado de acordo com testes foi

```
encoding = self.processor(images=image, padding="max_length", return_tensors="pt")
```

Junto com Input_ids e label: "Question: <pergunta> Answer: <resposta>"

Não foram encontradas documentações sobre como dar finetune nesse modelo para VQA ou em outros generativos. Ainda está longe de funcionar, mas também é preciso testar um novo dataset.

Após pesquisas parece que deveria ser considerado um conjunto de possíveis respostas e gerar um dataset com múltiplas opções e Cross-entropy loss. Avaliar também utilizar um modelo mais fácil como ViLT com esse possível dataset.

O dataset atual não é recomendado para isso. A classe da roupa não está bem definida (por exemplo classe "denim"), atributos teriam muitas respostas possíveis. Cores estão muito dívidas e com muitas cores com poucas imagens, maior parte inclusive tendo 1 imagem apenas. Pode ser reavaliado gerar perguntas através de labels de datasets tradicionais de classificação ou detecção de objetos.

Próximos passos:

- Revisar dataset (caption)
- Gerar novo dataset de VQA:
Opção 1: Trocar abordagem um transformer para gerar perguntas através do texto
Opção 2: Continuar a gerar um dataset baseado em problemas de classificação atributos. Procurar um dataset mais adequado.
- Avaliar possibilidade de mesclar um dataset de VQA com um de caption e ver se eles

mantem a capacidade
- Utilizar um modelo multilinguagem ou um transformer para traduzir prompts e respostas para português
- Retestar experimento 2 com learning rate menor (parece estar dando overfit)

Github: https://github.com/leoreigoto/VQA_finetune/

HF Space: [leoreigoto/Data2_V2_Blip2_Finetune_Caption · Hugging Face]