

基于图像分类规划学习的视觉故事生成模型

王元龙, 张宁倩, 张虎

山西大学计算机与信息技术学院, 山西 太原 030006

摘要

视觉故事生成旨在通过解析和整合一系列给定的图像, 创作出逻辑连贯、情节流畅的故事。现有的解决方案大多直接处理图像信息, 虽引入了主题分析、知识图谱等方法, 但在图像信息的处理上存在视角单一、弱化故事生成过程、缺乏结构化设计等问题。针对上述问题, 提出了一种基于图像分类规划学习的视觉故事生成模型 (IC-PLVS 模型), 引入图像分类和规划学习方法, 将图像划分为人、动物、食物、自然景观、建筑、室内场景、其他 7 个类型, 并针对每个类型设定对应的问题, 利用视觉问答预训练语言模型生成设定问题的答案, 完成规划设计, 从而引导视觉故事生成。模型分为 4 个阶段: 第一阶段从图片中提取视觉信息; 第二阶段利用预训练语言模型进行分类, 并引导规划信息生成; 第三阶段更新数据集词汇信息; 第四阶段融合以上阶段生成的视觉和规划信息, 完成视觉故事生成任务。在公开数据集 VIST 上验证所提模型的效果, 与现有模型 COVS 相比, 其在 BLEU-1、BLEU-2、CIDEr、Distinct-3、Distinct-4 和 TTR 指标上分别提升了 2.07%、4.29%、0.44%、1.78%、0.91% 和 1.07%。

关键词

视觉故事生成; 图像分类; 规划学习; 视觉问答

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025076

Visual storytelling based on image classification planning learning

WANG Yuanlong, ZHANG Ningqian, ZHANG Hu

School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China

Abstract

Most of the existing solutions directly process image information, although topic analysis, knowledge graph, and other methods are introduced, there are problems in the processing of image information, such as a single perspective, weakening the story generation process, and a lack of structured design. To solve the above problems, this paper proposes a visual story generation model based on image classification planning learning, introduces image classification and planning learning methods, divides images into seven types: people, animals, food, natural landscapes, architecture, indoor scenes, and others, and sets corresponding questions for each type, and uses the Visual Question pre-trained language model to generate answers and complete planning design, to guide visual story generation. The model is divided into four stages: the first stage extracts visual information from pictures; The second stage is to use the pre-trained language model to classify and guide the generation of planning information; the third

stage updates the vocabulary information of the dataset; the fourth stage integrates the visual and planning information generated in the previous stages to complete the visual storytelling task. Compared with the existing COVS, the results of the proposed model are improved by 2.07%, 4.29%, 0.44%, 1.78%, 0.91% and 1.07% on BLEU-1, BLEU-2, CIDEr, Distinct-3, Distinct-4, and TTR.

key words

storytelling, image classification, planning learning, visual question

0 引言

视觉故事生成作为一项综合性的任务，融合了计算机视觉与自然语言处理技术两大领域，旨在将序列图像转化为生动、连贯且富含情感和逻辑的故事叙述^[1-2]。相比图像描述任务，此过程中不仅需要识别并解析单张图像的内容，还需要捕捉和理解多张图像之间的元素以及元素之间的关系，并将获取的视觉信息转化为流畅、富含逻辑的故事^[3-4]。

传统的故事生成方法大多采用编码器-解码器框架，首先利用计算机视觉技术提取图像特征表示，然后将其输入语言生成模型以解码生成故事。为了降低句子的冗余度，Hus等^[5]在解码阶段应用了句间多样化束搜索技术，该技术会参考先前生成的句子，避免在给定相似图片序列时生成重复的句子。Hu等^[2]提出利用强化学习框架及复合奖励函数，优化了故事质量。以上方法虽然在一定程度上实现了从图像到故事的初步转换，但用词单一导致故事丰富度不足，难以全面捕捉图像背后的含义。为了缓解此类问题，研究者们引入多阶段故事生成方法，此方法大多包括视觉特征提取、概念信息引入、故事生成3个阶段，在概念信息引入阶段根据图像中的目标引入相关概念，其不仅关注图像中直接可见

的对象和场景，还提供知识库^[6]、知识图谱^[3]等概念。Gu等^[7]引入主题模型，提出了并行自上而下视觉和主题注意力的连贯视觉叙事模型（COVS），使故事围绕主题展开，提高故事的一致性。多阶段故事生成方法也面临一些挑战，如加入词汇缺乏针对性、未能紧扣主题或情景，导致获取的信息与故事核心或后续的句子脱节。根据对COVS等模型视觉故事生成结果的分析可知，目前视觉故事生成任务存在描述内容与图片不相符、冗余信息较多以及缺乏细节描述等问题。

基于上述问题，受Cheng等^[8]在评估视觉-语言模型任务中提出的6个评估维度（对象、动作、地点、推理、预测和计划）启发，并借鉴司亚中等^[9]提出的图像视觉增强思路，本文依据图片主题类别不同而其维度也有一定差异的思路，采用问答的方式从不同维度规划所生成的故事，提出了基于图像分类规划学习的视觉故事生成（visual storytelling based on image classification planning learning, IC-PLVS）模型。IC-PLVS模型先将图片划分为7个类别（人、动物、食物、自然景观、建筑、室内场景、其他），针对图像类型共设定了12个维度的问题，其中包括3个通用问题（主题、对象、地点）和9个不同图片类别对应的针对性问题，以此引导预训练视觉问答模型基于图像回答问题，规划故事核心要素。本文主要贡献如下。

- 引入图像分类方法：首次将数据集划分为7个类型，通过此方法，简化图像理解的复杂度，提升了视觉内容与文本描述之间的关联度，从而避免内容与图像脱节的问题。

- 设定全面的规划机制：通过为不同类型图像规划不同的方案，确定出故事的核心要素，后续通过组织相应的行为和环境描述，融入细节描述，丰富故事内容。

- 在公开数据集 VIST (visual storytelling)^[10]上的实验结果表明，本文方法效果在 BLEU-1、BLEU-2、Distinct-3、Distinct-4 和 TTR 中优于基线模型。

1 相关工作

1.1 视觉故事生成

视觉故事生成侧重于将序列图像转化为内容丰富、逻辑连贯的故事。与针对单幅图像进行即时解读和概括的图像字幕任务不同，视觉故事生成需要探索每一幅图像的内涵及它们之间的联系，从而理解图像流的整体概念。Huang 等^[10]构建了视觉故事生成的数据集 VIST，使用序列到序列的循环神经网络故事生成模型，为视觉故事生成领域奠定基础。鉴于自动评估标准在衡量故事质量上的不足，Wang 等^[1]提出了对抗性奖励学习框架，通过从训练数据集中学习隐式奖励函数来优化策略搜索，提高模型的泛化能力。为了提升故事的连贯性和多样性，Chen 等^[11]引入知识图谱，并建立图片序列内部的跨图像关系。为提高图像的利用率，Malakan 等^[12]结合 ViT 和 Bi-LSTM 构建视觉故事生成模型，将每个图像划分为 16 块，从多个视觉特征中提取信息，提高故事的连贯性和相关性。Chen 等^[13]为了丰富故事的情感内涵，引入

了情绪概念生成器，首先处理图像特征，然后提取情绪，最后生成文本，通过将概念词融入句子，丰富故事内容。以上方法从不同角度解决了故事连贯性、丰富性等问题。

近年来，随着多模态大模型 (large language model, LLM) 的发展，基于其生成故事的研究逐渐增多，利用大模型生成故事成为一个新方向。Yang 等^[14]利用 CLIP 大模型进行视觉与文本对齐，通过多任务训练和记忆模块生成风格化故事，但是图像与风格化文本不完全匹配，导致生成故事的语义相关性下降。Song 等^[15]利用 CLIP 和 GPT-2，通过引入上下文感知机制和对比学习目标，提升了故事生成的质量和多样性。Lin 等^[16]提出结合 LLM 和 LVLM (large vision-language model) 的新方法，通过指令调整来解决视觉叙事中维护叙事连贯性和上下文相关性的问题。而视觉故事生成的研究不应仅局限于上述方向。视觉问答 (VQA) 模型通过理解和回答与视觉内容相关的问题，能够更好地捕捉图像中的细节和语义信息。将 VQA 模型与视觉故事生成相结合，为故事生成提供了更丰富的背景知识和逻辑推理能力，从而提升故事的质量和连贯性。

1.2 规划学习

在故事生成任务中，通过设定故事情节确定图像关键对象是一项重要的内容。Hsu 等^[3]首次引入了规划机制，融合图像和知识图谱构建故事图，为故事情节设定规划。但该方法采用的网络架构较复杂，为了降低复杂性，Wang 等^[17]提出 SCO-VIST 框架，将图像序列表示为包含对象和关系的图，然后通过语义和发生概率的边权重构建情节节点的桥梁，并利用

Floyd-Warshall 算法生成事件序列的故事线。Liu 等^[4]另辟蹊径,引入了问答对作为故事生成的蓝图。问答对的使用不仅降低了计算的复杂度,还提高了故事生成的精准度。

基于上述问题,本文提出基于图像分类规划学习的视觉故事生成(IC-PLVS)模型(如图1所示)。IC-PLVS模型先将图片划分为7个类别(人、动物、食物、自然景观、建筑、室内场景、其他),并针对每一个类别设定了12个相应的问题,之后引导预训练问答语言模型剖析图像,基于图像本身回答问题,从而确定故事核心要素。

应图像特征提取、规划信息生成、更新数据集词汇、故事生成4个模块。

- 阶段1: 图像特征提取阶段。CNN作为图像序列编码器,提取每张图像的特征。

- 阶段2: 规划信息生成阶段。利用视觉问答模型识别图像类别,并根据设定的问题获取答案,拼接问题与答案作为解码器的输入信息。

- 阶段3: 更新数据集词汇阶段。将生成的新词汇添加到词典中。

- 阶段4: 故事生成阶段。分别结合视觉和规划信息生成故事。

2.1 图像特征提取模块

本模型采用卷积神经网络(CNN)作为图像编码器。为了有效整合图像特征引入平均池化机制,通过计算每个特征维度

2 IC-PLVS模型

IC-PLVS模型分为4个阶段,分别对

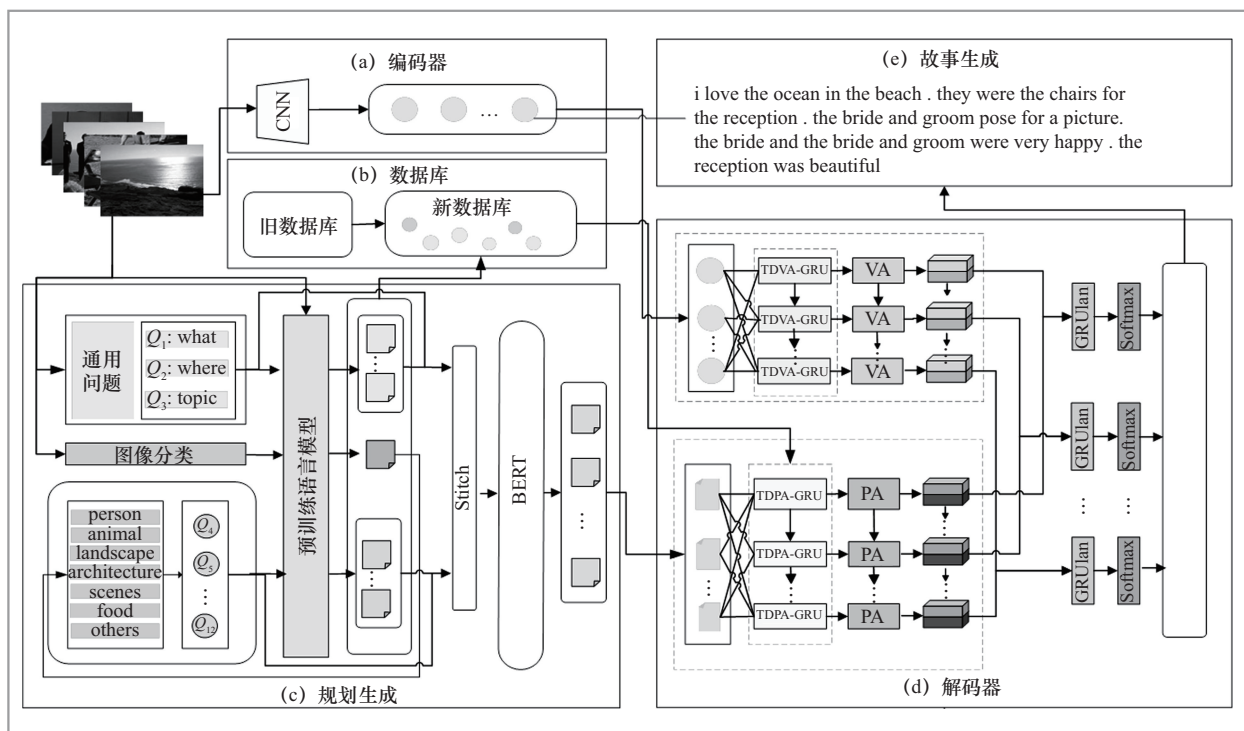


图1 基于图像分类规划学习视觉故事生成模型架构

上所有图像嵌入的平均值实现全局特征的累积与融合。上述过程获取的视觉信息为 $V(V_1, V_2, V_3, V_4, V_5)$ ，其中数字下标表示图片序号。

2.2 规划信息生成模块

规划学习模块采用预设问题驱动的信息提取方法，将预先设定的问题与图像输入 Monkey^[18] 中获取规划信息。本文受 Cheng 等^[8] 从6个维度出发评估方法的启发，结合当前任务的特殊性，构建了基于通用和特定两个方向、细化为12个子维度的方法提取信息。

2.2.1 通用规划信息提取

通用规划信息提取旨在利用 Monkey^[18] 的泛化能力，从图像中识别并提取出图像的基本要素：主题、地点、主要对象（见表1）。

首先将序列图像 I^i 和设定的通用问题 Q_j^g 输入 Monkey 中，得到答案 A_j^i ，如式（1）所示：

$$A_j^i = \text{Monkey}(I^i, Q_j^g) \quad (1)$$

其中， I^i 表示第 i 个图像， Q_j^g 表示图像 i 的第 j 维度通用问题， Q_1^g, Q_2^g, Q_3^g 组成了能用问题集合 Q^g ， A_j^i 表示第 i 个图像第 j 维问答的答案。

表1 通用问题

维度	问题	目的
theme	What is the theme	主题
object	What is in the picture	对象
location	Where the events took place	地点

2.2.2 图像分类

本文在参考 Deng 等^[19] 划分图像类型方法以及曹峰等^[20] 图像分类算法思路的基础上，将图像分为7个类别：人、动物、自然景观、食物、建筑、室内场景、其他。在图像分类时，设定以下优先级顺序：人>动物>自然景观>食物>（建筑=室内场景）>其他。具体规则为，若图像中同时包含人物与其他类别（如动物或自然景观），则该图像被归类为人类别；此逻辑同样适用于其他类别间的比较，依此类推。在后续的特定问题中会围绕图像类别针对性地分析周围的核心元素，旨在更好地捕捉图像中的核心元素，提升分类的准确性和实用性。图像分类具体过程为：首先将序列图像 I^i 和问句 Q^w ：“What category do images belong to: people, animals, landscapes, architecture, indoor, food, others” 输入 Monkey 中，得到图像类别 C^i ，如式（2）所示：

$$C^i = \text{Monkey}(I^i, Q^w) \quad (2)$$

其中， Q^w 表示图像类别问题， C^i 表示第 i 个图像的类别。

2.2.3 特定规划信息提取

Cheng 等^[8] 提出的评估法设定 What is around me、What am I doing、Where am I、What about the situation around me、What will happen to me、How will I do 这6个问句判断模型是否具有相应的能力。本文在此基础上，从每个类型的特点出发，设定9个具体的问题，为提取规划所需的具体信息。各类信息预先设定的特定问题 Q^s 见表2。根据图像类型，找到对应规划问题，如式（3）所示：

$$Q^{si} = Q^s(C^i) \quad (3)$$

表2 特定问题

类型	问题	目的
人	Which words describe the people	对象
	How many characters are in the picture	个数
	What mood is the characters displaying	情绪
	What is the person doing in the picture	地点
	Which physical traits does the person have	外貌
	What is the interaction or relationship between the characters	推理
	What are the distinctive features of the characters' clothing styles	风格
	What activities people are attending or holding	动作
	What are people going to do	预测
动物	How many animals are in the picture	个数
	What word describes the animal	对象
	What animal is in the picture	对象
	What are the animal's habits	习性
	What kind of environment do the animals live in	环境
	What are the animals doing	动作
	What mood is the animal displaying	情绪
	How does the animal interact with the environment	联系
	What is animal going to do	预测
风景	What type of natural landscape is it	类型
	What specific landform features are shown	地形
	What is the atmosphere of the landscape	环境
	What are the climatic characteristics	天气
	What kind of atmosphere does landscape convey	氛围
	What season is it in the picture	季节
	What colors are predominantly found in this natural landscape	颜色
	What types of plants can be found in this natural landscape	对象
	What are the climatic characteristics of the landscape	特征
室内	What kind of interior scene is it	类型
	What is the background of the scene	背景
	Which interior decoration style does the picture belong to	风格
	What event or activity is happening	活动
	How do you describe the spatial structure of a scene	结构
	How to describe the emotional impact of the scene on the characters or readers	情绪
	What is the purpose of this interior room	用途
	What colors and patterns are used	颜色
	What kind of atmosphere the scene creates	氛围
食物	What is the food	对象
	What is the type of food	类型
	How would the taste of the food	口味
	Where the food is arranged	环境
	Which festival is the food related to	节日
	Which does the food feel in the mouth	口感
	How to describe the presentation and decoration of food	外观
	What is the cooking method of the food	烹饪
	Which event is the food related to	联系

续表

类型	问题	目的
建筑	What are the distinctive decorative elements of the building's exterior	装饰
	What is the architectural style of the building	类型
	What is the descriptive features of the building	特征
	What is around the building	环境
	What is the atmosphere and feeling that the building creates	氛围
	What materials were used to construct this building	素材
	What is the impact of the building on the surrounding environment	影响
	What is the exterior of the building	外观
	What type of building	风格
其他	What atmosphere is created	氛围
	What color is it	颜色
	What is the use of the object	作用
	What event was held in the picture	活动
	What technical methods or tools were used to create the image	材料
	What emotions does the image convey	情感
	What are the creative techniques or features of the image	特征
	What else is in the picture	对象
	What cultural elements are depicted	文化

其中， $Q^{si}(Q_4^{si}, Q_5^{si}, \dots, Q_{12}^{si})$ 表示第*i*个图像的特定规划问题集合，特定问题一共有9个，从第4个到第12个，其中 Q_4^{si} 表示第*i*个图像的第4个问题。

将问题集合 Q^{si} 与图像 I^i 输入Monkey模型获取答案。针对每个类别，预设一系列特定问题，并将问题与图像一同输入Monkey模型，获取详细答案，如式（4）所示：

$$A_j^i = \text{Monkey}(I^i, Q_j^{si}) \tag{4}$$

2.2.4 规划信息整合

将规划信息输入解码器前需进一步处理：首先将两类规划的问答对分别拼接，组成question: Q_j^{si} [SEP]answer: A_j^i 、question: Q_j^{si} [SEP]answer: A_j^i 的形式，后通过分隔符[SEP]将第*i*个图像的每维规划拼接，并在首部添加[CLS]标记，组成第*i*张图片的规划信息字符串，如图2所示，形成规划信息： $P(P_1, P_2, \dots, P_{12})$ 。整合后将规划信息输入BERT进行语言特征提取，最终转化为词嵌入作为输入数据。

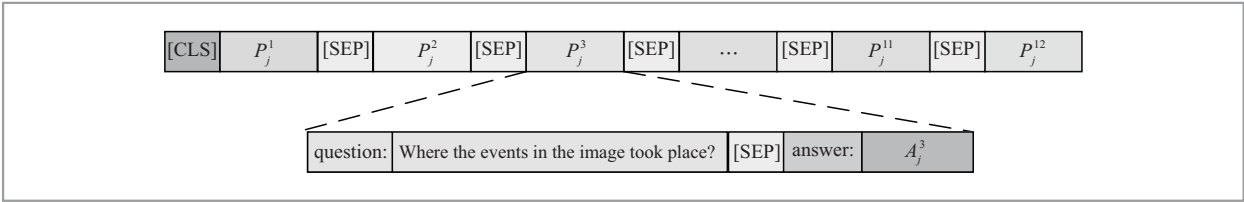


图2 规划拼接

2.3 更新数据集词汇模块

鉴于解码器通过词嵌入技术进行文本解码,为了更有效地引入新词汇,将未包含在原词汇表的词汇按照(序号:单词)的格式加入扩展后的词典中,构成新的词典,进而增强故事的丰富性和多样性。

2.4 故事生成模块

故事生成模块由信息解码器 (GRU) 和连贯语言生成器组成。其中信息解码器由视觉解码器 (TDVA-GRU) 和规划解码器 (TDPA-GRU) 分别解码视觉和规划信息。首先利用任务特定的上下文预测注意分布;然后用加权平均特征计算注意特征向量;分别向 TDVA-GRU、TDPA-GRU 中输入步长 $t-1$ 的隐藏状态 $h_{j,t-1}^{\text{lan}}$ 、对应的视觉信息 $v_{j,t}$ 或规划信息 $p_{j,t}$ 、上一步长的单词信息 $W_{e_{\chi_{t-1}}}$, 获得此步长的隐藏状态,如式 (5)、式 (6) 所示:

$$h_{j,t}^{\text{VA}} = \text{TDVA-GRU}(h_{j,t-1}^{\text{lan}}, v_{j,t}, W_{e_{\chi_{t-1}}}) \quad (5)$$

$$h_{j,t}^{\text{PA}} = \text{TDPA-GRU}(h_{j,t-1}^{\text{lan}}, p_{j,t}, W_{e_{\chi_{t-1}}}) \quad (6)$$

其中, $h_{j,t-1}^{\text{lan}}$ 表示第 j 个图像序列在时间步长 $t-1$ 时 GRU_{lan} 的隐藏状态; $v_{j,t}$ 表示在时间步长 t 下 CNN 对第 j 个图像序列的输出。 $W_{e_{\chi_{t-1}}}$ 表示上一个单词的编码, $W_e \in R^{E \times \Omega}$, $R^{E \times \Omega}$ 是词汇表, E 为词的嵌入大小, χ_{t-1} 为输入词在时间步长 $t-1$ 时的单次编码。

在时间步长 t 处,图像归一化注意力权重 ($v_{j,t}^*$)、规划归一化注意力权重 ($p_{j,t}^*$) 分别如下:

$$v_{j,t}^* = \tanh(W_v v_{j,t} \oplus W_a h_{j,t-1}^{\text{VA}}) \quad (7)$$

$$p_{j,t}^* = \tanh(W_p p_{j,t} \oplus W_b h_{j,t-1}^{\text{PA}}) \quad (8)$$

其中, W_v 和 W_a 为变换矩阵, $\tanh(\cdot)$ 为双曲正切函数。 $h_{j,t-1}^{\text{TA}}$ 、 $h_{j,t-1}^{\text{PA}}$ 分别为 TDVA-

GRU、TDPA-GRU 对第 j 个图像序列在时间步长 $t-1$ 处的输出。 W_v 、 W_a 和 W_p 是变换矩阵。

因此,输入 GRU_{lan} 的第 j 个图像序列的视觉、规划注意力分布分别如下:

$$\hat{v}_{j,t} = \sum_{j=1}^J \text{softmax}(W_{hv}^T v_{j,t}^*) v_{j,t} \quad (9)$$

$$\hat{p}_{j,t} = \sum_{j=1}^J \text{softmax}(W_{hp}^T p_{j,t}^*) p_{j,t} \quad (10)$$

其中, W_{hv}^T 、 W_{hp}^T 为权重参数, $v_{j,t}$ 、 $p_{j,t}$ 分别为第 j 个图像序列在时间步长 t 的视觉、规划的注意力分布。

连贯语言生成器 GRU_{lan} 负责在视觉、规划关注特征引导下生成故事。之后用 GRU_{lan} 生成词序列,并在每个时间步更新 GRU_{lan} 的隐藏状态。GRU_{lan} 的隐藏状态在每一时间步按 $h_{j,t-1}^{\text{lan}}$ 更新计算,如式 (11) 所示。整个故事词汇的概率分布由 $p(y_{j,t})$ 表示,如式 (12) 所示:

$$h_{j,t}^{\text{lan}} = \text{GRU}_{\text{lan}}(h_{j,t-1}^{\text{lan}}, \hat{v}_{j,t}, \hat{p}_{j,t}, h_{j,t}^{\text{VA}}, h_{j,t}^{\text{PA}}) \quad (11)$$

$$p(y_{j,t}) = \text{softmax}(\tanh(W_p h_{j,t}^{\text{lan}} + b_p)) \quad (12)$$

其中, W_p 和 b_p 为学习权值和偏差。

IC_PLVS 模型的优化目标是最小化以视觉、规划为条件的交叉熵损失,如式 (13) 所示:

$$L(\psi) = \sum_{j=1}^J \sum_{m=1}^M \sum_{n=1}^{N_M} \log(p_{\psi}(W_{j,m,n}^* | W_{j,m,1:n-1}^*)) \quad (13)$$

其中, $W_{j,m,n}^*$ 表示参考答案的单词, $y_{j,m}^* = \langle W_{j,m,1}^*, W_{j,m,2}^*, \dots, W_{j,m,n}^* \rangle$, ψ 是 GRU_{lan} 的参数。

3 实验分析

3.1 数据集和实验环境

本次实验将 Ubuntu 18.04 64 位系统

作为开发平台，并依托PyTorch框架完成模型的训练与测试；采用硬件Intel i9-9900k处理器及NVIDIA GeForce RTX 4090显卡；在VIST数据集^[10]进行实验，经过筛选处理，除去损坏或版权受限的图片后，保留10 033个完整的相册集合，共计80 245张图片，共关联14 526个词汇。本次实验从总数据集中划分出39 659个样本用于模型训练，4 988个样本作为验证集以调整超参数，另有5 050个样本构成测试集，用于最终评估模型在未知数据上的泛化能力。在模型配置层面，笔者设置了多项参数以优化性能：批大小为128，以平衡内存使用与训练效率；图像通过编码处理转化为2 048维的向量；注意力机制则配置为2层4头；学习率为0.0004；规划向量与词嵌入的维数均被设定为512。

3.2 评价指标

本次实验采用多种评估指标评估提炼算法模型的效果，包括BLEU- n (B n)^[21]、METEOR (ME)^[22]、ROUGE-L (RL)^[23]和CIDEr (CIr)^[24]。以上方法主要评估模型的准确度，其中BLEU- n 侧重于生成文

本的准确性和流畅度；METEOR考虑单词的变形；ROUGE-L侧重于召回率；CIDEr主要评估能否捕捉图像的关键信息。此外，实验采用多样性指标：Distinct- n ^[25] (D n)和TTR^[26]评估生成故事的多样性。Distinct- n 计算文本中不同 n -gram的数量与总 n -gram数量的比率；TTR计算文本中不同单词与总单词数量的比率。使用上述客观评估指标，能够全面评估本文提出的算法模型在视觉故事生成任务中的表现。

3.3 与主流方法实验结果对比

本文对多个主流方法进行了比较，具体包括AREL^[1]、CKACS^[11]、KE-VIST^[6]、PR-VIST^[3]、COVS^[7]、ViT-model^[12]和SCO-VIST^[17]，比较结果见表3。AREL模型通过强化学习优化通用评价指标的表现。KE-VIST模型预测与图像序列相关的概念术语，并结合外部知识输入Transformer生成故事。PR-VIST模型借鉴人类写作，设计故事脉络策划模型，以提升故事连贯性。COVS^[4]模型利用主题模型挖掘相册主题分布，确保生成的叙述与

表3 本文模型与其他模型对比

模型	B1	B2	RL	CIr	ME	D3	D4	TTR
AREL	0.6041	0.2483	0.2782	0.0641	0.3393	0.5719	0.7020	0.6056
KE-VIST	0.6407	0.3509	0.2891	0.0789	0.3481	--	--	--
PR-VIST	0.6397	0.3527	0.261	0.0406	0.314	--	--	--
COVS	0.6410	0.3591	0.3012	0.0853	0.3543	0.5658	0.6916	0.5514
ViT-model	0.6300	0.3750	0.3100	0.0540	0.3536	--	--	--
SCO-VIST	--	0.347	0.221	0.0591	0.2750	--	--	--
qwen-nl2sql*	0.2556	0.0859	0.1384	0.016	0.0821	0.5743	0.7181	0.6079
Llama3.2-vision*	0.2145	0.0928	0.1241	0.023	0.1733	0.5557	0.7164	0.7233
IC-PLVS	0.6617	0.4020	0.2994	0.0897	0.3538	0.5836	0.7202	0.6197

图像主题一致。ViT-model 通过分割图像、提取特征并利用 Bidirectional-LSTM 进行编码，生成故事。SCO-VIST 模型通过图结构优化，构建标题、常识和主题的异构图，捕捉并发展故事情节。

根据表 5 可知，在 VIST 测试集上，IC_PLVS 模型相对其他主流方法在 5 个评价指标上表现相对较好。与 COVS 模型相比，IC_PLVS 在 B1、B2 和 C1r 指标上分别提升了 3.23%、11.95% 和 5.15%，显示出在故事准确性和连贯性方面的优化；在 D3、D4 和 TTR 指标上分别提升 3.15%、4.14% 和 12.39%，显示出在丰富故事内容和增强表达多样性方面的优势。此外，IC_PLVS 在各项指标上均优于 AREL、PR-VIST 等模型，表明引入规划与概念模块有助于捕捉并发展故事情节，生成低冗余度和细节丰富的故事。另外，本文从数据集中随机挑选 100 个示例，在通义千问模型（qwen-nl2sql）和 Llama3.2-vision 上进行了实验，实验结果用*标出。

3.4 消融实验

为了验证引入规划模块信息和图像分类对模型生成的有用性，在保持基线模型结构不变的前提下，本文设计了多组对比

实验，见表 4。其中，I 指进行图像处理；D 指引用 Hsu 等^[3]提出的概念抽取方法从图像中提取概念信息；C 指引入图像分类方法；P 指引入规划方法；U 指将新的单词加入词库中。从实验效果发现，引入概念信息或规划信息都能丰富故事，其中规划信息强于概念信息，但两者结合效果减弱。此外引入图像分类方法可提升故事生成质量。

为了验证规划模块信息维度选取、信息组合方法以及解码器搭建方法的有效性，本文设计了多组对比实验，具体结果见表 5。针对每个图像包含的 12 个维度规划信息，选取了从 5~11 共 7 个维度，确定规划信息的最佳引入维度。此外，为了估规划信息的选取方式对实验结果的影响，本文设置了两种策略：随机选取 n 维信息（以“IC_PLVS_r”为前缀）；按顺序选取前 n 维度信息（以“IC_PLVS_fix”为前缀）。最后，为了评估解码器模块中解码结构的影响，设置了一个对照组实验 IC_PLVS(2plan)：使用通用规划信息解码器、特定规划信息解码器，分别解码规划信息。对比实验结果发现，解码结构对规划信息的影响并不显著，从而验证了当前解码器设计的合理性。

表4 消融实验1

方法					B1	B2	RL	C1r	ME	D3	D4	TTR
I	D	C	P	U								
✓					0.6060	0.2572	0.2793	0.0715	0.3326	0.5718	0.7013	0.5852
✓	✓			✓	0.6237	0.3839	0.2919	0.0700	0.3419	0.5820	0.7137	0.6060
✓			✓	✓	0.6412	0.3951	0.3017	0.0820	0.3563	0.5726	0.7028	0.5708
✓		✓	✓		0.6562	0.4005	0.2985	0.0880	0.3530	0.5810	0.7170	0.6170
✓	✓	✓	✓	✓	0.6568	0.3860	0.3052	0.0845	0.3456	0.5880	0.7177	0.6074
✓		✓	✓	✓	0.6617	0.4020	0.2994	0.0897	0.3535	0.5836	0.7202	0.6197

表5 消融实验2

方法	方法	B1	B2	RL	CIr	ME	D3	D4	TTR
随机 选取	IC_PLVS_r5	0.6534	0.3987	0.2991	0.0873	0.3498	0.5882	0.7212	0.5962
	IC_PLVS_r6	0.6504	0.3911	0.2952	0.0823	0.3468	0.5893	0.7287	0.6333
	IC_PLVS_r7	0.6536	0.3959	0.2984	0.0879	0.3505	0.5835	0.7203	0.6189
	IC_PLVS_8	0.6551	0.4002	0.3010	0.0894	0.3536	0.5821	0.7168	0.6045
	IC_PLVS_r9	0.6530	0.3968	0.2990	0.0898	0.3502	0.5804	0.7153	0.6054
	IC_PLVS_r10	0.6507	0.3988	0.3022	0.0889	0.3522	0.5765	0.7112	0.5870
	IC_PLVS_r11	0.6573	0.3974	0.2938	0.0871	0.3484	0.5850	0.7239	0.6300
顺序 抽取	IC_PLVS_fix5	0.6426	0.3921	0.2964	0.0849	0.3485	0.5767	0.7123	0.6236
	IC_PLVS_fix6	0.6490	0.3939	0.2969	0.0837	0.3532	0.5858	0.7213	0.6224
	IC_PLVS_fix7	0.6453	0.3945	0.2989	0.0858	0.3529	0.5830	0.7175	0.6195
	IC_PLVS_fix8	0.6538	0.3963	0.2968	0.0921	0.3521	0.5781	0.7138	0.5997
	IC_PLVS_fix9	0.6572	0.4006	0.2985	0.0880	0.3536	0.5845	0.7204	0.6193
	IC_PLVS_fix10	0.6565	0.4026	0.2995	0.0918	0.3543	0.5781	0.7134	0.6012
	IC_PLVS_fix11	0.6576	0.3978	0.2997	0.0913	0.3531	0.5864	0.7199	0.6189
解码 结构	IC_PLVS(plan)	0.6617	0.4020	0.2994	0.0897	0.3535	0.5836	0.7202	0.6197
	IC_PLVS(2plan)	0.6607	0.4031	0.3008	0.0925	0.3539	0.5726	0.7183	0.6169

3.5 可视化故事生成示例

本节对比 IC_PLVS 与 COVS^[7]、PR-VIST^[3]在评估实验中的具体表现。首先从图像内容贴近度方面来看，表6中COVS生成描述侧重于对人和景色的描写，例如“we”和“tree”等表达，但忽略了关键元素——狗，因此生成的故事存在一定的偏差。而PR-VIST缺少对细节的描述，未提及“dog”“birds”等核心元素。相比之下，IC_PLVS能较好地规避以上情况，通过将图像序列的第2、3、5张图片归类为animal并引入规划信息，能更贴近图像内容。因此引入图像分类可有效提升故事与图像的匹配度。在故事的丰富度方面，表7中的数据显示，COVS生成的故事中

“beautiful”一词被多次使用，显得故事单调，缺乏变化。而IC_PLVS通过引入更多样化的词汇和表达方式，如以“couple”“room”“nice”等作为修饰，提高了句子的丰富度。

4 结束语

本文首先对输入图像进行分类；接着利用Monkey完成预先设定的两类共12个问题，将问答作为中间规划步骤，有效筛选关键概念，指导故事构建；最后通过将规划与视觉特征一起输入解码器，生成与图像内容紧密相关的故事。实验证明，该模型在生成

表6 图像分类有用性分析

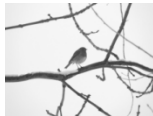
图像序列					
图像类型	people	animals	animals	landscape	animals
COVS	we decided to go to a walk yesterday. the trees were covered in snow. the is a picture of a snow. we had to go to the park to see the snow. the birds was in the air				
PR-VIST	i was out for a walk in the city. the ground was covered in snow. we were able to take a walk through the snow. we made it through the snow. it was very cold outside. the trees were covered in snow				
IC_PLVS	plan	snow ; shoes; snow; elderly; standing; tall; some; delighted; feet; boots; snow; walk ; winter; dog ; winter; 1; dog; fierce; hunt; snow; standing; curious; walk; pee; winter; prints; snow ; 1; dog; fierce; hunting; snow; walking; happy; walking; run; winter; playground; park ; winter; forests; peaceful; stunning; snow; winter; cold; white; trees; winter; bird ; outside; 1; bird; fierce; perch; natural; sitting; calm; eats; fly;			
	story	i had to go to the snow last week. the dog were very friendly . there was a snow on the ground . we had to go out to the park . the birds were so beautiful .			
Gold 文本	i awoke to snow this morning. frankly my dog decided he wanted to take a run through it. it wasn't very deep. the playground next to my home looked so quiet and serene. even the birds were chirping a little happier this morning. It's so beautiful when the snow falls				

表7 冗余度分析

图像序列					
图像类型	people	indoor	landscape	landscape	building
COVS	we went to visit the old town. it is a beautiful place to be. the was a very beautiful day. this is a picture of a build - ing. the hotel was beautiful and beautiful				
PR-VIST	my parents were from out of the country. there was a small apartment in our room. we went for a walk around the apartment . the area looked very nice. i saw a lot of water that had everything in it.				
IC_PLVS		walk ; people ; building; elderly; trip ; blond; several; happy; friendship; color; fireplace; living room ; modern; sitting; room; calm; brown; relaxation			
	plan	fall; bench; forest ; elderly; standing; white hair; several; happy; friendship; light; picnic; eat; fall; bench; outside; forest; peaceful; stunning; sunny; mild; green; trees ;			
	story	restaurant; tables; downtown; arches; modern ; fence; peaceful; imposing; concrete; pollution; building ; the couple was excited to be a day trip . the inside was a nice room . and, we went on a walk through the woods . this is a picture of a modern building . i had a great time there			
Gold 文本	we arrived at the lodge for vacation. indoor was very quaint, with a fireplace that looked cozy. the deck was beautiful, wrapped around a large tree. the view was so peaceful. it was hard to believe that we were just across the water from the city				

符合图像内容、捕捉图片细节信息、丰富内容上的表现与预期相符。在实际应用场景中，一幅图片可能同时隶属于多个类别，后

续计划开展图像多分类的研究，进一步优化视觉故事生成方法，以在应对复杂图像类型时展现出更卓越的性能与适应性。

参考文献:

- [1] WANG X, CHEN W H, WANG Y-F, et al. No metrics are perfect: adversarial reward learning for visual storytelling[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 899-909.
- [2] HU J J, CHENG Y, GAN Z, et al. What makes a good story? designing composite rewards for visual storytelling[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7969-7976.
- [3] HSU C Y, CHU Y-W, HUANG T-H, et al. Plot and rework: modeling storylines for visual storytelling[C]//Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg: ACL, 2021: 4443-4453.
- [4] LIU D Y, LAPATA M, KELLER F. Visual storytelling with question-answer plans [C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: ACL, 2023: 5800-5813.
- [5] HSU C C, CHEN S M, HSIEH M H, et al. Using inter-sentence diverse beam search to reduce redundancy in visual storytelling[EB]. arXiv preprint, 2018, arXiv: 1805.11867.
- [6] HSU C C, CHEN Z-Y, HSU C Y, et al. Knowledge-enriched visual storytelling [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7952-7960.
- [7] GU J J, WANG H L, FAN R C. Coherent visual storytelling *via* parallel top-down visual and topic attention[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(1): 257-268.
- [8] CHENG S J, GUO Z C, WU J, et al. Ego-Think: evaluating first-person perspective thinking capability of vision-language models[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 14291-14302.
- [9] 司亚中, 张旭龙, 杨帆, 等. 沙尘图像视觉增强技术综述[J]. 大数据, 2025, 11(1): 117-134.
- [10] SI Y Z, ZHANG X L, YANG F, et al. A survey of sand-dust image visual enhancement techniques[J]. Big Data Research, 2025, 11(1): 117-134.
- [10] HUANG T K, FERRARO F, MOSTAFAZADEH N, et al. Visual storytelling[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2016: 1233-1239.
- [11] CHEN H, HUANG Y F, TAKAMURA H, et al. Commonsense knowledge aware concept selection for diverse and informative visual storytelling[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 999-1008.
- [12] MALAKAN Z M, HASSAN G M, MIAN A. Vision transformer-based model for describing a set of images as a story[C]//Proceedings of AI 2022: Advances in Artificial Intelligence. Cham: Springer, 2022: 15-28.
- [13] CHEN W, LIU X F, NIU J W. SentiStory: a multi-layered sentiment-aware generative model for visual storytelling[C]//Proceedings of the IEEE Transactions on Circuits and Systems for Video Technology. Piscataway: IEEE Press, 2022: 8051-8064.
- [14] YANG D Y, JIN Q. Attractive story-

- teller: stylized visual storytelling with unpaired text[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 11053–11066.
- [15] SONG Y J, PAPERNO D, GATT A. Context-aware visual storytelling with visual prefix tuning and contrastive learning[C]//Proceedings of the 17th International Natural Language Generation Conference. Stroudsburg: ACL, 2024: 384–401.
- [16] LIN X C, CHEN X Y. Improving visual storytelling with multimodal large language models[EB]. arXiv preprint, 2024, arXiv:2407.02586.
- [17] WANG E, HAN C R, POON J. SCOVIST: social interaction commonsense knowledge-based visual storytelling[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 1602–1616.
- [18] LI Z, YANG B, LIU Q, et al. Monkey: image resolution and text label are important things for large multi-modal models[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 26753–26763.
- [19] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248–255.
- [20] 曹峰, 李文涛, 骆剑承, 等. 融合光谱度量标记迁移和 Tri-training 的高光谱遥感图像半监督分类算法[J]. 大数据, 2023, 9(6): 72–89.
- CAO F, LI W T, LUO J C, et al. Semi-supervised classification algorithm for hyperspectral remote sensing images fusing spectral measure-based label transfer and tri-training[J]. Big Data Research, 2023, 9(6): 72–89.
- [21] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Morristown: ACL, 2001: 311.
- [22] BANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Stroudsburg: ACL, 2005: 65–72.
- [23] LIN C Y, OCH F J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. New York: ACM, 2004: 605–608.
- [24] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr: consensus-based image description evaluation[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 4566–4575.
- [25] LI J W, GALLEY M, BROCKETT C, et al. A diversity-promoting objective function for neural conversation models[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2016: 110–119.
- [26] CONNORS L H, LIM A, PROKAEVA T, et al. Tabulation of human transthyretin (TTR) variants, 2003[J]. Amyloid, 2003, 10(3): 160–184.

作者简介



王元龙（1982- ），男，博士，山西大学计算机与信息技术学院副教授，主要研究方向为自然语言处理、图形图像处理。



张宁倩（1998- ），女，山西大学计算机与信息技术学院硕士生，主要研究方向为视觉故事生成。



张虎（1979- ），博士，山西大学计算机与信息技术学院教授，主要研究方向为自然语言处理。

收稿日期: 2025-05-19
基金项目: 国家自然科学基金项目(No.62176145, No.62476161)
Foundation Items: The National Natural Science Foundation of China (No.62176145, No.62476161)