# 基于编解码框架下的图像描述生成中期报告

## 一、题目

1. 实现如下模型结构：CNN/ViT整体表达+GRU,视觉Transformer+Transformer解码器
2. 实现如下测评标准：ROUGE-L,CIDEr-D
3. 完成任选任务中的：调多模态预训练模型或多模态大语言模型，并测试性能

## 二、详细设计

### （一）模型架构

1. **视觉 Transformer（ViT）+ Transformer 解码器**
   1. **编码器（ViT）**：将输入图像划分为固定大小的图像块，对每个图像块进行线性嵌入后，通过多层自注意力机制捕捉图像块间的全局依赖关系，获取图像的整体特征表示。ViT适用于具有较强细节和整体形状相互关系的服装图像，能够精确地捕捉到细节特征，为生成准确丰富的描述提供有力支持
   2. **解码器（Transformer）**：采用多头注意力机制与位置编码，以编码器输出的特征为上下文信息，自回归地生成文本描述。在生成每个单词时，依据图像特征与已生成的文本信息动态调整注意力分布，确保生成描述的逻辑性与连贯性。
2. **CNN 整体表示 + GRU**
   1. **编码器（CNN）**：通过卷积层、池化层的堆叠，提取服装图像的细粒度特征，如材质纹理、颜色分布及主要样式轮廓等。其卷积核设计可有效捕捉不同尺度的局部特征，经池化层压缩后形成图像的整体特征向量。
   2. **解码器（GRU）**：基于门控循环单元，处理编码器输出的特征向量，以逐字生成简洁的文本描述。GRU 的门控机制能有效筛选并保留关键信息，避免长序列处理中的梯度消失问题，在生成短文本服装描述时高效且计算成本低。

### （二）数据集处理

采用 DeepFashion-MultiModal 数据集，仅选取其中的 image 和 textual descriptions 部分。对图像数据进行归一化、裁剪、缩放等预处理操作，确保输入模型的图像规格统一、数据分布合理，提升模型训练的稳定性与收敛速度；对文本数据进行清洗、分词、构建词汇表及编码操作，将文本转化为模型可处理的数值形式，便于后续的描述生成与评估。

### （三）描述生成流程

1. **图像特征提取**：ViT 途径下，图像经分块嵌入与多层自注意力计算提取全局特征；CNN 途径则利用卷积与池化操作提取关键细节特征，二者并行处理输入图像，为后续不同风格的描述生成奠定基础。
2. **特征解码与描述生成**：ViT 提取特征交予 Transformer 解码器，经复杂的多头注意力交互与位置编码引导，生成语义丰富、逻辑连贯的长文本描述；CNN 特征由 GRU 解码器处理，凭借其高效的短文本生成能力，快速产出简洁直接的服装描述，以适配不同应用场景需求。

### （四）评价指标

1. **CIDEr**：基于 n-gram 语言模型，从词汇选择、语义匹配及描述结构多维度衡量生成描述与参考描述的相似性。对服装图像描述任务，其加权 n-gram 计算聚焦服装关键细节词语，精准评估描述质量，契合任务注重细节与简洁性要求。
2. **ROUGE-L**：以最长公共子序列为核心，考量生成文本与参考文本在短语层级的重叠度，有效捕捉服装描述的关键短语结构，补充 CIDEr 评价视角，全面评估生成描述对图像内容的准确表达与合理组织能力。

## 三、已完成工作

### （一）数据预处理模块构建

实现 DeepFashion-MultiModal 数据集图像数据的归一化、裁剪、缩放以及文本数据的清洗、分词、词汇表构建与编码程序，确保数据格式与模型输入要求精准匹配，奠定高效训练基础。经处理，图像数据分布归一至特定区间，文本数据转化为词索引序列，提升模型收敛效率。

### （二）模型架构初步搭建与验证

依据设计完成 ViT + Transformer 解码器及 CNN + GRU 模型架构编码，初步运行测试确保模型结构稳定，各组件间数据流通顺畅、维度匹配合理。运用小型样本子集进行前向传播测试，验证模型计算图无逻辑与维度错误，为大规模训练准备就绪。

### （三）基础训练与评估框架搭建

基于 PyTorch 与 Transformers 库，搭建模型训练循环与评估逻辑框架，实现多轮次训练迭代、参数更新及指标计算记录功能。于训练循环中集成 CIDEr 与 ROUGE-L 指标计算模块，实时监控模型性能变化趋势，为优化策略调整提供数据依据，经测试框架运行稳定、指标计算准确可靠。

### （四）部分模型训练实验开展

在简化数据集与有限计算资源下，启动 ViT + Transformer 解码器模型的初步训练探索，观测模型在服装图像特征提取与描述生成任务中的学习曲线、损失收敛态势及指标波动趋势，收集训练过程中的中间结果与模型参数，为后续深度优化储备经验数据。

## 四、初步结论

### （一）模型可行性验证

初步实验表明所选编解码框架下的两种模型架构在图像描述生成任务中具理论可行性，训练过程中模型能依据图像与文本数据逐步调整参数，尝试学习图像特征与自然语言间的映射规律，损失值呈波动下降趋势，显示模型对任务的适应潜力。

### （二）数据与模型适配性观察

经数据预处理流程优化与模型结构微调，数据集与模型适配性逐步提升，表现为训练初期模型在处理归一化后图像数据及编码文本时稳定性增强，未出现因数据异常导致的梯度爆炸或停滞现象，为进一步挖掘模型性能创造条件。

### （三）评价指标有效性确认

CIDEr 与 ROUGE-L 指标于服装图像描述任务中切实有效，能精准反映生成描述与参考描述的语义与结构相似性，为模型优化提供直观且具区分度的量化导向，依指标波动可精准定位模型性能瓶颈与提升方向。

## 五、问题及可能的解决方案

### （一）模型训练效率瓶颈

伴随训练数据量与模型复杂度增加，训练速度减缓、资源消耗剧增。解决方案为引入分布式训练技术，借助多 GPU 并行计算分摊训练负载；优化模型架构超参数，如调整层数、头数及隐藏单元数量，平衡性能与效率；采用混合精度训练，以半精度浮点数存储部分数据，削减内存占用与计算量同时维持精度。

### （二）过拟合现象显现

在有限数据训练阶段，模型于验证集表现波动下滑，出现过拟合倾向。应对策略为实施数据增强技术，如旋转、翻转、裁剪图像及添加噪声，扩充数据集规模与多样性；引入 L1、L2 正则化或 Dropout 机制，于训练中随机抑制神经元活性，削弱模型对特定样本特征的过度依赖，增强泛化能力。

### （三）生成描述语义准确性不足

生成的服装描述偶现关键元素缺失、语义模糊或逻辑矛盾状况。拟采用强化学习算法优化模型训练目标，直接以评价指标为奖励信号引导模型生成策略优化；增强注意力机制效能，引入自适应注意力策略，依图像内容动态分配注意力权重，聚焦关键区域提升语义准确性；扩充与细化词汇表，纳入更多服装专业术语及描述性词汇，提升模型表达精度。

## 六、后续工作计划

### （一）模型深度优化与拓展

持续调整模型架构超参数，依实验反馈精准优化；探索融合视觉 - 语言预训练模型技术，迁移学习其优质特征表示与跨模态对齐能力，提升模型性能上限；构建模型融合框架，整合 ViT + Transformer 与 CNN + GRU 优势，依图像特征复杂程度自适应切换或协同生成描述，增强模型适应性与准确性。

### （二）大规模训练与性能提升

拓展至完整 DeepFashion-MultiModal 数据集，运用分布式训练系统加速训练进程；精细调整优化器参数、学习率调度策略及批次大小，依据训练曲线精准调控模型收敛节奏；引入课程学习策略，依图像复杂度或文本长度渐进式提升训练难度，助力模型稳健掌握复杂映射关系，提升整体性能。

### （三）系统集成与应用测试

完善模型训练、评估与部署全流程自动化脚本，实现一键式操作；开发便捷用户交互界面，支持图像上传、描述生成展示及用户反馈收集；于实际应用场景展开测试，收集用户反馈数据优化模型与系统，确保系统易用性、稳定性及描述生成质量满足多样化需求。