北京郵電大學
Beijing University of Posts and Telecommunications

神经网络与深度学习课设

开题报告

题目：图像描述生成：基于编解码框架的方法

组长：　邓圣曦 2022212222

组员 1：　董光硕 2022212155

组员 2：　张宇昊 2022212264

日期：2024/11/17

1. 任务描述

本项目旨在实现图像描述生成任务，通过深度学习模型自动为图像生成流畅且关联的自然语言描述。具体任务包括：

(1) 实现两种模型结构：CNN/ViT+GRU 和 网格/区域表示、Transformer 编码解码。

(2) 通过 METEOR 和 ROUGE-L 评测标准。

(3) 调用多模态预训练模型或多模态大语言模型，以提升模型性能。

2. 预期目标

1. **模型完成度：**
   (1) CNN/ViT + GRU 模型：100%。
   (2) 网格/区域表示 + Transformer 模型：80%-100%（视开发进度）。
2. **评测数据效果：**
   METEOR 和 ROUGE-L 指标初步完成。
3. **评测观感效果：**
   可视化效果直观，图像描述与图像内容基本吻合，生成句子衔接流畅。
4. **文档完善度：**
   包含设计逻辑、模型架构、评测结果与分析。

3. 相关工作

阶段一：数据集预处理

对图像数据进行统一大小调整和标准化等处理。

生成图像特征向量，作为后续输入。

整理文本描述，进行分词和词表生成。

阶段二：模型开发与优化

1. 实现 CNN/ViT + GRU 模型（工作量：★★★★）
   (1) 图像编码器：使用 CNN 提取图像特征或 ViT 生成图像的自注意力表示。
   (2) 文本解码器：使用 GRU 生成描述。
2. 实现网格/区域表示 + Transformer 编码解码框架（工作量：★★★★★）
   (1) 图像表示:使用 Faster R-CNN 或 DETR 提取区域特征
   (2) 编码/解码器：实现 transformer，基于区域表示生成描述。

(3) 训练优化:结合多头注意力机制，优化生成描述的语义一致性和语义理解能力。

3. 中期汇报材料准备（工作量： ★★）

阶段三：模型评估与改进

1. 模型评测与生成效果提升和性能优化记录，评测数据记录可视化展示。(工作量：★★)
2. 调用多模态大语言模型适配和调整，提升模型表现，对比评测结果。(工作量：★★★)

阶段四：文档撰写与展示优化

1. 整理项目流程与代码，撰写模型设计与结果分析报告。。 (工作量： ★★)
2. 撰写结题报告，拍摄演示视频/做口头报告展示。 (工作量： ★★)

4. 技术方案（初步预设）

模型结构

1. CNN/ViT+GRU：
   (1) 图像特征提取：使用预训练的 CNN 模型（如 ResNet、VGG）或 Vision Transformer 提取图像的整体表示。
   (2) 文本生成：使用门控循环单元（GRU）生成文本描述，每个时间步依赖前一时间步的隐藏状态和图像特征。
2. 网格/区域表示、Transformer 编码解码：
   (1) 图像特征提取：使用 CNN 提取图像的网格表示或目标检测模型（如 Faster R-CNN）提取图像的区域表示。
   (2) Transformer 编码器：使用多头自注意力机制处理图像特征，提取上下文信息。
   (3) Transformer 解码器：使用多头自注意力和交叉注意力机制生成文本描述，每个时间步依赖前一时间步的隐藏状态和图像特征。

评测标准

1. METEOR：
   实现词到词的映射和 chunk 计数，计算一元组的准确率和召回率，结合 chunk 计数计算 METEOR 值。
2. ROUGE-L：
   计算候选句子和参考句子的最长公共子序列（LCS），计算 LCS 的召回率和准确率，结合 β 值计算 ROUGE-L 值。

调用多模态预训练模型或多模态大语言模型：

选择 CLIP、BERT 等多模态预训练模型或多模态大语言模型。将预训练模型集成到现有模型中，提升图像和文本的联合表示能力。测试集成后的模型性能，评估其在图像描述生成任务中的表现。

5. 开发环境

编程语言：Python 3.8 及以上版本。

开发工具：Jupyter Notebook 用于代码编写和实验记录，VSCode 作为 IDE。

版本控制工具：Git，代码托管平台为 GitHub。

深度学习框架：PyTorch 1.8 及以上版本，用于构建和训练深度学习模型。

图像处理框架：Caffe，用于图像预处理和特征提取。

评估指标库：集成 BLEU、METEOR、ROUGE、CIDEr 和 SPICE 等评估指标库，用于模型性能评估。

CPU：Intel i7 或更高级别的处理器。

内存：至少 16GB RAM。

存储：SSD，至少 128GB。

GPU：NVIDIA RTX 系列或更高级别的 GPU，至少 12GB 显存，支持 CUDA ,用于加速深度学习模型的训练和推理。

开发环境：本地&云端 linux 开发环境，配置高性能 GPU。

测试环境：云服务或远程服务器，用于大规模模型训练和测试。

数据集：Flickr8k、Flickr30k、COCO、AIC-ICC 和 AI Challenger 等，用于模型训练和评估。

6. 组内分工

检查点	任务内容	组长	组员 1	组员 2
1	数据集预处理与方法调研	√	√	√
2	CNN/ViT + GRU 模型	√	√	
	CNN/ViT + GRU 模型			√

检查点	任务内容	组长	组员 1	组员 2
	网格/区域表示 ＋Transformer 编码解码框架		√	
	网格/区域表示 ＋Transformer 编码解码框架			√
	网格/区域表示 ＋Transformer 编码解码框架	√		
3	模型生成效果优化与展示	√	√	√
	模型评测	√	√	
	调用多模态大模型优化对比	√		√
4	文档撰写	√	√	√

7. 时间安排