

大数据原理与技术作业六

图文多模态问答 (VQA)

【学号】 22336259

【姓名】 谢宇桐

【专业】 计算机科学与技术

【实验要求】

- 使用ViLT预训练模型 (Hugging Face库)
- 尝试调用, 输入图片+问题 (如“图中有什么动物? ”, 也可以自定义问题), 输出答案

【实验内容】

- 由于网络问题, 无法在线访问hugging face, 于是我在镜像hf网站上下载了dandelin/vilt-b32-finetuned-vqa模型到本地。
- **ViLT (Vision-and-Language Transformer)** : 一种多模态深度学习模型, 专门设计用于视觉与语言的联合理解任务。它通过统一的Transformer架构直接处理图像和文本, 无需复杂的视觉特征提取步骤 (如Faster R-CNN), 显著提升了效率和性能。

```
from transformers import ViltProcessor, ViltForQuestionAnswering

# 模型所在本地路径
model_path = "./models/vilt-b32-finetuned-vqa"

processor = ViltProcessor.from_pretrained(model_path)
model = ViltForQuestionAnswering.from_pretrained(model_path)
```

图片我选择如下:



- 定义视觉问答函数 vqa()

```
def vqa(image_path, question):  
    try:  
        # 加载图片  
        image = Image.open(image_path).convert("RGB")  
  
        # 预处理图片和问题  
        encoding = processor(image, question, return_tensors="pt",  
truncation=True, max_length=40)  
  
        # 模型推理  
        with torch.no_grad():  
            outputs = model(**encoding)
```

```

# 获取预测答案
idx = outputs.logits.argmax(-1).item()
return model.config.id2label[idx]
except Exception as e:
    print(f"推理错误: {str(e)}")
    return "无法回答问题"

```

- 因为此模型经尝试之后需要用英文提问回答才有效，因此我用英文进行提问：

```

if __name__ == "__main__":
    local_image =
r"C:\Users\85013\Desktop\a298552618de948b0c1554a7fe36d72.png"

    en_question = "What kind of dog is in the picture?"
    answer = vqa(local_image, en_question)
    print(f"Q: {en_question}\nA: {answer}")

```

输出如下：

```

Q: What is the animal in the picture?
A: dog

```

进程已结束，退出代码为 0

```

Q: What's behind the dog?
A: flowers

```

进程已结束，退出代码为 0

```

Q: What kind of flowers are in the picture?
A: sunflowers

```

进程已结束，退出代码为 0

Q: What kind of dog is in the picture?

A: husky

可以看到都回答正确。

至此，实验结束。