基于全局与序列混合变分Transformer的多样化图像描述生成方法

刘 兵[1,2],李 穗[1,2],刘明明[1*],刘 浩[1,2]
(1. 中国矿业大学计算机科学与技术学院,江苏徐州 221116;2. 矿山数字化教育部工程研究中心,江苏徐州 221116)

摘 要: 多样化图像描述生成已成为图像描述领域研究热点. 然而,现有方法忽视了全局和序列隐向量之间的依赖关系,严重限制了图像描述性能的提升. 针对该问题,本文提出了基于混合变分Transformer的多样化图像描述生成框架. 具体地,首先构建全局与序列混合条件变分自编码模型,解决全局与序列隐向量之间依赖关系表示的问题. 其次,通过最大化条件似然推导混合模型的变分证据下界,解决多样化图像描述目标函数设计问题. 最后,无缝融合Transformer和混合变分自编码模型,通过联合优化提升多样化图像描述的泛化性能. 在MSCOCO数据集上实验结果表明,与当前最优基准方法相比,在随机生成20和100个描述语句时,多样性指标m-BLEU(mutual overlap-BiLingual Evaluation Understudy)分别提升了4.2%和4.7%,同时准确性指标CIDEr(Consensus-based Image Description Evaluation)分别提升了4.4%和15.2%.

关键词: 图像理解;图像描述;变分自编码;隐嵌入;多模态学习;生成模型
基金项目: 国家自然科学基金(No.62276266,No.61801198)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)04-1305-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20231155

Diverse Image Captioning Based on Hybrid Global and Sequential Variational Transformer

LIU Bing[1,2], LI Sui[1,2], LIU Ming-ming[1*], LIU Hao[1,2]
(1. *School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*;
2. *Ministry of Education Engineering Research Center of Mine Digitization, Xuzhou, Jiangsu* 221116, *China*)

Abstract: Diverse image captioning has become a research hotspot in the field of image description. Existing methods generally ignore the dependency relationship between global and sequential latent vectors, which seriously limits the performance improvement. To address this problem, this paper proposes a hybrid variational Transformer based diverse image captioning framework. Firstly, we construct a hybrid conditional variational autoencoder to effectively model the dependency between global and sequential latent vectors. Secondly, the evidence lower bound is derived by maximizing the conditional likelihood of the hybrid autoencoder, which serves as the objective function for diverse image captioning. Finally, we seamlessly combine the Transformer model with the hybrid conditional variational autoencoder, which can be jointly optimized to improve the generalization performance of diverse image captioning. The experimental results on MSCOCO dataset show that compared with the state-of-the-art methods, when randomly generating 20 and 100 captions, the diversity metric m-BLEU (Mutual overlap Bilingual Evaluation Under study) has improved by 4.2% and 4.7%, respectively, while the accuracy metric CIDEr (Consensus based Image Description Evaluation) has improved by 4.4% and 15.2%, respectively.

Key words: image understanding; image captioning; variational autoencoding; latent embedding; multi-modal learning; generative model
Foundation Item(s): National Natural Science Foundation of China (No.62276266, No.61801198)

收稿日期:2023-12-12;修回日期:2024-02-26;责任编辑:覃怀银
*通讯作者:刘明明

1 引言

图像描述生成，是一项具有挑战性的条件生成任务，旨在生成语法正确且与图像语义匹配的描述语句，是图像理解领域的基础性热点研究课题[1]. 随着近年来深度学习技术的兴起，受神经机器翻译启发的编解码框架成为图像描述生成的主流方法，并广泛应用于图像检索和分类[2]、视觉辅助[3]和智慧医疗[4]等领域.

传统的图像描述生成模型侧重于提升描述语句的准确性指标. 例如，李志欣等[5]提出结合视觉特征和场景语义的图像描述生成方法，通过主题词指导单词的准确生成. 周东明等[6]提出基于强化学习的多层级视觉融合网络模型，通过将视觉特征转化为视觉知识的特征集，从而生成更加流畅的描述语句. 刘茂福等[7]利用视觉关联与上下文双注意力机制，指导生成准确的图像描述文本. 宋井宽等[8]通过视觉区域聚合与双向协作学习，以促进模型生成更加细粒度的图像描述文本. 尽管这些模型有效提升了图像描述的准确性，但模型大多局限于学习图像到文本域的确定性映射，仅关注如何提高平均描述的准确性，仍未从根本上解决确定性映射导致的模式坍塌问题，导致无法生成多样化的描述语句.

为了解决模式坍塌问题，最近一些研究者开始探索多样化的图像描述生成方法. Dai 等人[9]首次提出了一种基于条件生成对抗网络的多样化图像描述框架. Shetty 等人[10]进一步将对抗样本与近似耿贝尔采样[11]相结合，以生成更自然的描述. 然而，这类方法生成的描述与真实描述之间差异性较大，精确性指标较差. 为了兼顾图像描述的准确性和多样性，Wang 等人[12]将条件变分自编码引入图像描述生成. 在此基础上，Aneja 等人[13]提出了基于序列化隐空间的条件变分自编码方法，提升了细粒度描述生成的能力. Mahajan 等人[14]对描述文本的上下文和目标单词进行建模，进一步增强了描述的多样性. Wang 等人[15]结合检索奖励，提出了一种兼顾准确性和多样性的评价指标. 尽管这些模型缓解了准确性指标下降问题，但其忽视了全局和序列隐向量之间的依赖关系，严重限制了隐向量空间的表示能力和描述性能的提升. 此外，这些模型大多基于传统的长短时期记忆网络（Long Short Term Memory，LSTM）构建，导致不能充分利用图像和文本的全局信息，也无法提供并行训练支撑.

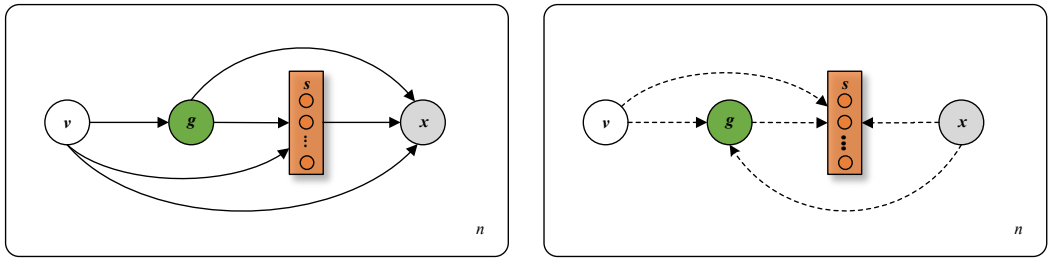
针对上述存在的问题，本文首先提出一种新的混合条件变分自编码模型，通过建模全局和序列隐向量之间的依赖关系，进一步扩展了隐空间表示能力，并推导了该模型的变分证据下界. 然后，以变分证据下界为优化目标，构建基于 Transformer 的多样化图像描述生成框架，通过同时建模句子级和单词级的多样性，克服当前模型隐向量空间表示能力不足的局限性，进一步提高多样化图像描述泛化性能. 在 MSCOCO 标准数据集上，进行了定量和定性实验对比分析. 结果表明，相比现有方法，本文方法能够同时提升图像描述的准确性和多样性.

2 混合条件变分自编码模型

2.1 条件生成模型

为增强图像描述的多样性表征能力，引入句子级全局隐向量 $\boldsymbol{g}$ 建模语法结构的多样性. 同时，通过单词级序列隐向量表征每个时间步单词的多样性. 假设 $\boldsymbol{v}$ 表示图像视觉特征，$\boldsymbol{x}=(x_1,x_2,\cdots,x_m)$ 表示对应的描述，n 表示描述语句个数，$\boldsymbol{\theta}$ 为模型参数. 给定句子级全局隐向量 $\boldsymbol{g}$ 和单词级序列隐向量 $\boldsymbol{s}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_m)$，其中 m 表示描述语句中单词个数，所构建的混合条件变分自编码生成模型如图 1 所示，联合概率分布表示如下：

$$\begin{aligned}p_\theta(\boldsymbol{x},\boldsymbol{g},\boldsymbol{s}\,|\,\boldsymbol{v})&=p_\theta(\boldsymbol{g},\boldsymbol{s}\,|\,\boldsymbol{v})p_\theta(\boldsymbol{x}\,|\,\boldsymbol{v},\boldsymbol{g},\boldsymbol{s})\\&=p_\theta(\boldsymbol{g}\,|\,\boldsymbol{v})p_\theta(\boldsymbol{s}\,|\,\boldsymbol{v},\boldsymbol{g})p_\theta(\boldsymbol{x}\,|\,\boldsymbol{v},\boldsymbol{g},\boldsymbol{s})\end{aligned}\tag{1}$$



(a) 生成模型 $p_\theta(\boldsymbol{g}\,|\,\boldsymbol{v})p_\theta(\boldsymbol{s}\,|\,\boldsymbol{v},\boldsymbol{g})$　　(b) 变分近似后验模型 $q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x})$

图 1 混合条件变分自编码模型

如图 1 所示，生成模型 $p_\theta(\boldsymbol{g}\,|\,\boldsymbol{v})p_\theta(\boldsymbol{s}\,|\,\boldsymbol{v},\boldsymbol{g})$ 用于建模描述语句的生成过程，用实线表示. 变分近似后验模型 $q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x})$ 用于近似真实的后验概率分布，用虚线表示. $\boldsymbol{x}$ 的随机生成过程包括三个步骤：(1) 首先从先验条

件分布$p_\theta(\boldsymbol{g}|\boldsymbol{v})$中采样一个句子级全局隐向量$\boldsymbol{g}$;(2)然后从先验条件分布$p_\theta(\boldsymbol{s}|\boldsymbol{v},\boldsymbol{g})$中采样单词级序列隐向量$\boldsymbol{s}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_m)$;(3)最后从条件分布$p_\theta(\boldsymbol{x}|\boldsymbol{v},\boldsymbol{g},\boldsymbol{s})$中生成一个描述语句.

2.2　变分证据下界

图1生成模型条件似然的变分表示形式为

$$\begin{aligned}\log p_\theta(\boldsymbol{x}|\boldsymbol{v})=&D_{\mathrm{kl}}\left(q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x})\|p_\theta(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x})\right)\\&+L\left(q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x});\boldsymbol{x}\right)\end{aligned}\tag{2}$$

其中,$q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x})$表示近似后验概率分布.

由于$D_{\mathrm{kl}}\left(q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x})\|p_\theta(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x})\right)\geqslant 0$,则

$$\log p_\theta(\boldsymbol{x}|\boldsymbol{v})\geqslant L\left(q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x});\boldsymbol{x}\right)\tag{3}$$

其中,$L\left(q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x});\boldsymbol{x}\right)$称为对数条件似然$p_\theta(\boldsymbol{x}|\boldsymbol{v})$的变分证据下界.根据图1所示生成模型,$q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x})=q_{\varphi_1}(\boldsymbol{g}|\boldsymbol{v},\boldsymbol{x})q_{\varphi_2}(\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x},\boldsymbol{g})$,则变分证据下界可等价表示为

$$\begin{aligned}L\left(q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x});\boldsymbol{x}\right)=&E_{q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{v},\boldsymbol{s},\boldsymbol{g})\right]\\&-D_{\mathrm{kl}}\left(q_{\varphi_1}(\boldsymbol{g}|\boldsymbol{v},\boldsymbol{x})\|p_\theta(\boldsymbol{g}|\boldsymbol{v})\right)\\&-D_{\mathrm{kl}}\left(q_{\varphi_2}(\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x},\boldsymbol{g})\|p_\theta(\boldsymbol{s}|\boldsymbol{v},\boldsymbol{g})\right)\end{aligned}\tag{4}$$

使用乘积规则和时间因子分解,式(4)可等价表示为

$$\begin{aligned}L\left(q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x});\boldsymbol{x}\right)=&E_{q_\varphi(\boldsymbol{g},\boldsymbol{s}|\boldsymbol{v},\boldsymbol{x})}\left[\sum_t\log p_\theta\left(x_t|\boldsymbol{x}_{<t},\boldsymbol{v},\boldsymbol{s}_{\leqslant t},\boldsymbol{g}\right)\right]\\&-D_{\mathrm{kl}}\left(q_{\varphi_1}(\boldsymbol{g}|\boldsymbol{v},\boldsymbol{x})\|p_\theta(\boldsymbol{g}|\boldsymbol{v})\right)\\&-\sum_t D_{\mathrm{kl}}\left(q_{\varphi_2}(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x},\boldsymbol{v},\boldsymbol{g})\|\right.\\&\left.p_\theta(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x}_{<t},\boldsymbol{v},\boldsymbol{g})\right)\end{aligned}\tag{5}$$

其中,式(5)中第一项表示用于生成单词序列的对数似然,第二项表示句子级全局隐向量$\boldsymbol{g}$的后验分布$q_{\varphi_1}(\boldsymbol{g}|\boldsymbol{v},\boldsymbol{x})$和条件先验$p_\theta(\boldsymbol{g}|\boldsymbol{v})$之间的KL(Kullback-Leibler)散度,第三项表示每个时间步单词对应的序列隐向量$\boldsymbol{s}_t$的后验$q_{\varphi_2}(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x},\boldsymbol{v},\boldsymbol{g})$和条件先验$p_\theta(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x}_{<t},\boldsymbol{v},\boldsymbol{g})$之间的KL散度之和.

3　基于变分证据下界优化的图像描述生成

将变分证据下界作为优化目标,其本质在于最大化描述语句的条件概率分布,从而增强描述语句的多样性.在此基础上,提出一种混合变分自编码多样化图像描述生成框架(Hybrid Conditional Variational Autoencoder and Transformer based Image Captioning Framework,HCVA-T-ICF).如图2所示,该框架包括三个模块:全局条件变分编码网络(Global Conditional Variational Encoder Network,G-CVEN)、序列条件变分编码网络(Sequential Conditional Variational Encoder Network,S-CVEN)和图像描述解码网络.

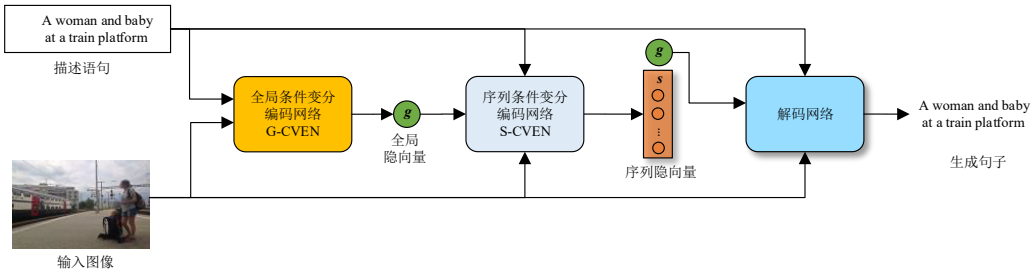


图2　基于混合条件变分自编码的图像描述生成架构HCVA-T-ICF

3.1　基于Transformer的条件全局变分编码网络

如图3所示,首先面向句子级全局隐向量,构建条件全局变分编码网络G-CVEN,其中后验推断分支网络$q_{\varphi_1}(\boldsymbol{g}|\boldsymbol{v},\boldsymbol{x})$和先验分支网络$p_\theta(\boldsymbol{g}|\boldsymbol{v})$组成了双分支的编码网络.对于输入图像,首先采用预训练的Swin-T(Transformer using Shifted Windows)模型[16]提取图像特征$\boldsymbol{v}$,然后输入到由l个注意力块组成的编码器中得到视觉特征$\boldsymbol{v}_l$.针对描述$\boldsymbol{x}=(x_1,x_2,\cdots,x_m)$,首先通过词嵌入和位置编码将其转换为$\boldsymbol{w}\in\mathbf{R}^d$,随后输入至由$l$个注意力块组成的编码器中得到文本特征$\boldsymbol{w}_l$.$\boldsymbol{v}_l$和$\boldsymbol{w}_l$分别表示如下:

$$\begin{aligned}\bar{\boldsymbol{v}}&=\mathrm{AN}\left(\mathrm{MHSA}\left(\boldsymbol{v}_{l-1},\boldsymbol{v}_{l-1},\boldsymbol{v}_{l-1}\right)\right)\\\bar{\boldsymbol{w}}&=\mathrm{AN}\left(\mathrm{MHSA}\left(\boldsymbol{w}_{l-1},\boldsymbol{w}_{l-1},\boldsymbol{w}_{l-1}\right)\right)\end{aligned}\tag{6}$$

$$\begin{aligned}\boldsymbol{v}_l&=\mathrm{AN}\left(\mathrm{FFN}\left(\bar{\boldsymbol{v}}\right)\right)\\\boldsymbol{w}_l&=\mathrm{AN}\left(\mathrm{FFN}\left(\bar{\boldsymbol{w}}\right)\right)\end{aligned}\tag{7}$$

其中,MHSA表示多头自注意力(Multi-Head Self-Attention,MHSA)模块,AN表示相加归一化(Add&Normalization,AN)模块,FFN表示前馈网络(Feed Forward Network,FFN).

为了抽取图像和描述的全局特征表示，引入一个可学习向量作为查询向量，并通过交叉注意力（Cross Attention，CA）模块自适应地进行融合，具体如下式所示：

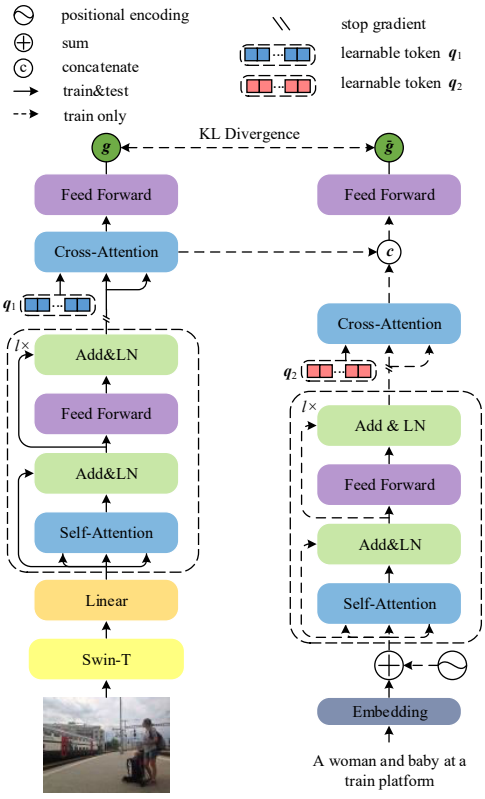


图3 基于Transformer的条件全局变分编码网络

$$\begin{aligned}\boldsymbol{v}_{l+1}&=\mathrm{CA}\left(\boldsymbol{q}_1,\boldsymbol{v}_l,\boldsymbol{v}_l\right)\\ \boldsymbol{w}_{l+1}&=\mathrm{CA}\left(\boldsymbol{q}_2,\boldsymbol{w}_l,\boldsymbol{w}_l\right)\end{aligned}\tag{8}$$

其中，$\boldsymbol{q}_1\in\mathbf{R}^d,\boldsymbol{q}_2\in\mathbf{R}^d,\boldsymbol{v}_{l+1}$和$\boldsymbol{w}_{l+1}$分别表示图像和描述的全局特征．两者拼接后经前馈层生成后验全局隐变量$\bar{\boldsymbol{g}},\boldsymbol{v}_{l+1}$输入前馈层生成全局隐变量$\boldsymbol{g}$．具体地，将$q_{\varphi_1}(\boldsymbol{g}|\boldsymbol{v},\boldsymbol{x})$建模为高斯分布$q_{\varphi_1}(\boldsymbol{g}|\boldsymbol{v},\boldsymbol{x})=N\left(\boldsymbol{g};\mu_1(\boldsymbol{x},\boldsymbol{v}),\sigma_1(\boldsymbol{x},\boldsymbol{v})\right)$，其中，$\mu_1(\boldsymbol{x},\boldsymbol{v})$和$\sigma_1(\boldsymbol{x},\boldsymbol{v})$分别表示均值和标准差．使用前馈网络将$\boldsymbol{v}_{l+1}$和$\boldsymbol{w}_{l+1}$映射为均值$\mu_2(\boldsymbol{x},\boldsymbol{v})$与标准差$\sigma_2(\boldsymbol{x},\boldsymbol{v})$，并通过重参数技巧$\bar{\boldsymbol{g}}=\mu_2(\boldsymbol{x},\boldsymbol{v})+\sigma_2(\boldsymbol{x},\boldsymbol{v})\odot\boldsymbol{\varepsilon}$采样得到$\bar{\boldsymbol{g}}$，其中，$\boldsymbol{\varepsilon}$服从标准多元正态分布．先验分支网络$p_\theta(\boldsymbol{g}|\boldsymbol{v})$使用类似的采样方法．

3.2 基于Transformer与混合变分自编码的多样化图像描述生成

序列变分网络模型架构如图4所示．编码器仍采用Swin-T[16]提取图像视觉特征$\boldsymbol{v}$，将其输入到编码器中得到视觉特征$\boldsymbol{v}_l'$．然后$\boldsymbol{v}_l'$和文本特征一起输入后验推断和先验近似网络，执行双路径的变分推断．

首先将单词嵌入后的向量进行位置编码得到初始输入向量．随后，将其与全局隐向量$\boldsymbol{g}$逐一相加得到融合向量$\boldsymbol{w}'$，输入多头自注意模块MHSA和AN层可得

$$\boldsymbol{w}_l'=\mathrm{AN}\left(\mathrm{MSA}\left(\boldsymbol{w}',\boldsymbol{w}',\boldsymbol{w}'\right)+\boldsymbol{w}'\right)\tag{9}$$

然后，通过多头交叉注意模块和残差归一化层，将语义特征$\boldsymbol{w}_l'$与视觉特征$\boldsymbol{v}_l'$进行融合，表示为

$$\boldsymbol{h}=\mathrm{AN}\left(\mathrm{CA}\left(\boldsymbol{w}_l',\boldsymbol{v}_l',\boldsymbol{v}_l'\right)+\boldsymbol{w}_l'\right)\tag{10}$$

后验概率$q_{\varphi_2}\left(\boldsymbol{s}_t|\boldsymbol{s}_{t-1},\boldsymbol{x},\boldsymbol{v},\boldsymbol{g}\right)$进行神经网络参数化服从均值为$\bar{\mu}_t\left(\boldsymbol{s}_{t-1},\boldsymbol{x},\boldsymbol{v},\boldsymbol{g}\right)$和标准差为$\bar{\sigma}_t\left(\boldsymbol{s}_{t-1},\boldsymbol{x},\boldsymbol{v},\boldsymbol{g}\right)$的多元高斯分布，表示为

$$q_{\varphi_2}\left(\boldsymbol{s}_t|\boldsymbol{s}_{t-1},\boldsymbol{x},\boldsymbol{v},\boldsymbol{g}\right)=N\left(\boldsymbol{s}_t;\bar{\mu}_t(\boldsymbol{x},\boldsymbol{v}),\bar{\sigma}_t(\boldsymbol{x},\boldsymbol{v})\right)\tag{11}$$

具体地，分别使用两个前馈网络将$\boldsymbol{h}$映射成均值与标准差向量，并通过重参数技巧采样得到$\boldsymbol{s}_t$．

类似于后验推断网络，先验推断网络将提取的语义特征$\boldsymbol{w}_l'$先后输入共享的多头交叉注意力模块、AN和FFN层，实现对先验概率的参数化：

$$p_\theta\left(\boldsymbol{s}_t|\boldsymbol{s}_{t-1},\boldsymbol{x}_{<t},\boldsymbol{v},\boldsymbol{g}\right)=N\left(\boldsymbol{s}_t;\mu_t(\boldsymbol{x},\boldsymbol{v}),\sigma_t(\boldsymbol{x},\boldsymbol{v})\right)\tag{12}$$

其中，先验序列隐变量服从均值为$\mu_t\left(\boldsymbol{s}_{t-1},\boldsymbol{x}_{<t},\boldsymbol{v},\boldsymbol{g}\right)$、标准差为$\sigma_t\left(\boldsymbol{s}_{t-1},\boldsymbol{x}_{<t},\boldsymbol{v},\boldsymbol{g}\right)$的多元高斯分布．

解码网络与Transformer解码结构类似，但全局隐向量$\boldsymbol{g}$和序列隐向量$\boldsymbol{s}$需要与词嵌入向量进行逐个拼接，并作为解码网络的输入．输入特征首先经过线性层降维，然后利用MSA和AN模块提取文本语义特征，并与图像视觉特征$\boldsymbol{v}_l'$一同输入CA模块，利用交叉注意力机制获得加权视觉特征．然后通过AN与FFN层融合图像与文本特征．最后通过线性层和Softmax函数预测词汇表中单词出现的概率．

3.3 模型训练与推断

已知图像视觉特征$\boldsymbol{v}$和对应的成对描述句子$\boldsymbol{x}=\left(x_1,x_2,\ldots,x_m\right)$，所提出的多样化图像描述生成模型的优化目标如下：

$$\begin{aligned}L\left(\boldsymbol{\theta},\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2;\boldsymbol{x}\right)=&-\sum_{t=1}^{m}\log p_\theta\left(x_t|\boldsymbol{x}_{<t},\boldsymbol{v},\boldsymbol{s}_{\leqslant t},\boldsymbol{g}\right)\\&+\alpha D_{\mathrm{kl}}\left(q_{\varphi_1}\left(\boldsymbol{g}|\boldsymbol{v},\boldsymbol{x}\right)\|p_\theta\left(\boldsymbol{g}|\boldsymbol{v}\right)\right)\\&+\beta\sum_{t=1}^{m}D_{\mathrm{kl}}\left(q_{\varphi_2}\left(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x},\boldsymbol{v},\boldsymbol{g}\right)\|\right.\\&\left.p_\theta\left(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x}_{<t},\boldsymbol{v},\boldsymbol{g}\right)\right)\end{aligned}\tag{13}$$

其中，α和β表示超参数，第一项表示交叉熵损失函数，第二项为条件全局变分编码网络中的先验和后验概率分布之间的KL散度，第三项为条件序列先验和后验概率分布之间的KL散度．

在训练阶段，后验分支网络作为教师网络指导先

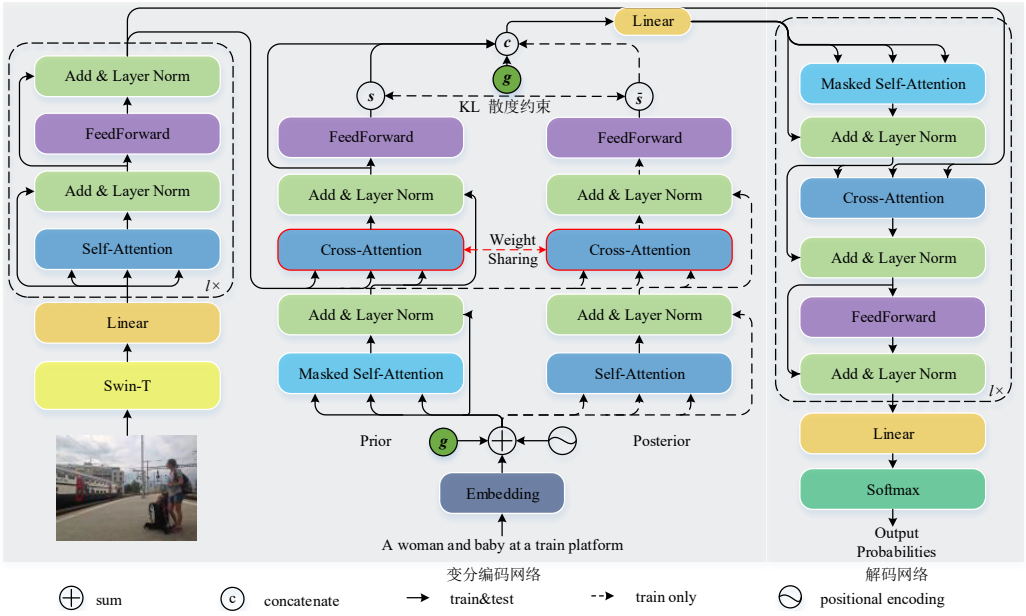


图4　融合全局与序列变分Transformer的多样化图像描述生成模型

验网络,从而实现先验隐变量与后验隐变量的对齐. 在测试阶段,从全局和序列先验分支中采样对齐后的全局与序列隐向量用于解码. 解码过程中,使用束搜索策略提升生成句子的准确性.

4　实验

4.1　数据集和评价标准

4.1.1　数据集

定量与定性实验使用标准MSCOCO数据集进行模型的训练与测试. 为了公平对比,采用多样化图像描述中常用的m-RNN(Multimodal Recurrent Neural Networks)数据集划分[1],其中训练集118 287张图像,验证集4 000张图像,测试集1 000张图像,且每张图像均有5条人工标注的描述语句.

4.1.2　准确性指标

实验采用了四种广泛使用的准确性指标,包括BLEU-N (BiLingual Evaluation Understudy for N-gram)[17]、METEOR(Metric for Evaluation of Translation with Explicit ORdering)[18]、ROUGE-L(Recall-Oriented Understudy for Gisting Evaluation: Longest common subsequence)[19]、CIDEr(Consensus-based Image Description Evaluation)[20]. 为公平对比,采用常用的Oracle重排序方法[13]计算最优准确性指标. Oracle重排使用测试图像的真实描述作为参考,在生成的一组描述中,选择以上准确性指标得分最高的描述作为最优描述,然后计算所有测试图像最优描述的准确性平均值.

4.1.3　多样性指标

目前图像多样化描述方法大多采用一致性重排序方法[13]计算多样性指标. 其中,对于一张生成了n个描述的测试图像,首先计算其与训练集中相似度最高的k个图像,然后将n个描述分别与k个相似图像的真实描述计算CIDEr得分,分值最高的描述被选为最优描述. 最终选取单张图片得分最高的5个描述,分别计算多样性指标:(1)Uniqueness:测试集所有图像生成的最优描述中,不重复的描述所占比例;(2)Novel:测试集生成的描述与训练集中真实描述不重复的描述个数;(3)m-BLEU(mutual overlap-BiLingual Evaluation Understudy):对于每一幅测试图像最优的5个候选语句,计算每个描述与其余四个描述的BLEU-4分数,取单张图像五个描述分数的平均后,再取测试集平均;(4)Div-1(1-gram Diversity):计算每一张测试图像最优的5个候选语句中不重复的1-gram在总1-gram长度中所占比例;(5)Div-2(2-gram Diversity):使用2-gram替换1-gram,计算方法同Div-1.

4.2　实验设置

模型训练时图像特征、单词嵌入和隐变量的维度均设置为512. 在视觉编码器中,使用预训练的Swin-T模型来提取每幅图像的1 536维视觉特征,并将其线性映射到512维向量. 视觉编码器和生成器均是由3层的注意力块组成的,其中多头注意力的头数设置为8. 平衡因子α和β分别设置为1和0.1. 训练阶段,设置批大小为10,首先利用Adam(Adaptive Moment Estimation)优化算法和学习率预热技巧来优化提出的模型,然后对全局与序列先验分布网络进行预训练. 最后,使用学习率5×10^{-6}对模型训练30个回合. 准确性和多样性评价时束搜索宽度分别设置为2和1. 本文的实验环境为

PyTorch 3.8.2和1个Nvidia GTX 3080 GPU.

4.3 实验结果定量分析

首先将本文方法与主流多样化图像描述方法进行对比. 这些方法包括Div-BS(Diverse Beam Search)[21]、PoS (Part-of-Speech)[22]、AG-CVAE (Conditional Variational Auto-Encoder with Additive Gaussian)[12]、Seq-CVAE (Sequential Conditional Variational Auto-Encoder)[13]、COS-CVAE (Context-Object Split Conditional Variational AutoEncoder)[14]、DCL-CVAE (Conditional Variational AutoEncoder with Dual Contrastive Learning)[23]和DivCon (Diverse Concepts)[24]. 表1列出了各方法在MSCOCO数据集上使用M-RNN划分和Oracle重排序后统计的准确性指标结果,其中,各指标的最优结果加粗表示,"↑"表示数值越大性能越好,"↓"反之.

为公平对比,实验设置与对比方法保持一致,分别利用先验分支网络采样20和100个隐变量,然后输入解码网络生成多样化的描述语句. 如表1所示,HCVA-T-ICF在两种采样下获得的各个准确性指标得分均优于其他方法. 在与人工评价相关性较好的CIDEr指标上,HCVA-T-ICF显著优于其他方法. 具体地,在采样20个全局与序列隐向量的情况下,相比于DCL-CVAE与DivCon,分别获得了16.5和4.4的CIDEr性能提升. 特别在采样100个隐向量的情况下,相比DCL-CVAE与DivCon,准确性分别提升了18.1和15.2.

表2进一步对比了不同模型使用一致性重排后获得的最优五个句子的多样性指标得分. 其中,HCVA-T-ICF模型在所有多样性指标中的综合性能更好. 在两种采样下,HCVA-T-ICF的Uniqueness指标分别达到了99.3%和98.4%. 在m-BLEU、Div-1和Div-2指标上,HCVA-T-ICF显著优于其他方法,在20和100个采样中分别获得0.72和0.60的Div-2指标得分,相比于最优的DCL-CVAE分别提升了10.7%和9%. 从而证实了HCVA-T-ICF模型能够有效提升图像描述的多样性. 此外,与DivCon相比,在mBleu指标上分别显著提升了4.2%和4.7%,这表明生成的描述之间有着显著的差异性. 此外,相比于COS-CVAE、DivCon和DCL-CVAE,HCVA-T-ICF无需进行数据增强、概念提示以及模型预训练工作,具有更好的适用性.

表1 MSCOCO数据集上的M-RNN划分和Oracle重新排序条件下各方法准确性对比

采样数量	方法	BLEU-4↑	CIDEr↑	ROUGE-L↑	METEOR↑
20	Div-BS[21]	38.3	140.5	65.3	35.7
	PoS[22]	44.9	146.8	67.8	36.5
	AG-CVAE[12]	47.1	130.8	63.8	30.9
	Seq-CVAE[13]	44.5	144.8	67.1	35.6
	COS-CVAE[14]	49.2	160.4	70.0	38.2
	DCL-CVAE[23]	45.9	150.2	67.8	35.8
	DivCon[24]	47.9	162.3	70.0	38.7
	HCVA-T-ICF	**49.7**	**166.7**	**70.5**	**39.1**
100	Div-BS[21]	40.2	144.8	66.6	37.2
	PoS[22]	55.0	166.1	72.5	40.9
	AG-CVAE[12]	55.7	151.7	69.0	34.5
	Seq-CVAE[13]	57.5	169.5	73.3	41.0
	COS-CVAE[14]	59.7	181.2	74.8	43.6
	DCL-CVAE[23]	61.1	182.3	75.2	42.8
	DivCon[24]	60.2	185.2	75.5	44.6
	HCVA-T-ICF	**66.1**	**200.4**	**78.4**	**46.2**

表3列出了所提出的HCVA-T-ICF模型在不同采样下的消融实验结果. 其中,IC-G-CVEN(Image Captioning with Global Conditional Variational Encoder Network)和IC-S-CVEN(Image Captioning with Sequential Conditional Variational Encoder Network)分别表示仅使用句子级全局隐向量和单词级序列隐向量的图像描述生成模型. 在CIDEr准确性指标上IC-G-SVEN优于IC-S-CVEN,在多样性指标上均低于IC-S-CVEN. 这是由于IC-G-CVEN侧重于句子语法结构多样性的建模,而IC-S-CVEN关注单词级多样性的表征,因此IC-S-CVEN倾向于生成更加多样的语句. 然而,IC-S-CVEN在提升多样性的同时无法兼顾准确性. 相比于IC-G-CVEN和IC-S-CVEN,HCVA-T-ICF模型融合了两种模型的优点,在绝大部分实验指标上均获得了更好的性能,能够同时提升图像描述的准确性和多样性.

表4给出了采样20个句子条件下,不同α和β设

表2　MSCOCO数据集上使用M-RNN划分和一致性重排序后多样性的性能对比　　单位:%

采样数量	方法	Uniqueness ↑	Novel ↑	m-BLEU↓	Div-1 ↑	Div-2 ↑
20	Div-BS[21]	**100**	3 106	0.81	0.20	0.26
	PoS[22]	96.3	3 394	0.64	0.24	0.35
	AG-CVAE[12]	69.8	3 189	0.66	0.24	0.34
	Seq-CVAE[13]	94.0	4 266	0.52	0.25	0.54
	COS-CVAE[14]	96.0	4 249	0.52	0.33	0.52
	DCL-CVAE[23]	97.9	**4 899**	0.54	0.39	0.65
	DivCon[24]	98.7	4 426	0.48	0.38	0.62
	HCVA-T-ICF	99.3	4 590	**0.46**	**0.44**	**0.72**
100	Div-BS[21]	**100**	3 421	0.82	0.20	0.25
	PoS[22]	91.5	3 446	0.67	0.23	0.33
	AG-CVAE[12]	47.4	3 069	0.70	0.23	0.32
	Seq-CVAE[13]	84.2	4 215	0.64	0.33	0.48
	COS-CVAE[14]	95.9	4 411	0.67	0.34	0.50
	DCL-CVAE[23]	92.1	**4 607**	0.66	0.35	0.55
	DivCon[24]	97.2	3 982	0.64	0.33	0.47
	HCVA-T-ICF	98.4	4 001	**0.61**	**0.37**	**0.60**

置对所提出模型的多样性和准确性的影响. 可以看到,当α取1.2和β取0.1时,模型尽管取得了最优的BLEU-4、METEOR和CIDEr准确性指标,但是在Div-1和Div-2多样性指标上的得分最低;当α取1.0和β取0.05时,模型取得了最优的Div-1和Div-2指标,然而对应的准确性指标最低;当α取1.0和β取0.4时,模型训练不稳定;当β取0.1,同时α分别取0.8和0.6时,模型准确性呈现下降趋势,这是由于损失函数中全局先验和后验概率分布之间的KL散度权重变小,从而降低了模型对全局隐空间特征表示学习的性能;当α取1.0和β取0.1时,模型在多样性和准确性之间达到了一个较好的平衡点.

表3　MSCOCO数据集上的M-RNN划分下准确性和多样性消融实验

采样数量	方法	全局	序列	BLEU-4 ↑	CIDEr ↑	METEOR ↑	Div-1 ↑	Div-2 ↑
20	IC-G-CVEN	√	×	45.6	162.7	38.2	0.39	0.63
	IC-S-CVEN	×	√	48.8	161.1	38.8	**0.44**	**0.72**
	HCVA-T-ICF	√	√	**49.7**	**166.7**	**39.1**	**0.44**	0.71
100	IC-G-CVEN	√	×	59.3	192.6	44.9	0.33	0.52
	IC-S-CVEN	×	√	65.2	191.0	43.9	**0.37**	**0.61**
	HCVA-T-ICF	√	√	**66.1**	**200.4**	**46.2**	**0.37**	0.60

表4　MSCOCO数据集采样20个句子条件不同超参取值对性能的影响

α	β	BLEU-4 ↑	METEOR ↑	CIDEr ↑	Div-1 ↑	Div-2 ↑
1.0	0.05	46.2	36.6	151.4	0.51	0.78
1.0	0.1	49.7	39.1	166.7	0.44	0.72
1.0	0.2	49.5	39.0	164.4	0.37	0.60
1.0	0.4	—	—	—	—	—
1.2	0.1	52.4	41.2	171.4	0.35	0.58
0.8	0.1	48.2	38.3	162.4	0.46	0.74
0.6	0.1	48.3	38.0	160.2	0.49	0.77

注:“—”表示模型训练不稳定

4.4　实验结果定性分析

图5展示了在每个时间步生成的单词及解码网络最后一层的注意力权重热图. 图中红色越深的区域表示生成相应单词时对该区域的特征关注度越高. 从图中可以看出,除了非视觉单词(例如“a”),所提出的方法解码时可以关注到与单词语义最相关的图像区域,即视觉单词推断时,模型会自动为相关图像视觉特征分配较大的注意力权重,从而验证了注意力机制的有效性和权值的可解

释性. 图6进一步定性对比了各方法生成的描述语句质量. 直观地说,与其他方法相比,本方法生成的描述更加准确和多样. 如图6所示,HCVA-T-ICF可以准确识别出图像中的鸟的数量,而其他方法则生成了不准确的量词和错误的单词. 此外,对比方法倾向于生成高频短语,而提出的HCVA-T-ICF方法均可以生成更自然和多样的描述,例如,生成的描述中包含了形容词“brown and white”,以及不常见的单词“identical”和“wading across”等.

(*a*) 原始图像 (*b*) 生成“a” (*c*) 生成“man” (*d*) 生成“kicking” (*e*) 生成“a”

(*f*) 生成“soccer” (*g*) 生成“ball” (*h*) 生成“over” (*i*) 生成“a” (*j*) 生成“field”

图5 描述语句生成过程中的视觉注意力可视化示意图

(*a*) 测试图像1

Div-BS: - a black and white cat is sitting in a suitcase - a black and white cat siting on top of a piece of luggage - a cat is sitting in a suitcase	POS: - a cat is sitting in a suitcase on a bed - a cat sitting on top of a suitcase on a bed - a cat that is laying down on a suitcase	Seq-CVAE: - a cat is sitting on a suitcase on a bed - cat sitting on a piece of luggage - a small cat sitting on the back of a suitcase
COS-CVAE: - a cat sitting on top of luggage on the floor - a cat curled up on a piece of luggage - a close up of a very cute cat in a red suitcase	DCL-CVAE: - a cat lays down in a packed piece of luggage - a cat is sitting inside of a piece of blue luggage - a cat laying in a piece of luggage on a floor	HCVA-T-IC: - a cat laying on a small piece of luggage - there is a c at that is sitting inside of a piece of luggage - a brown and white cat laying inside of a suitcase

(*b*) 各方法对测试图像1的描述结果对比

(*c*) 测试图像2

Div-BS: - a couple of birds standig on top of a river - a couple of birds standig on top of a pond - a couple of birds that are standing next to each other	POS: - two white birds are starding in the water - two large white birds standing in the water - two birds are standing in the water together	Seq-CVAE: - the birds are swimming in the water and one is on the top - two birds are standing in the water and drinking - a group of birds on some water near water
COS-CVAE: - two birds are standing on the water at the beach - a couple of birds standing on top of a lake - two red and white birds standing next to each other	DCL-CVAE: - a group of birds with orange beaks in some water - a flock of three birds wading on the water - a couple of birds on a body of water	HCVA-T-IC: - three identical birds are seen wading in the water - a flock of three birds wading across a pond - a group of three birds standing near each other on a body of water

(*d*) 各方法对测试图像2的描述结果对比

图6 各种模型生成描述结果的定性对比

5 结论

本文提出了多样化图像描述生成框架HCVA-T-ICF. 该框架将混合条件变分自编码与Transformer模型进行无缝融合,通过引入全局和序列隐向量分别建模了图像描述句子级和单词级的多样性. 与现有方法相比,所提出的模型在无需数据增强、概念提示以及模型预训练的情况下,能够在描述准确性与多样性上取得更好的平衡,并增强了多样化描述生成过程的可解释性. 定量和定性实验验证了所提出方法的有效性. 下一步工作中,我们将引入扩散模型进行语言建模,进一步提高多样化图像描述过程的可控性.

参考文献

[1] STEFANINI M, CORNIA M, BARALDI L, et al. From show to tell: A survey on deep learning-based image captioning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 539-559.

[2] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2018: 6077-6086.

[3] YANG X, ZHANG H W, CAI J F. Deconfounded image

captioning: A causal retrospect[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 12996-13010.

[4] 石义乐, 杨文忠, 杜慧祥, 等. 基于深度学习的图像描述综述[J]. 电子学报, 2021, 49(10): 2048-2060.
SHI Y L, YANG W Z, DU H X, et al. Overview of image captions based on deep learning[J]. Acta Electronica Sinica, 2021, 49(10): 2048-2060. (in Chinese)

[5] 李志欣, 魏海洋, 黄飞成, 等. 结合视觉特征和场景语义的图像描述生成[J]. 计算机学报, 2020, 43(9): 1624-1640.
LI Z X, WEI H Y, HUANG F C, et al. Combine visual features and scene semantics for image captioning[J]. Chinese Journal of Computers, 2020, 43(9): 1624-1640. (in Chinese)

[6] 周东明, 张灿龙, 李志欣, 等. 基于多层级视觉融合的图像描述模型[J]. 电子学报, 2021, 49(7): 1286-1290.
ZHOU D M, ZHANG C L, LI Z X, et al. Image captioning model based on multi-level visual fusion[J]. Acta Electronica Sinica, 2021, 49(7): 1286-1290. (in Chinese)

[7] 刘茂福, 施琦, 聂礼强. 基于视觉关联与上下文双注意力的图像描述生成方法[J]. 软件学报, 2022, 33(9): 3210-3222.
LIU M F, SHI Q, NIE L Q. Image captioning based on visual relevance and context dual attention[J]. Journal of Software, 2022, 33(9): 3210-3222. (in Chinese)

[8] 宋井宽, 曾鹏鹏, 顾嘉扬, 等. 基于视觉区域聚合与双向协作的端到端图像描述生成[J]. 软件学报, 2023, 34(5): 2152-2169.
SONG J K, ZENG P P, GU J Y, et al. End-to-end image captioning via visual region aggregation and dual-level collaboration[J]. Journal of Software, 2023, 34(5): 2152-2169. (in Chinese)

[9] DAI B, FIDLER S, URTASUN R, et al. Towards diverse and natural image descriptions via a conditional GAN[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2970-2979.

[10] SHETTY R, ROHRBACH M, HENDRICKS L A, et al. Speaking the same language: Matching machine to human captions by adversarial training[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 4135-4144.

[11] HUIJBEN I A M, KOOL W, PAULUS M B, et al. A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1353-1371.

[12] WANG L W, SCHWING A G, LAZEBNIK S. Diverse and accurate image description using a variational auto-encoder with an additive Gaussian encoding space[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 5758-5768.

[13] ANEJA J, AGRAWAL H, BATRA D, et al. Sequential latent spaces for modeling the intention during diverse image captioning[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 4261-4270.

[14] MAHAJAN S, ROTH S. Diverse image captioning with context-object split latent spaces[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2020: 3613-3624.

[15] WANG J, XU W, WANG Q, et al. On distinctive image captioning via comparing and reweighting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 2088-2103.

[16] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows[C]// 2021 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2021: 9992-10002.

[17] 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. New York: ACM, 2002: 311-318.

[18] 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: Association for Computational Linguistics, 2005: 65-72.

[19] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//In Text summarization branches out: Proceedings of the ACL-04 workshop. Stroudsburg: Association for Computational Linguistics, 2004: 74-81.

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

[21] VIJAYAKUMAR A, COGSWELL M, SELVARAJU R, et al. Diverse beam search for improved description of complex scenes[C]//Proceedings of the AAAI Conference

on Artificial Intelligence. Melbourne: AAAI Press, 2018: 7371-7379.

[22] DESHPANDE A, ANEJA J, WANG L W, et al. Fast, diverse and accurate image captioning guided by part-of-speech[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 10695-10704.

[23] XU J, LIU B, ZHOU Y, et al. Diverse image captioning via conditional variational autoencoder and dual contrastive learning[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2024, 20 (1): 29.

[24] ZHENG Y, LI Y L, WANG S J. Divcon: Learning concept sequences for semantically diverse image captioning [C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2023: 1-5.

作者简介

刘 兵 男,1981年8月出生于河南省永城市. 现为中国矿业大学计算机科学与技术学院副教授. 主要研究方向为机器学习和人工智能.
E-mail: liubing@cumt.edu.cn

李 穗 男,1997年6月出生于江西省赣州市. 现为中国矿业大学计算机科学与技术学院硕士研究生. 主要研究方向为计算机视觉和人工智能.
E-mail: suili@cumt.edu.cn

刘明明 女,1985年4月出生于安徽省宿州市. 现为中国矿业大学计算机科学与技术学院博士后. 主要研究方向为深度学习和图像处理.
E-mail: jsjzi_lmm@126.com

刘 浩 男,1994年12月出生于河南省永城市. 现为中国矿业大学计算机科学与技术学院博士. 主要研究方向为深度学习.
E-mail: TB20170007B4@cumt.edu.cn