计算机科学
COMPUTER SCIENCE

基于规划学习的视觉故事生成模型

王元龙, 张宁倩, 张虎

引用本文

王元龙, 张宁倩, 张虎. 基于规划学习的视觉故事生成模型[J]. 计算机科学, 2025, 52(9): 269-275.

WANG Yuanlong, ZHANG Ningqian, ZHANG Hu. Visual Storytelling Based on Planning Learning[J]. Computer Science, 2025, 52(9): 269-275.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700136

基于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　太原 030006

摘　要　近年来,视觉故事生成受到越来越多的计算机视觉和自然语言处理领域学者的关注。现有模型大多侧重于增强图像表示,例如引入外部知识、场景图等,虽然取得了一些进展,但生成的故事仍存在内容重复使用和细节描述少的问题。针对上述问题,提出了基于规划学习的视觉故事生成模型[1)],引入规划学习方法,从主题、对象、动作、地点、推理、预测6个维度设定对应的问题,利用视觉问答预训练语言模型生成答案,完成规划设计,引导视觉故事生成。模型分为4阶段:第一阶段从图片中提取视觉信息;第二阶段通过概念生成器抽取并选择相关概念;第三阶段利用预训练语言模型引导规划信息生成;第四阶段融合前3个阶段生成的视觉、概念和规划信息,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,ROUGE_L,Distinct-3,Distinct-4和TTR指标上提升了1.58百个分点、2.7百个分点、0.4百个分点、2.2百个分点、3.6百个分点和5.6百个分点。

关键词:视觉故事生成;规划学习;视觉问答
中图分类号　TP391

Visual Storytelling Based on Planning Learning

WANG Yuanlong,ZHANG Ningqian and ZHANG Hu
School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China

Abstract　Visual storytelling is a growing area of interest for scholars in computer vision and natural language processing. Current models concentrate on enhancing image representation,like using external knowledge and scene diagrams. Despite some advancements have been made,they still suffer from content reuse and lack of detailed descriptions. To address these issues,this paper proposes a visual story generation model that incorporates planning learning. It poses questions across six key dimensions—theme,object,action,place,reasoning,and prediction—and uses a pretrained visual question answering language model to generate detailed answers. This approach guides the planning and designs process,leading to more nuanced visual story generation. The model is divided into four stages. The first stage extracts visual information from pictures. The second stage extracts and selects relevant concepts through the concept generator. The third stage is used pre-trained language models to guide the generation of planning information. In the fourth stage,it integrates the visual,conceptual and planning information generated in the above three stages to complete the visual story generation task. The model's effectiveness is validated on the VIST dataset,outperforming the COVS model with improvements in BLEU-1,BLEU-2,ROUGE_L,Distinct-3,Distinct-4,and TTR scores by 1.58 percentage points,2.7 percentage points,0.4 percentage points,2.2 percentage points,3.6 percentage points,and 5.6 percentage points respectively.

Keywords　Visual storytelling,Planning learning,Visual question answering

1　引言

视觉故事生成任务是一项具有挑战性的视觉理解任务,旨在从一系列图片中提取关键视觉特征,并据此创作出符合图像特征且语句流畅、逻辑严密的故事。相比于其他的视觉识别任务,本任务除了要求模型对单个图像的视觉元素进行准确捕捉和语义理解外,还需挖掘并表达图片之间复杂的逻辑关系和叙事脉络。

近年来,关于视觉故事生成的研究主要集中在提升视觉表示的质量和精确度。Wang等[1]为了提升故事生成的细节层次和准确性,利用场景图来增强图像的视觉表示,将每张图像及关系转换成图形化的语义表示。Hsu等[2]打破视觉故事生成信息来源的局限,通过融入外部知识,为故事增添了新的维度。Li等[3]集合场景图和外部常识知识图谱,通过图像自扩充和图像序列关联扩充技术,增强了图像的语义概念。Gu等[4]引入主题模型算法来引导故事生成,使得生成的故事更

1) 所提出模型的源代码见 https://github.com/anna171717/PLVS

到稿日期:2024-07-22　返修日期:2024-11-01
基金项目:国家自然科学基金面上项目(62176145)
This work was supported by the National Natural Science Foundation of China(62176145).
通信作者:王元龙(ylwang007@163.com)

加贴合图像的主题内容，提升了故事的相关性。Liu 等[5]提出了基于问答策略的故事生成模型，该模型将图像转换为文本描述，再基于文本描述生成一系列问答对，能够生成与图像序列紧密相关的故事。以上方法虽然取得了一些进展，但仍然存在一些挑战，如表 1 所列。1)内容重复使用，造成所生成的故事语义模糊不清。如表 1 示例 1“the band is getting ready to perform for the band”和“the band played a song for the band”中，“the band”多次出现。2)缺乏具体的细节描述，遗漏图片中的重要信息点，导致读者无法深入理解和感受图片所传递的信息。如表 1 示例 2 中提到“had a great time”，但是未与图片中的具体细节联系，如人物动作、图片对象(美味的食物等)、场景等，这会显得内容很空洞。

表 1 示例分析

Table 1 Example analysis

内容重复使用	示例 1：the band is getting ready to perform for the band. the band played a song for the band. the lead singer of the band is very talented. the band played on the guitar. the lead guitarist is very talented.
缺乏细节描述	示例 2：the girls had a great time at the party. everyone had a great time. it was a great time. it was a great day. i'm so glad i had a great time.

因此，基于上述问题，受到 Cheng 等[6]在评估视觉-语言模型任务中提出的第一人称视角思维的启发，从 6 个维度(主题、对象、动作、地点、推理和预测)引导预训练语言模型基于图像本身回答相关问题，再通过六维信息构思故事情节，提出基于规划学习的视觉故事生成模型(Planning Learning Visual Storytelling，PLVS)。模型通过全面规划故事的骨架，有效避免了内容的冗余和重复。另外，多角度的问题探讨不仅提升了视觉内容与文本之间的匹配精度，还丰富了故事的层次和细节，能够从不同视角深入探索和表达图像信息，生成具有丰富性的视觉故事。本文主要贡献体现在以下 3 个方面。

1)设计规划模块。从 6 个维度规划图像内容，构建故事框架，减少冗余信息的生成，使得故事更全面。

2)引入问答机制。通过视觉问答预训练语言模型，获取每个图像的细节信息，从而丰富故事内容，提升对图像信息的利用效率。

3)在公开数据集 VIST[6](Visual Storytelling)上的实验结果表明，本文模型在 BLEU-1，BLEU-2，ROUGE_L，Distinct-3，Distinct-4 和 TTR 中优于基线模型。

2 相关工作

故事生成任务属于计算机视觉和自然语言处理领域研究的交叉课题，是衡量人工智能发展水平的一个重要因素，对于推动人工智能在语言理解、图像识别以及创造性写作等方面的进步具有重要意义。Huang 等[7]首次提出了视觉故事生成的概念，并创建了 VIST 数据集，为后续的研究打下了基础。目前，视觉故事生成方法主要分为两大类：端到端模型和两阶段生成模型。端到端模型通常利用深度学习技术，通过构建图像到故事的直接映射完成视觉故事生成任务。Kim 等[8]为了保证生成故事的上下文一致性并提高句子连贯性，采用了全局-局部注意力和上下文级联机制来处理图像数据。Wang 等[9]提出了强化学习框架，通过两个鉴别器作为奖励，使得层次生成模型能够生成相关且具有故事风格的叙事段落；同时，使用对抗训练策略来进一步提升叙事能力，使故事生成器和鉴别器的学习相互加强。以上方法虽然在简洁性和效率上有优势，但因为知识有限，生成的故事不仅信息量较少，而且缺乏趣味性。

与端到端模型相比，两阶段生成模型首先对图像内容进行深入分析，以创建一个中间层的抽象表示。Chen 等[10]为每个图像序列创建了一个常识知识图谱，并引入概念选择模块。该模块在两阶段视觉故事系统中用于选择相关概念，将选定的概念与图像特征结合起来，生成一个完整的故事。另一方面，Hsu 等[11]构建故事图，通过提取图像中的各种元素(如术语节点、对象节点)并使用外部知识(如 Visual Genome 图等)将元素链接起来，找到构成最佳故事情节的最优路径，帮助模型捕捉图像中的关键信息，生成更为丰富和结构化的故事。然而，两阶段生成模型大多包含多个阶段和组件，导致生成故事中句子内部重复信息过多，降低了故事整体的清晰度和理解度。

在人类写作故事的过程中，规划学习已被证明是一种非常有效的策略。其不仅能够捕获故事的核心内容，还能够确立一个符合逻辑顺序的故事结构。在视觉故事生成领域，Hsu 等[11]首次引入规划学习的概念，通过基于训练数据和外部资源构建的结构图表示图像序列，并从中识别出得分最高的路径作为理想的故事情节线。Narayan 等[12]通过将问答对作为生成中间规划，实现了对生成内容的细致控制，并为模型的预测提供了解释。Liu 等[5]进一步将此类规划策略扩展至多模态场景，通过结合预训练的语言模型和视觉信息表示，在生成完整故事前加入中间规划步骤。该方法将阅读理解数据集转换为问题生成数据集，并微调序列到序列的转换器模型，以预测与给定答案和上下文相匹配的问题，从而引导故事的生成。然而，模型难以匹配视觉信息和相应的文本语料，无法全面捕捉到图像中的所有关键信息，影响了故事的清晰度和可理解性。

因此本文提出了一种新的规划学习方法，该方法以图像为核心，将图像与预设的问题输入预训练语言模型中生成答案。随后，将规划内容与概念选择器生成的概念相结合并输入解码器中，以此来完成视觉故事的生成任务。在构建规划问题的过程中，采用了 Cheng 等[6]在评估视觉-语言模型任务中提出的第一人称视角思维的方法，将规划的问题划分为 6 个维度：主题、对象、动作、地点、推理和预测。六维度是内容选择(即确定故事需要讲述哪些内容)和规划的基础。本文方法的核心优势在于：允许模型在生成故事时，始终保持对整体结构和情节发展的宏观把控；通过定向规划，模型能够避免使用重复的词汇和冗余的描述，从而提升了故事的质量和可读性；确保了故事各个部分之间的逻辑相连，提供了流畅的叙述体验；故事进展有序，每个情节都建立在前一个情节的基础上，保证了故事的逻辑性和完整性。

3 模型

本文模型分为 4 个阶段。阶段 1：图像特征提取阶段。

由 CNN 作为图像序列编码器，提取每张图像的特征。阶段 2：从每张图像中提取一组概念述语，并结合外部知识概念词，最终筛选出与图像最相关的概念词。阶段 3：利用规划学习策略来引导视觉问答预训练语言模型生成相关内容。首先，逐步将 5 张图片输入预训练模型，每次提出 6 组问题：图像的主题是什么？图像中有哪些对象？图像中的事件是在哪里发生的？图像中的人或动物在做什么？图像中的物体下一步计划做什么？图像中可能发生什么？然后，将问题和生成的答案相结合，作为规划信息输入解码器中。阶段 4：故事生成阶段。图 1 概述了基于规划学习的视觉故事生成模型的工作流程，该模型主要包括图像特征提取模块、概念选择模块、规划学习模块和故事生成模块，其分别对应上述的 4 个阶段，以下将进行详细介绍。

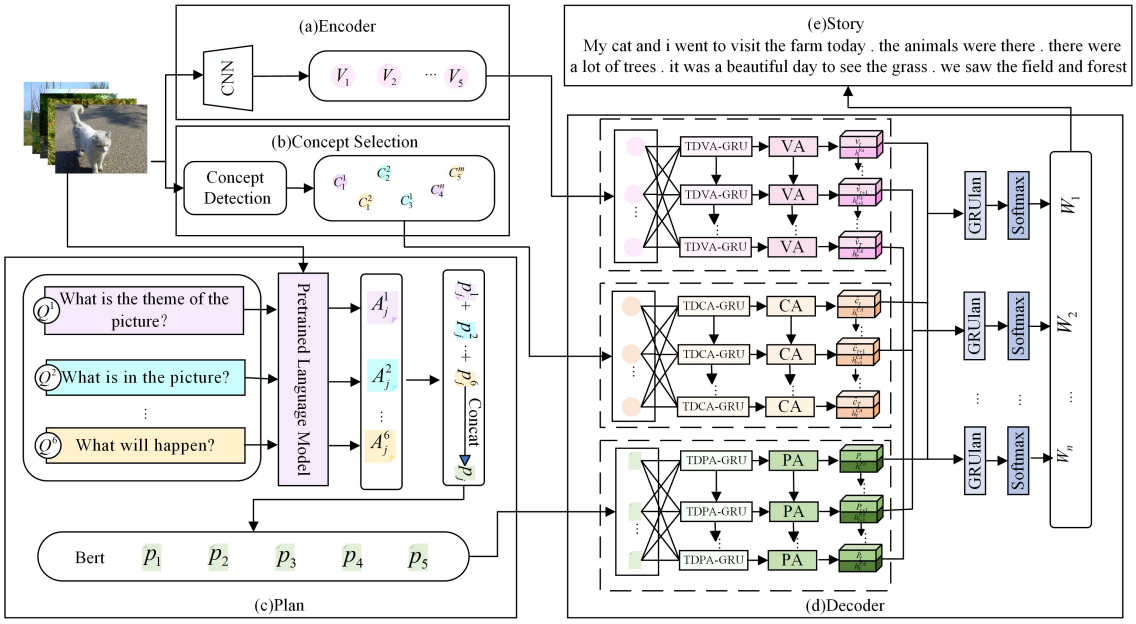

图 1 基于规划学习的视觉故事生成模型

Fig. 1 Visual storytelling based on planning learning

3.1 图像特征提取模块

在处理顺序输入的相册时，利用卷积神经网络(CNN)作为图像序列的编码器，以提取和编码每张图像的深层特征。为了有效地整合图像的特征，引入平均池化(Mean Pooling，MP)，通过计算每个特征维度上所有图像嵌入的平均值，进行全局性的累积与融合，实现图像的特征嵌入。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。上述方法能够有效减少信息的冗余，提高特征表示的可靠性和准确性，从而为后续任务奠定基础。

3.2 概念选择模块

概念选择模块采用了 Hsu 等[11]提出的方法，包括 3 个步骤。首先，将 Faster R-CNN[13](Faster Region-based Convolutional Neural Network)作为对象检测器，Transformer-GRU[2](Transformer Gated Recurrent Unit)作为概念预测器，从图像中提取概念。其次，为了关联不同的概念词，利用 Visual Genome 和 VIST 两个知识图谱收集概念词之间的动词关系，从而扩充概念集。最后，为故事生成模型选择合理的概念述语集。本次研究中，使用基于门控循环单元(Gated Recurrent Unit，GRU)的模型在所有可用的文本故事上进行训练，对概念集进行选择和过滤，得到最终的概念预测结果。此过程中，数字上标表示图片序号，数字下标表示图片中的具体概念，得出最终的概念预测结果：$C(C_1^1,C_1^2,\cdots,C_3^n,\cdots,C_5^m)$。上述方法结合了图像特征提取、概念预测和知识库扩充的步骤，为故事生成模型提供了更丰富的概念信息。

3.3 规划模块

规划模块的核心任务是利用 VQA(Visual Question Answering)预训练语言模型[14]生成规划内容，分为两个关键步骤。首先，VQA 模型接收一组事先设定的问题，并据此生成相应的答案。在问题设定方面，本文借鉴了 Cheng 等[6]提出的从第一人称视角出发的评估法，从 6 个维度即主题、对象、地点、动作、计划和预测出发，将问题设定为不同形式，如表 2 所列。

表 2 规划维度

Table 2 Planning dimensions

维度	问题	目的
theme	What is the theme of the picture?	主题
object	What is in the picture?	对象
location	Where the events in the image took place?	地点
active	What is the person or animal in the picture doing?	动作
plan	What is people or animal going to do?	推理
forecast	What will happen?	预测

通过设定多维度问题，模型能够获取图像的关键信息，如故事背景(地点)、故事细节(主要对象、动作)、故事发展(推理、预测)，从而全面地理解和解释图像内容，并生成一个结构化的规划框架，为下一步的故事生成提供基础。此方法不仅提高了故事的丰富度，也使得生成的故事更加贴近人类的

认知和叙事方式。

从 VQA 获得图像对应问题的答案后，首先将问题和答案拼接，组成 question:Q^i[SEP]answer:A_j^i 的形式，其中 Q^i 代表第 i 维规划的问句，分别对应于表 2 中的问句，A_j^i 代表第 j 个图像的 i 维规划规划的答案，后由[SEP]将第 j 个图像的每维规划拼接，并在首部添加[CLS]组成第 j 张图片的规划信息字符串 P_j，如图 2 所示。上述过程最终形成的规划结果为 $P(P_1, P_2, \cdots, P_5)$，随后将其输入 BERT(Bidirectional Encoder Representations from Transformer)中进行深度语言特征提取，以捕捉图片内容的丰富语义并兼顾上下文关系。规划信息最终转换为词嵌入作为输入数据，被送入故事生成模块，为故事构建基础框架。

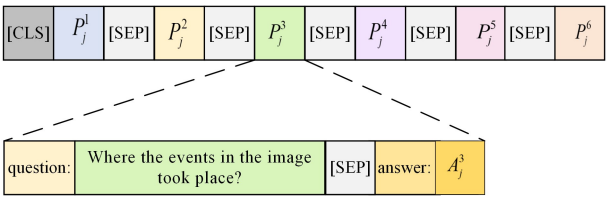

图 2 规划拼接

Fig. 2 Plan the mosaics

此模块的优势在于：通过将 5 张图片的信息整合成统一的规划内容的方法，为故事生成构建了一个全面而连贯的信息框架；能够较好地丰富故事生成的背景和细节，使得生成的故事在内容上更丰富，在逻辑和情节上更紧密。通过上述规划和整合流程后，后续的故事生成模块能更有效地利用视觉信息，丰富故事细节，减少冗余情况的产生。

3.4 故事生成模块

故事生成模块由信息解码器和连贯语言生成器组成，其主要任务是利用视觉信息、概念信息及规划信息引导模型生成上下文连贯的故事。此模块的构成如图 1 所示。

信息解码器：分别解码视觉、概念和规划输入中的不同信息。首先利用任务特定的上下文预测注意分布，然后以加权平均特征计算注意特征向量。GRU 模型由视觉解码器(TDVA-GRU)、规划解码器(TDPA-GRU)、概念解码器(TDCA-GRU)3 部分组成。

向 TDVA-GRU 中输入时间步长 $t-1$ 的隐藏状态 $\boldsymbol{h}_{j,t-1}^{\text{lan}}$、视觉信息 $\boldsymbol{v}_{j,t}$ 及时间步长 $t-1$ 的单词信息 $\boldsymbol{W}_{e\chi_{t-1}}$，获得时间步长 t 的视觉隐藏状态，如式(1)所示。

$$\boldsymbol{h}_{j,t}^{\text{VA}} = TDVA\text{-}GRU(\boldsymbol{h}_{j,t-1}^{\text{lan}}, \boldsymbol{v}_{j,t}, \boldsymbol{W}_{e\chi_{t-1}}) \tag{1}$$

其中，$\boldsymbol{h}_{j,t-1}^{\text{lan}}$ 表示第 j 个图像序列在时间步长 $t-1$ 时，GRU_{lan} 的隐藏状态；$\boldsymbol{v}_{j,t}$ 表示在时间步长 t 下，CNN 对第 j 个图像序列的输出；$\boldsymbol{W}_{e\chi_{t-1}}$ 表示上一个单词的编码，$\boldsymbol{W}_e \in R^{E\times\Omega}$，$R^{E\times\Omega}$ 是一个词汇表，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，词嵌入使用 word2vec 计算。

向 TDCA-GRU 中输入时间步长 $t-1$ 的隐藏状态 $\boldsymbol{h}_{j,t-1}^{\text{lan}}$、时间步长 t 时概念模块的输出 $\boldsymbol{c}_{j,t}$ 及时间步长 $t-1$ 的单词信息 $\boldsymbol{W}_{e\chi_{t-1}}$ 后，获得时间步长 t 的概念隐藏状态，如式(2)所示。

$$\boldsymbol{h}_{j,t}^{\text{CA}} = TDCA\text{-}GRU(\boldsymbol{h}_{j,t}^{\text{CA}}, \boldsymbol{c}_{j,t}, \boldsymbol{W}_{e\chi_{t-1}}) \tag{2}$$

向 TDPA-GRU 中输入时间步长 $t-1$ 的隐藏状态 $\boldsymbol{h}_{j,t-1}^{\text{lan}}$、时间步长 t 时规划模块的输出 $\boldsymbol{p}_{j,t}$ 及时间步长 $t-1$ 的单词信息 $\boldsymbol{W}_{e\chi_{t-1}}$ 后，获得时间步长 t 的规划隐藏状态，如式(3)所示。

$$\boldsymbol{h}_{j,t}^{\text{PA}} = TDPA\text{-}GRU(\boldsymbol{h}_{j,t-1}^{\text{lan}}, \boldsymbol{p}_{j,t}, \boldsymbol{W}_{e\chi_{t-1}}) \tag{3}$$

在时间步长 t 处，图像归一化注意力权重($\boldsymbol{v}_{j,t}^*$)、规划归一化注意力权重($\boldsymbol{p}_{j,t}^*$)、概念归一化注意力权重($\boldsymbol{c}_{j,t}^*$)分别如式(4)－式(6)所示。

$$\boldsymbol{v}_{j,t}^* = \tanh(\boldsymbol{W}_v \boldsymbol{v}_{j,t} \oplus \boldsymbol{W}_a \boldsymbol{h}_{j,t-1}^{\text{VA}}) \tag{4}$$

$$\boldsymbol{p}_{j,t}^* = \tanh(\boldsymbol{W}_p \boldsymbol{p}_{j,t} \oplus \boldsymbol{W}_b \boldsymbol{h}_{j,t-1}^{\text{PA}}) \tag{5}$$

$$\boldsymbol{c}_{j,t}^* = \tanh(\boldsymbol{W}_c \boldsymbol{c}_{j,t} \oplus \boldsymbol{W}_c \boldsymbol{h}_{j,t-1}^{\text{CA}}) \tag{6}$$

其中，$\boldsymbol{W}_v$，$\boldsymbol{W}_a$ 等是变换矩阵；$\tanh(\cdot)$为双曲正切函数；$\boldsymbol{h}_{j,t-1}^{\text{TA}}$，$\boldsymbol{h}_{j,t-1}^{\text{PA}}$，$\boldsymbol{h}_{j,t-1}^{\text{CA}}$ 分别为 TDTA-GRU，TDPA-GRU，TDCA-GRU 对第 j 个图像序列在时间步长 $t-1$ 处的输出。

因此，输入 GRU_{lan} 的第 j 个图像序列的视觉、规划和概念注意力分布分别如式(7)－式(9)所示。

$$\hat{\boldsymbol{v}}_{j,t} = \sum_{j=1}^{J} \text{softmax}(\boldsymbol{W}_{hv}^{\text{T}} \boldsymbol{v}_{j,t}^*) \boldsymbol{v}_{j,t} \tag{7}$$

$$\hat{\boldsymbol{p}}_{j,t} = \sum_{j=1}^{J} \text{softmax}(\boldsymbol{W}_{hp}^{\text{T}} \boldsymbol{p}_{j,t}^*) \boldsymbol{p}_{j,t} \tag{8}$$

$$\hat{\boldsymbol{c}}_{j,t} = \sum_{j=1}^{J} \text{softmax}(\boldsymbol{W}_{hc}^{\text{T}} \boldsymbol{c}_{j,t}^*) \boldsymbol{c}_{j,t} \tag{9}$$

其中，$\boldsymbol{W}_{hv}^{\text{T}}$，$\boldsymbol{W}_{hp}^{\text{T}}$，$\boldsymbol{W}_{hc}^{\text{T}}$ 为权重参数；$\hat{\boldsymbol{v}}_{j,t}$，$\hat{\boldsymbol{p}}_{j,t}$，$\hat{\boldsymbol{c}}_{j,t}$ 分别为第 j 个图像序列在时间步长 t 的视觉、规划和概念词的注意力分布。

连贯语言生成器：GRU_{lan} 负责在信息解码器处理的视觉、规划和概念信息的引导下生成连贯的故事。GRU_{lan} 生成词序列，在每个时间步更新 GRU_{lan} 的隐藏状态，如式(10)所示。整个故事词汇的概率分布由 $p(y_{j,t})$ 表示，如式(11)所示。

$$\boldsymbol{h}_{j,t}^{\text{lan}} = GRU_{\text{lan}}(\boldsymbol{h}_{j,t-1}^{\text{lan}}, \hat{\boldsymbol{v}}_{j,t}, \hat{\boldsymbol{p}}_{j,t}, \hat{\boldsymbol{c}}_{j,t}, \boldsymbol{h}_{j,t}^{\text{VA}}, \boldsymbol{h}_{j,t}^{\text{PA}}, \boldsymbol{h}_{j,t}^{\text{CA}}) \tag{10}$$

$$p(y_{j,t}) = \text{softmax}(\tanh(\boldsymbol{W}_p \boldsymbol{h}_{j,t}^{\text{lan}} + \boldsymbol{b}_p)) \tag{11}$$

其中，$\boldsymbol{W}_p$ 和 $\boldsymbol{b}_p$ 为学习权值和偏置。

PLVS 模型的优化目标是最小化以视觉、规划、概念词为条件的交叉熵损失，如式(12)所示。

$$L(\psi) = \sum_{j=1}^{J} \sum_{m=1}^{M} \sum_{n=1}^{N_M} \log(p_\psi(\boldsymbol{W}_{j,m,n}^* | \boldsymbol{W}_{j,m,1:n-1}^*)) \tag{12}$$

其中，$\boldsymbol{W}_{j,m,n}^*$ 表示词典中的单词，$\boldsymbol{y}_{j,m}^* = \langle \boldsymbol{W}_{j,m,1}^*, \boldsymbol{W}_{j,m,2}^*, \cdots, \boldsymbol{W}_{j,m,n}^* \rangle$，$\psi$ 是 GRU_{lan} 的参数。

4 实验与分析

4.1 数据集和实验环境

本次实验选择 Ubuntu 18.04 64 位操作系统作为开发环境，并采用 PyTorch 2.0.1 框架进行模型的训练与测试。硬件配置包括 Intel i9-9900k 处理器和 NVIDIA GeForce RTX 4090 显卡(24 GB)。实验使用了公开的 VIST 数据集[6]来验证算法模型的性能。因为有些图片已经损坏或是版权失效，最后筛选出 10033 个相册和 80245 张照片，共涉及 12977 个词汇。

实验分为训练、验证和测试 3 个阶段，分别使用了 39659 个、4988 个和 5050 个样本。本模型的批次大小设置为 128，编码器、解码器的失分率分别为 0.2 和 0.5。规划和词嵌入的维数均设为 512，图像编码维度为 2048，注意力层配置为 2 层 4 头，学习率为 0.0004。在本次实验中，训练的 EPOCHS 值设定为 100，进行 100 次迭代(Epoch)。每轮迭代平均耗时约 16 min，因此整个训练过程大约需要 26 h 40 min(不包括可

能的暂停和中断时间）。此外，每次测试平均耗时 18 min。通过上述实验设置，能够全面评估算法模型在视觉故事生成任务中的性能表现。

4.2 评价指标

采用多种量化评估指标来评估提出的算法模型在视觉故事生成领域的效果，包括 BLEU-N（Bilingual Evaluation Understudy）[15]，METEOR[16]（Metric for Evaluation of Translation with Explicit ORdering），ROUGE-L[17]（Recall-Oriented Understudy for Gisting Evaluation）和 CIDEr[18]（Consensus-based Image Description Evaluation）。BLEU-N 和 METEOR 是衡量文本生成质量的指标。BLEU-N 侧重于评估生成文本的准确性和流畅度，通过比较生成故事中的 n 元词组与参考文本的匹配程度来评估。而 METEOR 会考虑单词的变形，并通过 WordNet 等知识库扩展同义词集，允许更多变化。ROUGE-L 与 BLEU 类似，但侧重于召回率，即生成文本包含多少参考文本的信息，而 BLEU 则更注重准确率。CIDEr 针对图像描述生成任务，主要评估是否能捕捉到图像中的关键信息。

此外，实验还采用多样性指标 Distinct-N[19] 和 TTR[20] 来评估生成故事的多样性。Distinct-N（其中 N 代表 n-gram）通过计算文本中不同 n-gram 的数量与总 n-gram 数量的比率，来评估语言的多样性。TTR 通过计算文本中不同单词与总单词数量的比率，来衡量语言的多样性和丰富度。上述客观评估指标能够全面评估提出的算法模型在视觉故事生成任务中的表现，并从多个角度评价模型生成文本的质量。

4.3 与主流方法对比

将所提模型与几个主流模型进行了比较，包括 AREL，CKACS，KE-VIST，PR-VIST，COVS，VP-BART 和 SCO-VIST。

1）AREL[21] 模型：用端到端的训练方式进行训练，并通过强化学习的方法优化模型在通用评价指标上的表现。

2）CKACS[10] 模型：为每个图像序列创建常识知识图，考虑候选概念间的相关性和图像-概念的相关性，用 Bart 生成故事。

3）KE-VIST[2] 模型：通过概念选择模块预测与输入图像序列相关的概念述语，并结合外部知识丰富概念信息，最后将其输入 Transformer 生成故事。

4）PR-VIST[11] 模型：借鉴人类写作方法，在丰富概念集基础上设计故事脉络线策划模型，用于策划概念述语的顺序和组合关系，以提升故事连贯性。

5）COVS[4] 模型：利用主题模型挖掘并学习相册的主题分布，确保生成的叙述与图像主题一致。结合 PBS 算法，避免重复，促进故事多样性。

6）VP-BART[5] 模型：结合预训练语言模型和规划的视觉叙事框架，通过图像序列生成连贯、有趣且自然的叙事。

7）SCO-VIST[22] 模型：将社会互动常识融入故事，通过图结构优化，建立标题、常识、主题的异构图，捕捉并发展故事情节，生成视觉故事。

比较结果如表 3 所列。

表 3 模型对比分析

Table 3 Model comparison analysis

模型	*BLEU-1*	*BLEU-2*	*ROUGE_L*	*CIDEr*	*METEOR*	*Distinct-3*	*Distinct-4*	*TTR*
AREL	0.6041	0.2483	0.2782	0.0641	0.3393	0.5719	0.7020	0.6056
CKACS	—	—	0.2887	0.0806	0.3501	0.5789	0.7111	**0.6090**
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
VP-BART	—	—	—	0.0550	0.3360	—	—	—
SCO-VIST	—	0.347	0.221	0.0591	0.2750	—	—	—
PLVS(ours)	**0.6568**	**0.3860**	**0.3053**	0.0845	0.3456	**0.5880**	**0.7277**	0.6074

由表 3 可得出，在 VIST 测试集上，PLVS 模型相对其他主流方法在 5 个评价指标上表现较好。与 COVS 模型相比，PLVS 在 BLEU-1，BLEU-2 和 ROUGE_L 指标上提升了 1.58 个百分点、2.69 个百分点和 0.41 个百分点，显示出在故事准确性和连贯性方面的优化；在 Distinct-3，Distinct-4 和 TTR 指标上分别提升 2.22 个百分点、3.61 个百分点和 5.60 个百分点，显示出在丰富故事内容和增强表达多样性方面的优势。

与 CKACS 模型相比，PLVS 在 Distinct-3 和 Distinct-4 指标上分别提升了 0.91 个百分点和 1.66 个百分点，展现出 PLVS 在生成多样化和丰富表达方面的优势。PLVS 尽管在 TTR 指标上略低于 CKACS，但在 BLEU，ROUGE_L 和 CIDEr 指标上更优，表明其在保持文本多样性的同时能保持故事流畅。此外，PLVS 在各项指标上均优于 AREL，KE-VIST 和 PR-VIST 模型。以上结果表明，结合概念和规划模块有助于捕捉并发展故事情节，生成冗余度较低和细节丰富的故事。

4.4 消融实验

为验证概念和规划模块信息对模型生成的有用性，在不改变基线模型的前提下，设计对比实验，分别是：仅输入图片，通过图像特征提取模块和仅处理视觉信息的故事生成模块生成故事（PLVS-PC）；输入图像与概念述语，通过图像特征提取模块、概念模块、处理视觉和概念述语的故事生成模块生成故事（PLVS-P）；输入图像与规划信息，通过图像特征提取模块、规划模块和处理视觉和规划信息的故事生成模块生成故事（PLVS-C）；输入视觉和概念术语以及规划的答案，使用所有模块生成故事（PLVS(plan-answer)）；输入图像和概念术语以及规划信息，使用所有模块生成故事（PLVS）。

实验结果如表 4 所列。PLVS-P 在 BLEU，ROUGE_L，CIDEr 和 METEOR 等指标上优于 PLVS-PC，说明引入概念选择模块有助于扩充词汇内容，增加词汇选项空间，提升故事连贯性和多样性，进而生成内容丰富的故事。PLVS-C 在 BLEU-1，ROUGE_L，Distinct，TTR 指标上表现次优，且在

CIDEr 指标上高于 PLVS，表明引入规划学习方法能发掘细节内容、系统组织情节，提高故事质量。此外，PLVS 在各指标上均优于 PLVS(plan-answer)，表明引入问答策略可帮助模型发展故事线索，保持逻辑性和可读性。以上结果表明，结合概念和规划模块有助于生成语言流畅、冗余度较低和细节描述丰富的故事。

表 4 消融实验

Table 4 Ablation study

方法	*BLEU-1*	*BLEU-2*	*ROUGE_L*	*CIDEr*	*METEOR*	*Distinct-3*	*Distinct-4*	*TTR*
PLVS-PC	0.6060	0.2572	0.2793	0.0715	0.3326	0.5718	0.7013	0.5852
PLVS-P	0.6237	0.3839	0.2919	0.0700	0.3419	0.5820	0.7137	0.6060
PLVS-C	0.6412	**0.3951**	0.3017	0.0820	**0.3563**	0.5726	0.7028	0.5708
PLVS(plan-answer)	0.6048	0.3736	0.2985	0.0596	0.3450	0.5401	0.6487	0.3843
PLVS	**0.6568**	0.3860	0.3053	**0.0845**	0.3456	**0.5880**	**0.7277**	**0.6074**

4.5 可视化故事生成示例

本节将从两个方面分析 PLVS 的实现效果，并与上述评估实验中表现较好的模型进行对比分析。表 5 中 COVS[4] 生成的“friends and friends”的描述，影响了故事的流畅性和自然度，降低了叙述的质量。相比之下，本文提出的规划方法能较好地规避此类冗余情况。PLVS 在描述中能够贴近图像内容，例如“speakers”和“together for a group photo”等表达，与视觉内容相对应，既能反映图像信息，又填充了图像细节，使故事更具体、生动，通过全面的规划和概念选择，有效地提升了故事质量与图像的匹配度。

表 5 细节丰富对比

Table 5 Detailed comparison

图像序列	
COVS	I had a great time at the ceremony last weekend. [female] was very happy to be there. i had a great time at the conference. a group of **friends and friends**. we took a lot of pictures.
PLVS-C	the man was excited to see his friends. everyone was very happy to be there. [male] was aspeech to the audience. the whole family was there to celebrate. the whole family was there to celebrate
PLVS	this is a picture of a group. the woman is very happy to be there. we had a greatspeakers. the women of the community members gathered together for a group photo. the group of friends posed for a picture together.

当处理包含多张重复图像的场景时，COVS 倾向于在文本中多次生成重复内容，如表 6 中“had a great time”短语出现了 5 次之多，显得故事单调，缺乏变化。而在引入规划模块后，同样的场景下，“had a great time”出现的次数变少，尤其是 PLVS 中只出现两次且不在同一个句子中，减少了内容的重复。以上表明，引入概念和规划模块后，PLVS 能够较合理地分配和使用词汇资源，避免出现文本过度重复的现象，减少冗余信息，从而生成较多样化的故事。

表 6 冗余对比

Table 6 Redundancy comparison

图像序列	
COVS	we **had a great time** at the park last night. the sun was setting. it was a beautiful day. we **had a great time** and **had a great time**. we **had a great time** and **had a great time**.
PLVS-C	[male] and [female] were **having a great time** at the beach. the sun was setting. it was a lot of fun. we **had a great time**. we **had a great time**.
PLVS	we went to the park last night. the sunset was beautiful. it was a great day for the pool. the family **has a great time**. we **all had a great time**.

结束语 本次工作中引入了一种新的视觉故事生成方法：利用 VQA 的预训练语言模型引导生成规划后，与视觉特征、概念术语一起输入解码器生成故事。规划模块设定 6 个维度的问题，并通过 VQA 生成答案对作为中间规划步骤，有效筛选关键概念，指导故事构建。实验证明，该模型在降低冗余度、捕捉图片细节信息、丰富内容上的表现与预期相符。所提模型虽然细节有待完善，但有望实现个性化故事叙述，使生成的故事更加生动丰富。该方法不仅助力模型深入理解图像内容，还能融合规划生成连贯有趣的句子，提升故事的连贯性。

参 考 文 献

[1] WANG R, WEI Z, LI P, et al. Story telling from an Image Stream Using Scene Graphs[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 9185-9192.

[2] HSU C C, CHEN Z Y, HSU C Y, et al. Knowledge-Enriched Visual Storytelling[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 7952-7960.

[3] LI M M, JIANG A W, LONG Y Z, et al. Visual story generation

algorithm based on fine-grained visual features and knowledge graph[J]. Journal of Chinese Information Technology, 2022, 36(9):139-148.

[4] GU J, WANG H, FAN R. Coherent Visual Storytelling via Parallel Top-Down Visual and Topic Attention [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(1):257-268.

[5] LIU D, LAPATA M, KELLER F. Visual Storytelling with Question-Answer Plans[M]// Findings of the Association for Computational Linguistics: EMNLP 2023. ACL, 2023: 5800-5813.

[6] CHENG S, GUO Z, 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(CVPR). 2024: 14291-14302.

[7] HUANG T H, FERRARO F, MOSTAFAZAD EH 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. 2016: 1233-1239.

[8] KIM T, HEO M O, SON S, et al. GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation [J]. arXiv: 1805. 10973, 2018.

[9] WANG J, FU J, TANG J, et al. Show, Reward and Tell: Automatic Generation of Narrative Paragraph From Photo Stream by Adversarial Training[C]// AAAI Conference on Artificial Intelligence. 2018: 7396-74003.

[10] CHEN H, HUANG Y, TAKAMURA H, et al. Commonsense knowledge aware concept selection for diverse and informative visual storytelling[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 999-1008.

[11] HSU C Y, CHU Y W, HUANG T H K, et al. Plot and Rework: Modeling Storylines for Visual Storytelling[C]// Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and 11th International Joint Conference on Natural Language Processing. 2021: 4443-4453.

[12] NARAYAN S, MAYNEZ J, AMPLAYO R K, et al. Conditional generation with a Question-Answering Blueprint[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 974-996.

[13] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149

[14] 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 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 26753-26763.

[15] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002: 311-318.

[16] 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. 2005: 65-72.

[17] 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 of the Association for Computational Linguistics. 2004: 21-26.

[18] VEDANTAM R, ZITNICK C L, PARIKH D. C-IDEr: Consensus based image description evaluation [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 4566-4575.

[19] LI J, GALLEY M, BROCKETT C, et al. A Diversity-Promoting objective function for neural conversation conversation Models [C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 110-119.

[20] CONNORS L H, LIM A, PROKAEVAT, et al. Tabulation of human transthyretin(TTR) variants[J]. Amyloid, 2003, 10(3): 160-84.

[21] WANG X, CHEN W, 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. ACL, 2018: 899-909

[22] WANG E, HAN S C, POON J. SCO-VIST: Social Interaction Commonsense Knowledge-based Visual Storytelling[C]// Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. 2024: 1602-1616.

WANG Yuanlong, born in 1982, Ph.D, associate professor, is a member of CCF (No. 48432M). His main research interests include natural language processing and graphics image processing.

(责任编辑:柯颖)