智能感知认知实践实践项目：图片摘要生成

涂宇清 522030910152

1 引言

近年来，计算机视觉经历了从图像分类等简单任务到图像生成和理解等复杂任务的发展。同时，自然语言处理领域取得的重大进展使得结合视觉与语言来解决具有挑战性的跨模态任务成为可能。图像摘要生成是该领域的一项代表性任务，旨在为给定图像生成自然语言摘要。它不仅要求模型理解视觉和文本信息，还需要建立两者之间的关系。图像摘要生成在某种程度上弥合了计算机视觉与人类感知之间的鸿沟，并可应用于各种现实场景，如图像分类、图像检索和内容分析。

编码器-解码器（Encoder-Decoder）是解决图像摘要生成任务的主要方法之一，其使用卷积神经网络（CNN）编码图像，并使用循环神经网络（RNN）将特征解码成句子。同时，由于注意力机制的引入，使得模型在生成每个单词时聚焦于图像的不同部分，显著提升了模型性能。

在本项目中，我们将探索使用带有视觉注意力的编码器-解码器结构来解决图像摘要生成任务。同时，探究视觉大语言模型（VLM）在图像摘要生成中的应用。

2 方法

2.1 CNN-LSTM with attention 结构

图像摘要生成任务的主要挑战在于如何有效地将图像信息与文本信息结合起来。为此，我们采用了 CNN-LSTM with attention 结构，如图 2.1 所示。该结构由两个主要部分组成：图像编码器和文本解码器。

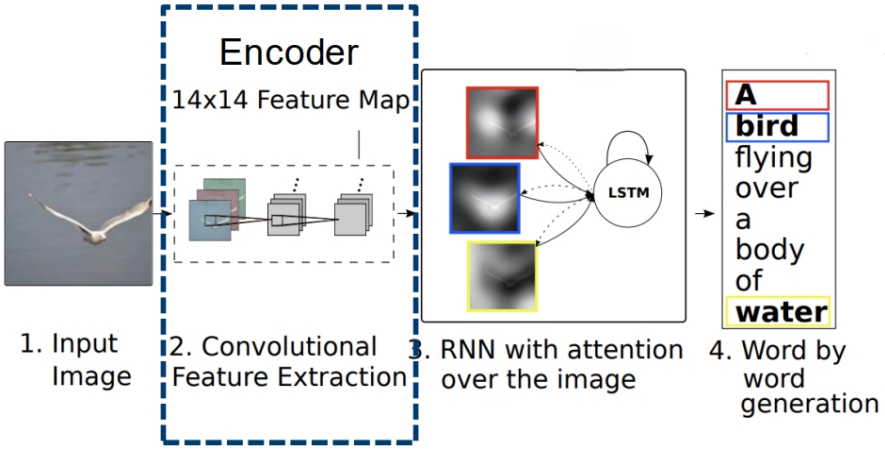

图 2.1　CNN-LSTM with attention 结构

2.1.1 图像编码器

图像编码器负责提取给定图像的特征，该特征包含生成摘要所需的视觉信息。我们采用在 ImageNet 数据集上预训练的 ResNet-101 模型作为图像编码器，以确保足够的表示能力。它将输入图像编码为一个具有 2048 个通道的 14×14 特征图 g。

2.1.2 文本解码器

文本解码器负责基于图像编码器提取的特征生成摘要文本。我们采用一个融合了视觉注意力机制的长短期记忆网络（LSTM）作为文本解码器。除了视觉信息，预测还应考虑文本信息，即句子中先前生成的单词。给定上一时刻的隐藏状态 h_{t-1}，计算权重 β_t：

$$\beta_t = \frac{e^{\psi(h_{t-1})}}{1 + e^{\psi(h_{t-1})}}$$

其中 ψ 是一个全连接层。然后，计算上下文向量 c_t：

$$c_t = \beta_t \cdot z_t$$

最终，上下文向量 c_t 将与预测的词嵌入 y_{t-1} 拼接起来，并输入 LSTM 单元，以生成下一个隐藏状态 h_t。h_t 作为预测头的输入，以生成下一个单词的概率分布。

2.2 计划采样

在传统设置中，模型被训练为在给定上一时间步的真实标签的条件下，最大化每个 token 的可能性。然而，在推理过程中，上一时间步的真实标签被替换为

最后预测的 token，这可能导致错误累积。

为了解决这个问题，计划采样技术通过在每个时间步随机选择是使用真实标签还是使用模型预测的 token 来修改训练策略：在第 i 个 epoch，我们以概率 ϵ_i 使用真实标签，以概率 $1-\epsilon_i$ 使用预测的 token。

并且，我们倾向于在训练过程中动态衰减采样概率 ϵ_i。这样，模型在训练过程中可以更快地收敛，并在推理过程中具有更好的泛化能力。

2.2.1 线性衰减

$$\epsilon_i = \max(\epsilon, c - ki)$$

其中 $0 \leq \epsilon < 1$ 决定了采样概率的下限，k 和 c 分别为衰减的斜率和截距，它们取决于预期的收敛速度。

2.2.2 指数衰减

$$\epsilon_i = k^i$$

其中 $k < 1$ 决定了采样概率的衰减率。

2.2.3 逆 sigmoid 衰减

$$\epsilon_i = \frac{k}{k + e^{i/k}}$$

其中 $k \geq 1$ 决定了采样概率的衰减率。

2.3 束搜索

束搜索是一种启发式搜索算法，广泛应用于序列生成任务中。在解码器生成描述文本时，我们需要搜索所有可能的单词序列以找到全局最优解。然而，搜索空间会随着序列长度呈指数级增长，这使得穷举法在计算上不可行，而贪心算法总是在每个时间步选择概率最高的单词，这往往会导致生成次优结果。

引入束搜索技术旨在平衡计算复杂度和生成描述质量之间的权衡。在每个

时间步，束搜索维护一组数量为 k 的部分假设序列。这些假设序列通过从词汇表中附加每个可能的单词进行扩展。然后，根据累积对数似然对所有假设序列进行排序，仅保留排名前 k 的假设序列用于下一个时间步的扩展。

2.4 评估指标

为了全面评估模型的性能，我们引入了多种评估指标，包括 Bleu、Rouge、METEOR、CIDEr、SPICE 和 SPIDEr。不同的指标在评估生成描述的质量时侧重于不同的方面。

2.4.1 Bleu

Bleu（双语评估替换得分）是机器翻译任务中常用的指标，它衡量预测结果与参考描述之间的 n 元词组重叠程度。其核心思想是衡量预测结果与参考描述的接近程度。Bleu 考虑 n 元词组而非单个单词，但所有 n 元词组被赋予同等权重，这可能会带来一定的偏差。

2.4.2 Rouge

Rouge（面向召回率的摘要评估替换得分）是一组用于评估文本摘要算法的指标，它衡量预测结果与参考描述之间最长公共子序列的长度。Rouge 与 Bleu 类似，但它更侧重于召回率而非精确率。

2.4.3 METEOR

METEOR （显式排序的翻译评估指标）同样是一个机器翻译任务指标。它会对齐预测结果和参考描述，并计算不同情况下的准确率、召回率和 F1 分数。需要注意的是，如果预测结果的词序与参考描述不同，METEOR 会进行惩罚，因此它与人类判断较为一致。

2.4.4 CIDEr

CIDEr（基于共识的图像描述评估）是专为图像标注任务设计的指标。它执行词频-逆文档频率 (TF-IDF) 分析，将句子表示为向量。然后，CIDEr 被计算为预测结果与参考描述向量之间的余弦相似度。这种方法弥补了 Bleu 的局限性。

2.4.5 SPICE

SPICE（语义命题图像描述评估）是专为图像描述任务设计的指标。它将句子解析为一个场景图，并计算提取出的元组集的精确率、召回率和 F1 分数。与 CIDEr 相比，SPICE 更侧重于句子的语义，因此与人类判断更为一致。

2.4.6 SPIDEr

SPIDEr 取 SPICE 和 CIDEr 分数的平均值。它提供了一种综合评估，平衡了生成描述中的词频信息和语义信息。

3 实验过程

3.1 模型结构

首先，我们将讨论不同的 embedding 维度和 decoder 维度对模型性能的影响。我们将 embedding 维度和 decoder 维度分别都设置为 128、256、和 512（embedding 维度-decoder 维度），其他参数保持默认。实验结果如表 1 所示。

Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Rouge	METEOR	CIDEr	SPICE	SPIDEr
128-128	0.592	0.403	0.267	0.174	0.406	0.191	0.469	0.137	0.303
128-256	0.590	0.404	0.270	0.175	0.406	0.190	0.465	0.136	0.301
256-128	0.583	0.393	0.259	0.168	0.399	0.189	0.450	0.134	0.292
256-256	0.596	**0.407**	0.269	0.173	0.408	0.191	0.479	**0.139**	**0.309**
256-512	**0.597**	0.406	**0.271**	**0.180**	0.409	0.194	**0.481**	0.138	**0.309**
512-256	0.592	0.403	0.266	0.171	**0.411**	0.193	0.471	0.138	0.304
512-512	0.591	0.405	**0.271**	0.179	0.404	**0.195**	0.471	**0.139**	0.305

表 1　不同 embedding 维度和 decoder 维度的实验结果

由上述结果可以得出，综合看来，256-512 的配置在大多数指标上表现最佳。接下来，我们将使用 256-512 的配置进行后续实验。

3.2 消融实验

接下来，我们将探究计划采用中不同动态衰减方法以及不同搜索策略对模型性能的影响。我们将使用 256-512 的配置，并分别使用线性衰减、指数衰减和逆 sigmoid 衰减作为计划采样的动态衰减方法，同时使用束搜索和贪心搜索作为

搜索策略。实验结果如表 2 所示。

Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Rouge	METEOR	CIDEr	SPICE	SPIDEr
Linear + Greedy	0.595	0.400	0.262	0.171	0.408	0.191	0.459	0.137	0.298
Linear + Beam	0.614	0.417	0.272	0.173	0.407	0.184	0.460	0.138	0.299
Exponential + Greedy	0.601	0.407	0.271	0.177	0.411	**0.193**	0.482	0.140	0.311
Exponential + Beam	0.621	**0.437**	**0.297**	**0.197**	**0.418**	0.192	**0.506**	**0.145**	**0.326**
Inverse Sigmoid + Greedy	**0.640**	0.422	0.273	0.171	0.415	0.180	0.451	0.128	0.290
Inverse Sigmoid + Beam	0.600	0.417	0.279	0.180	0.403	0.181	0.465	0.134	0.300

表 2　不同动态衰减方法和搜索策略的实验结果

由上述结果可以得出，指数衰减 + 束搜索组合展现出最优的综合性能，这表明指数衰减策略有效协调了训练稳定性与推理泛化性，而束搜索则通过多路径候选序列优化，显著提升了描述语句的连贯性与信息完整性。

3.3 视觉大语言模型

视觉大语言模型（VLMs）是近年来在图像理解和生成领域取得突破的关键技术。它们通过将视觉信息与语言模型相结合，能够更好地理解图像内容并生成相关的自然语言描述。在本次实验中，我们使用 Doubao-1.5-vision-pro 模型作为视觉大语言模型。实验结果如表 3 所示。

Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Rouge	METEOR	CIDEr	SPICE	SPIDEr
256-512 Exponential + Beam	0.621	0.437	0.297	0.197	0.418	0.192	0.506	0.145	0.326
Doubao-1.5-vision-pro	**0.728**	**0.559**	**0.416**	**0.305**	**0.569**	**0.296**	**0.981**	**0.249**	**0.615**

表 3　视觉大语言模型实验结果

可以看出，Doubao-1.5-vision-pro 模型在所有指标上均远远优于我们之前的模型。这表明视觉大语言模型在图像摘要生成任务中具有更强的能力，能够更好地理解图像内容并生成相关的自然语言描述。

4 结论

本次项目聚焦于图片摘要生成任务，旨在探索相关模型及方法。实验表明，在不同 embedding 维度和 decoder 维度的模型配置中，256 - 512 的配置综合表现最佳。在搜索策略和计划采样动态衰减方法的研究中，指数衰减与束搜索的组合展现出最优性能，有效提升了描述语句的连贯性和信息完整性。而引入视觉大语

言模型 Doubao-1.5-vision-pro 后，其在所有指标上都远超之前模型，凸显了视觉大语言模型在图像摘要生成领域的强大潜力。

5 代码及复现方式

详情见GitHub。