实验课题3-基于CNN+LSTM的图像描述生成

介绍

- 本实验课题任务应用卷积神经网络（CNN）和长短期记忆神经网络（LSTM）执行图像描述（字幕）生成任务。 将使用Flickr8K数据集为模型训练数据，该任务涉及数据预处理、构建CNN+LSTM模型、可视化结果和分析模型性能。

实验目的

- 学习获得基于CNN+LSTM的图像描述生成的实践经验。
- 学习针对图像和文本数据的数据预处理技术。
- 了解CNN和LSTM架构和工作原理。
- 分析图像描述生成模型的性能并解释结果。

数据集

- Flickr 8k数据集是一个基于句子的图像描述和搜索的新基准，由8000张图像组成，每张图像配对五个不同的标题，提供对图片中实体和事件的清晰描述。这些图像是从原始Flickr30k数据集中选择的，包括各种不同的生活场景。该数据集包含一个Images目录和一个caption.txt文件，我们主要针对Images目录和caption.txt文件进行分析处理，前者放置着图片，后者是每个图片对应的描述语句。因为存在数据缺失，图像和描述语句的对应关系只生成了40455组，且描述语句最长达到了33个单词，最短为2个单词。
- Dataset Source: Flickr8k
- 部分数据样本可视化：

任务

Image Captioning （图像描述生成）

What is Image Captioning ?

- 图像描述是生成图像文本描述的过程，它使用自然语言处理和计算机视觉来生成图像描述。
- 这项任务是结合计算机视觉和自然语言处理技术。大多数图像描述生成系统使用编码器-解码器框架，其中输入图像被编码为图像特征表示，然后解码为描述性文本序列。

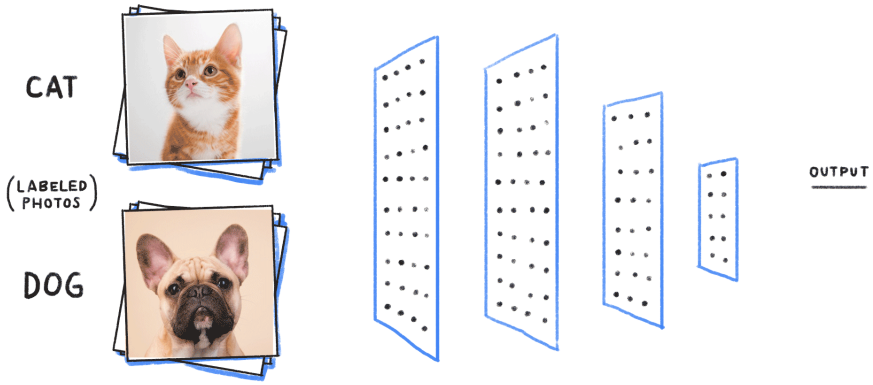
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

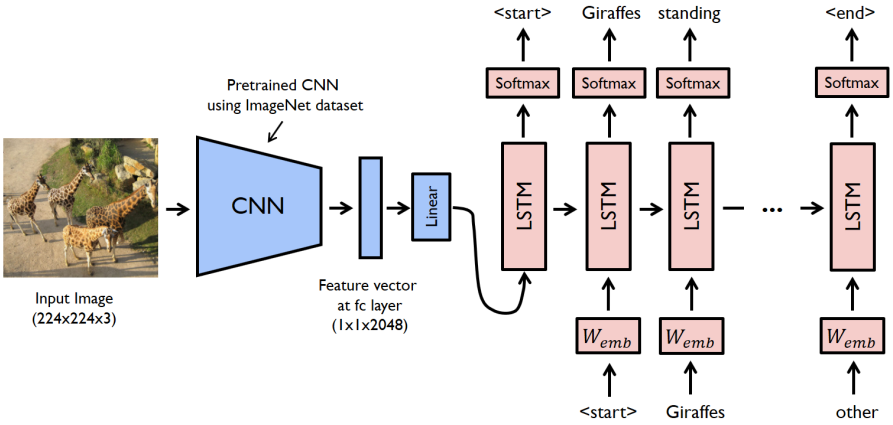
"two young girls are playing with lego toy."

模型：CNNs + RNNs (LSTMs)

- 图像描述生成模型一般由一个Encoder和一个Decoder两部分组成。
- Encoer: CNN从图像中提取图像特征，并表示为一个向量嵌入表示。这些图像特征的向量嵌入表示的维度取决于用于特征提取的预训练网络的类型（CNN类型,如ResNet）。



- Decoder: LSTM用于文本生成过程。图像特征嵌入与单词嵌入表示进行连接，并传递给LSTM模型以生成下一个单词。
- 模型结构:



实施步骤

1. **Data Preprocessing:**
   - 加载数据集并探索其结构.
   - 预处理图像：调整像素值的大小，图像分辨率调整，和归一化.
   - 预处理文本：标记描述，构建词汇表，并将描述转换为Embedding Representation.
   - 将数据集拆分为训练和验证集。
2. **图像描述生成模型（CNN+LSTM）：**
   - Build a CNN module and LSTM module using a framework like PyTorch.
   - Define the architecture of the CNN module, including convolutional layers, pooling layers, and fully connected layers.
   - Define the architecutre of the LSTM module, including LSTM layers, and fully connnected layers.
   - Combine the CNN and LSTM modules into one caption model.
3. **Training and Evaluation:**
   - 定义损失函数和优化器。
   - 使用训练数据集训练模型。监控训练过程，并输出损失函数指标。
   - *使用BLEU分数等指标在测试数据集上评估经过训练的模型*(已提供可选代码段，但需要调试。这一任务为可选加分任务)。
   - 评估训练数据与测试数据在训练过程中损失值的变化。
   - **请注意，在模型训练的时候，请根据自己计算机的计算资源，适当修改模型参数,例如bottleneck模块的线性层个数、神经元个数、以及LSTM层的神经元个数、以及classifier模块的线性层个数、神经元个数，以减少模型参数。**
4. **Visualization:**
   - 从测试集随机抽取一些示例图像，并输入到训练好的模型中生成描述。
   - 可视化图像及其生成描述和真实描述，以了解模型性能。
5. **Data Analysis:**
   - 分析数据集图像描述中单词频率的分布。
   - 探索示例图像及其生成描述，以讨论观察到的任何模式或挑战。
6. **Result Analysis:**

    - 解释从模型评估中获得的性能指标。
    - 讨论模型的优势和局限性。
    - 讨论潜在的改进方法和改进模型。

提交材料:

- 实验要求：
    - 本实验课题附件将提供源代码（Jupyter笔记本格式，包含详细的模型实现步骤和解释）和数据集。
    - 要求详细阅读代码并查阅相关文献，详细理解图像描述生成的实现原理，包括模型架构、数据预处理，文本处理、模型训练等。
    - 利用自己笔记本CPU和GPU、或者免费在线云计算平台（如谷歌的CoLab和阿里的天池）运行代码，并分析结果。
- 实验报告应包括：
    - 详细的图像描述生成原理说明：数据预处理步骤、模型架构、训练过程、评估结果、可视化结果和数据分析的报告。
    - 通过阅读代码，画出模型详细结构，包括网络层数、网络层类别、每层的神经元个数和激活函数、输出等。
    - 示例图像及其生成的描述可视化。讨论从分析中得出的见解和结论。
    - 如果可能，试图修改模型架构参数，并分析和观察评估结果。

Resources:

- CoLab, 包含免费的CPU和GPU资源。
- 阿里天池Notebook, 包含免费的CPU和GPU资源。
- PyTorch Documentation
- Matplotlib Documentation
- 离线模型下载地址：
    - ResNet18
    - ResNet34
    - ResNet50
    - VGG11
    - VGG16