基于条件变分推断与内省对抗学习的多样化图像描述生成

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: 现有多样化图像描述生成方法受到隐空间表示能力和评价指标制约,很难同时兼顾描述生成的多样性和准确性. 为此,本文提出了一种新的多样化图像描述生成模型,该模型由一个条件变分推断编码器和一个生成器组成. 编码器利用全局注意力学习每个单词的隐向量空间,以提升模型对描述多样化的建模能力. 生成器根据给定图像和序列隐向量生成多样化的描述语句. 同时,引入内省对抗学习的思想,条件变分推断编码器同时作为鉴别器来区分真实描述和生成的描述,赋予模型自我评价生成的描述语句的能力,克服预定义评价指标的局限性. 在MSCOCO数据集上的实验表明,与传统方法相比,在随机生成100个描述语句时,多样性指标mBLEU(mutual overlap-BiLingual Evaluation Understudy)提升了1.9%,同时准确性指标CIDEr(Consensus-based Image Description Evaluation)显著提升了7.5%. 与典型多模态大模型相比,所提出方法在较小参数量的条件下更适用于生成多样化的陈述性描述语句.

关键词: 图像描述;变分推断;对抗学习;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)07-2219-09
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20231156

Diverse Image Captioning via Conditional Variational Inference and Introspective Adversarial Learning

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. *Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou, Jiangsu* 221116, *China*)

Abstract: Limited by the latent space modeling ability and pre-defined diversity metrics, most diverse image captioning models fail to achieve a balance between diversity and accuracy. To this end, we propose a novel diverse image captioning framework, which consists of a transformer based variational inference encoder and a generator. Specifically, the variational inference network aims to learn a latent space for each word to enhance the ability of caption diversity modeling, while the generator network produces diverse captions conditioned on each image and a sequence of latent variables. To overcome the limitation of pre-defined metrics, we introduce introspective adversarial learning into the proposed model, where the variational inference network also serves as a discriminator to distinguish between the ground truth captions and those produced by the generator without extra discriminators. The proposed method is endowed the ability to self-evaluate the quality of generated captions. The experimental results on dataset MSCOCO show that compared with the conventional methods, the proposed method with 100 samples improves the mBLEU (mutual overlap-BiLingual Evaluation Understudy) scores by 1.9% and the CIDEr (Consensus-based Image Description Evaluation) scores by 7.5%, respectively. Compared with typical multimodal large models, the proposed method is more suitable for generating diverse declarative descriptive captions with smaller parameters.

Key words: image captioning; variational inference; adversarial learning; 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:2024-05-19;责任编辑:覃怀银

1 引言

图像描述生成作为计算机视觉与自然语言处理交叉领域的基础性研究课题[1~3],旨在让机器理解图像的内容,并且以人类语言的形式自动生成对应的描述语句. 图像描述生成致力于让机器拥有“看图说话”的能力,具有重要科学研究和应用价值. 在图像检索和分类领域,用于提升图像内容检索和分类的准确度[4]. 在视觉辅助领域,帮助视力障碍人群理解图像信息,弥补视觉缺陷[5]. 在智慧医疗领域,可以自动生成医学图像诊断报告,为智能诊疗提供技术支撑[6].

传统基于卷积神经网络以及基于Transformer架构的方法仅关注描述生成的准确性[7,8],而忽视了描述生成的多样性. 近年来,多样化的图像描述生成方法逐渐成为研究热点. 一些工作将生成对抗网络[9]和变分自编码器[10]引入传统的图像描述模型. Dai等人[9]首先提出了一种基于条件生成对抗性网络的多样化图像描述模型. 虽然这种方法提高了多样性,但它以损失图像描述的准确性为代价,并且训练过程不稳定. 此外,一些研究人员引入条件变分自编码模型[10~12],通过从学习的隐空间中采样的方式生成多样化描述. 尽管这些方法取得了较好的准确性指标,但描述生成的多样性受限于隐空间的表示能力. 为此,Wang等人[13]提出将多样化描述指标和强化学习相结合的方法提升描述多样性. Xu等人[14]利用双流条件变分自编码与对比学习,通过提升隐空间建模能力促进多样化描述生成. 然而,这类方法通常引入额外的判别网络或人工评价指标,模型性能严重依赖预定义的评价指标.

随着算力的提升,近年来先后出现了不同参数规模的多模态图文大模型. BLIP2(Bootstrapping Language-Image Pre-training-2)[15]利用冻结图像编码器和大型语言模型引导语言和图像大规模预训练,但其不具备多样化描述生成能力. LLaVA-V1.6(Large Language and Vision Assistant)[16]和GPT4V(Generative Pre-trained Transformer-4 Vision)[17]进一步增大了模型参数规模,通过输入提示词支持生成多样化的描述语句. 然而,它们存在生成不真实或虚构描述的倾向,即所谓的幻觉现象.

针对上述问题,本文提出一种新的基于条件变分推断和内省对抗学习的图像描述生成模型,该模型由一个条件变分推断编码器和一个生成器组成. 其中,编码器将序列变分自编码器与Transformer模型无缝融合,能够有效学习每个单词的隐向量空间. 生成器以每个图像和序列隐向量为条件,通过在隐空间采样生成多样化的描述语句. 为了自评价生成的多样化描述,条件变分推断编码器同时充当鉴别器,用来区分真实描述和生成的描述. 编码器和生成器以内省对抗学习的方式联合训练. 在无需引入额外鉴别网络的情况下,所提方法在准确性和多样性上优于传统方法. 与典型多模态大模型相比,所提出方法在较小参数量的条件下更适用于生成多样化的陈述性描述语句.

2 方法

受内省变分自编码模型模拟人类内省学习能力的启发[18],本文提出基于条件变分推断和内省对抗学习的图像描述生成模型,实现在内省对抗学习过程中自评价生成的描述语句. 所提出模型的整体结构如图1所示,主要包含序列化条件变分推断编码器和生成器两个模块. 其中变分推断编码器将序列化条件变分自编码器与Transformer模型相结合,将一幅图像和其对应的描述文本映射为一组序列隐变量,并同时作为鉴别器来评估真实描述和生成描述之间的差异. 生成器网络则是利用Gumbel Softmax近似从单词类别的概率分布中采样出整个描述语句,并将其反馈到变分推断网络中进行对抗学习,变分推断编码器和生成器以内省对抗学习方式进行联合交替优化.

2.1 序列化条件变分推断编码网络

给定输入图像特征$\boldsymbol{v}$,所提模型旨在生成多个描述语句$\boldsymbol{x}^k, k\in\{1,2,\cdots,K\}$. 每个长度为$T$的描述表示为$\boldsymbol{x}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_T)$,其中,单词$\boldsymbol{x}_t\in D, t\in\{1,2,\cdots,T\}$,$D$表示词汇表. 受序列变分自编码模型[12]的启发,将序列化隐变量$\boldsymbol{s}=(\boldsymbol{s}_1,\cdots,\boldsymbol{s}_T)$引入Transformer架构,以建模每个时间步生成单词的多样性. 通过使用时间因子分解,条件分布概率$p_\theta(\boldsymbol{x}|\boldsymbol{v})$可表示为

$$\begin{aligned} p_\theta(\boldsymbol{x}|\boldsymbol{v}) &= \sum_{\boldsymbol{s}} p_\theta(\boldsymbol{x},\boldsymbol{s}|\boldsymbol{v}) \\ &= \sum_{\boldsymbol{s}}\prod_t p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{<t},\boldsymbol{s}_{\leqslant t},\boldsymbol{v})p_\theta(\boldsymbol{s}_t|\boldsymbol{x}_{<t},\boldsymbol{s}_{<t},\boldsymbol{v}) \end{aligned} \tag{1}$$

其中,$\boldsymbol{\theta}$为模型参数. 进一步,通过最大化条件分布$p_\theta(\boldsymbol{x}|\boldsymbol{v})$的对数似然可以得到如下变分下界:

$$\begin{aligned} \log p_\theta(\boldsymbol{x}|\boldsymbol{v}) &\geqslant \mathrm{E}_{q_\phi(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x},\boldsymbol{I})}\left[\sum_t \log p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{<t},\boldsymbol{s}_{\leqslant t},\boldsymbol{v})\right] \\ &\quad -\sum_t \mathrm{KL}(q_\phi(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x},\boldsymbol{v})\|p_\theta(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x}_{<t},\boldsymbol{v})) \\ &= -L_{\mathrm{XE}}+L_{\mathrm{KL}}(\boldsymbol{x},\boldsymbol{v}) \end{aligned} \tag{2}$$

其中,$\boldsymbol{\theta}$和$\boldsymbol{\phi}$分别表示先验概率模型$p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{<t},\boldsymbol{s}_{\leqslant t},\boldsymbol{v})$和后验概率模型$q_\phi(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x},\boldsymbol{v})$的参数. 变分下界的第一项表示所生成单词序列数据的对数似然,第二项为每个时间步计算的后验和条件先验分布之间的KL散度之和.

如图1所示,变分推断编码器采用预训练Faster R-CNN[19]提取图像区域特征,然后输入到由n个注意力块组成的编码器中得到视觉特征$\boldsymbol{v}$. 分别对概率模型$q_\phi(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x},\boldsymbol{v})$、$p_\theta(\boldsymbol{s}_t|\boldsymbol{s}_{<t},\boldsymbol{x}_{<t},\boldsymbol{v})$、$p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{<t},\boldsymbol{s}_{\leqslant t},\boldsymbol{v})$进行神经网

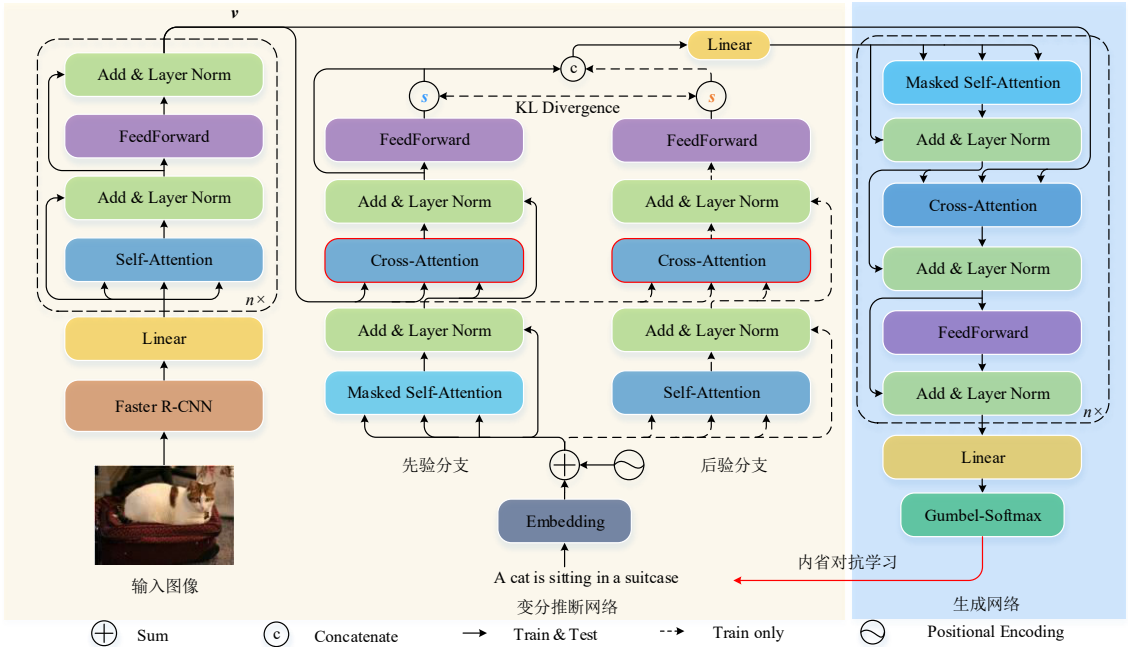

图 1　多样化图像描述生成模型架构

络参数化近似,分别对应后验推断子网络、先验近似子网络和生成网络. 模型具体实现细节如下:

图像视觉特征 $\boldsymbol{v}$ 分别输入到后验推断子网络和先验近似子网络,进行双分支变分推断. 为了降低计算复杂性,先验和后验模型分别简化为 $p_\theta(\boldsymbol{s}_t|\boldsymbol{s}_{t-1},\boldsymbol{x}_{<t},\boldsymbol{v})$ 和 $q_\phi(\boldsymbol{s}_t|\boldsymbol{s}_{t-1},\boldsymbol{x},\boldsymbol{v})$.

首先将单词嵌入向量进行位置编码得到文本输入向量 $\boldsymbol{w}_0$. 随后,使用多头自注意力模块对 $\boldsymbol{w}_0$ 进行处理,并将其输入残差归一化层,提取出以下文本语义特征:

$$\boldsymbol{w}_n=\text{LayerNorm}(\text{MultiHead}(\boldsymbol{w}_0,\boldsymbol{w}_0,\boldsymbol{w}_0)+\boldsymbol{w}_0) \tag{3}$$

其中,MultiHead(·)表示多头自注意力模块,LayerNorm(·)表示残差归一化模块,文本输入向量 $\boldsymbol{w}_0\in\mathbb{R}^{T\times d}$,文本语义特征 $\boldsymbol{w}_n\in\mathbb{R}^{T\times d}$,$T$ 表示描述长度,d 表示特征维度. 通过多头交叉注意力模块和残差归一化层,进一步将文本特征 $\boldsymbol{w}_n$ 与视觉特征 $\boldsymbol{v}$ 进行融合,融合特征 $\boldsymbol{u}$ 表示为

$$\boldsymbol{u}=\text{LayerNorm}(\text{MultiHead}(\boldsymbol{w}_n,\boldsymbol{v},\boldsymbol{v})+\boldsymbol{w}_n) \tag{4}$$

其中,$\boldsymbol{u}\in\mathbb{R}^{T\times d}$,$\boldsymbol{v}\in\mathbb{R}^{m\times d}$ 表示 m 个图像区域特征. 由于后验模型 $q_\phi(\boldsymbol{s}_t|\boldsymbol{s}_{t-1},\boldsymbol{x},\boldsymbol{v})$ 中的后验序列隐变量 $\boldsymbol{s}_t$ 依赖于图像特征 $\boldsymbol{v}$、整个句子 $\boldsymbol{x}$ 和前一个隐变量 $\boldsymbol{s}_{t-1}$,假设后验分布为多元独立高斯分布 $\text{N}(\boldsymbol{s}_t;\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)$,其均值和标准差分别为 $\boldsymbol{\mu}_t(\boldsymbol{s}_{t-1},\boldsymbol{x},\boldsymbol{v})$ 和 $\boldsymbol{\sigma}_t(\boldsymbol{s}_{t-1},\boldsymbol{x},\boldsymbol{v})$,具体表示如下:

$$q_\phi(\boldsymbol{s}_t|\boldsymbol{s}_{t-1},\boldsymbol{x},\boldsymbol{v})=\text{N}(\boldsymbol{s}_t;\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t) \tag{5}$$

其中,分别使用两个前馈网络生成均值与方差,$\boldsymbol{\Sigma}_t$ 表示协方差矩阵. 在训练过程中,通过重参数技巧随机采样得到后验序列隐变量 $\boldsymbol{s}_t$:$\boldsymbol{s}_t=\boldsymbol{\mu}_t+\boldsymbol{\sigma}_t\odot\boldsymbol{\varepsilon}$. 其中,$\boldsymbol{\varepsilon}$ 服从标准多元正态分布,“⊙”表示逐元素相乘. 先验分支网络使用类似的方法构建与采样先验隐变量.

2.2　生成网络

如图 1 所示,生成网络由标准的 Transformer 解码器和 Gumbel 采样器组成. 在训练阶段,首先将后验隐变量 $\boldsymbol{s}_t$ 与后验网络中的融合特征 $\boldsymbol{u}$ 进行拼接操作,然后利用全连接层进行降维,最后将降维后的特征输入生成器. 对先验隐变量和先验近似网络中的融合特征采用同样的处理,并将降维后的特征也输入生成器. 在测试阶段,仅利用先验近似网络为生成器提供输入特征.

在每个时间步,生成器通过类别分布 $p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{<t},\boldsymbol{s}_{\leqslant t},\boldsymbol{v})$ 对单词进行预测. 生成器的输出被重新反馈到变分推断编码器,以实现对生成描述的判别. 然而,通过 arg-max 函数进行采样操作是不可导的,直接将离散的单词输入变分推断编码器无法进行误差反向传播. 为此,引入 Gumbel-Softmax 近似,通过对离散类别分布进行采样,使得采样得到的样本在传入变分推断网络时可以进行误差的反向传播.

2.3　模型训练与推断

在联合训练阶段,模型通过式(2)中的优化目标进行训练. 在内省对抗学习阶段,为了将序列条件变分自编码模型中的优化目标与内省对抗学习相结合,分别设计了两种不同的损失来交替训练变分推断网络和生成网络:

$$L_E=L_{\text{KL}}(\boldsymbol{v},\boldsymbol{x})+\alpha[\delta-L_{\text{KL}}(\boldsymbol{v},\text{sg}(\boldsymbol{x}'))]^+\beta L_{\text{XE}} \tag{6}$$

$$L_G=\alpha L_{\text{KL}}(\boldsymbol{v},\boldsymbol{x}')+L_{\text{XE}} \tag{7}$$

其中,$[\cdot]^+=\max(0,-)$,δ 表示正的间隔参数,$\text{sg}(\cdot)$ 表示反向传播在此停止,$\boldsymbol{x}'$ 表示通过后验隐变量生成的描述,α 和 β 为平衡因子.

在训练阶段，当$L_{\mathrm{KL}}(\boldsymbol{v},\mathrm{sg}(\boldsymbol{x}'))\leqslant\delta$时，变分推断网络和生成网络进行最小-最大博弈．前者通过最大化$L_{\mathrm{KL}}(\boldsymbol{v},\mathrm{sg}(\boldsymbol{x}'))$，以区分真实描述与采样得到的描述，生成网络则通过对抗训练以生成具有较小$L_{\mathrm{KL}}(\boldsymbol{v},\boldsymbol{x}')$的描述语句．当变分推断网络不再能够区分真实描述和生成描述时，训练过程达到纳什均衡．因此，模型无需使用预定义的多样性指标来评估生成的描述，而且能够在不引入额外的判别器的情况下，以内省对抗方式稳定地进行训练．

在推断阶段真实描述不可见，从训练后的先验分支采样一组序列化隐向量作为生成器的输入．由于在训练阶段先验和后验隐向量已经进行了对齐，并且编码器对先验隐向量生成的描述语句进行了自评价，因此训练后的先验分支具备预测描述不可见部分的能力，从而保证在推断阶段从先验分支采样是有效的．在预测单词时，通过Softmax和Gumbel-Softmax两种方式进行集成推断，并结合束搜索策略生成多样化的描述．

3 实验

3.1 数据集和评价标准

3.1.1 数据集

实验中在MSCOCO数据集上对提出的模型进行训练与测试．使用常用的M-RNN(Multimodal Recurrent Neural Networks)数据集划分设置[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．

3.1.2 准确性指标

实验采用了五种广泛使用的准确性指标[20-22]，包括CIDEr(Consensus-based Image Description Evaluation)、BLEU-N(BiLingual Evaluation Understudy for N-gram)、METEOR(Metric for Evaluation of Translation with Explicit ORdering)、ROUGE-L(Recall-Oriented Understudy for Gisting Evaluation：Longest common subsequence)和SPICE(Semantic Propositional Image Caption Evaluation)．其中，BLEU(BiLingual Evaluation Understudy)通过计算生成和参考文本之间n-gram的精准率来评价生成文本的精确性，得分越高表示生成描述与参考描述的短语重合度越高．METEOR在BLEU的基础上进一步考虑了召回率，通过召回率和精确率计算调和平均值，其得分越高表示生成的句子与参考句子在词汇、语法和语义等多个层面上的匹配度较高．ROUGE(Recall-Oriented Understudy for Gisting Evaluation)是通过比较生成文本和人工标注文本相同的部分，实现对单词重复率以及排列顺序相似度的计算，高的ROUGE-L得分表示生成描述与参考描述重合度较好．CIDEr是通过计算n-gram的TF-IDF(Term Frequency-Inverse Document Frequency)指标来计算生成文本和多个人工标注文本之间的相似性．因此，其数值越高表示生成描述的质量较高，更接近人类描述．SPICE指标利用场景图计算了生成与真实描述之间的语义相似性，得分越高表示生成的句子更准确地描述了图像中的场景，并且在语义上与参考描述更为接近．

多样化描述生成模型则需要针对生成的一组描述的准确性进行评价．目前大多采用Oracle重排序方法[12,13]．具体地，使用测试图像的真实描述作为参考，在生成的一组描述中，选择准确性指标得分最高的描述，然后计算所有测试图像的最优准确性评价指标的平均值．

3.1.3 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[12,13]计算多样性指标．其中，对于一张生成了多个描述的测试图像，首先计算其与训练集中相似度最高的k个图像，然后将生成描述分别与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Best-5个描述中，不重复的描述所占比例．Uniqueness越高说明生成句子与参考句子的重复率低，即生成的句子更加多样化．(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．Novel值越高说明生成与参考描述不同的句子数量越多，生成的描述多样性越好．(3)mBLEU(mutual overlap-BiLingual Evaluation Understudy)：对于每一张测试图像的Best-5，分别计算其中一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．mBLEU越低表明生成句子之间的相似词组越少，生成描述反而具有较好的多样性．(4)Div-1(1-gram Diversity)：计算每一张测试图像的Best-5中不重复的1-gram在五个描述总1-gram长度中所占比例，并取测试集平均．Div-1指标越高表示在生成句子中存在的多样化的单词较多．(5)Div-2(2-gram Diversity)：使用2-gram替换1-gram，计算方法同Div-1．Div-2指标越高表示在生成句子中存在的多样化的2-grams词组更多．

3.2 实验设置

训练与测试环境配置如下：操作系统为Ubuntu22.04，CPU为Intel Core i9-10900X 3.70 GHz×20，GPU为NVIDIA GeForce RTX 3080，内存为16 GB，同时采用CUDA11.6进行加速处理．

实验中，所提模型中的图像特征、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Faster R-CNN来提取每幅图像的1 536维区域特征，并将其线性映射为512维向量．在生成器中，对单词嵌入进行位置编码，并将其作为后验推断网络和先验近似网络的输入．此外，视觉编码器和生成器均是由3个

注意力块组成,其中多头注意力的头数设置为8. 在训练阶段,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,并进行30个回合的模型训练. 超参数α、β和δ根据经验分别设置为1.0、1.0和100. 在测试阶段,为了对比的公平性,在进行准确性评价时束搜索宽度统一设置为2,而在多样性评价时束宽设置为1.

3.3　消融实验

首先通过消融实验验证所提模型中内省对抗以及SCST(Self-Critical Sequence Training)强化学习后处理对模型性能的影响. 为公平对比,首先训练每一种实验设置下的模型,然后针对每幅测试图像均随机采样20个描述语句进行性能评价. 表1给出了不同方法的描述准确性和多样性指标. 其中,Baseline表示不使用内省对抗以及使用softmax预测时的基准模型. 准确性指标BLEU和CIDEr通过Oracle重排序进行评估,多样性指标Div-1和Div-2通过Consensus重排序后得到的最优的五个描述语句进行评估. 各指标的最优结果加粗表示,"↑"表示数值越大性能越好,"↓"反之. 为简便起见,使用B-N、C、R、M和S分别表示BLEU-N、CIDEr、ROUGE-L、METEOR和SPICE指标.

表1　MSCOCO数据集M-RNN划分下的消融实验

Baseline	内省对抗	SCST	B-4↑	C↑	Div-1↑	Div-2↑
√	×	×	0.520	1.713	0.37	0.61
√	√	×	0.536	1.739	0.39	0.63
√	√	×	**0.549**	**1.760**	**0.40**	**0.65**
√	√	√	0.494	1.720	0.24	0.33

从表1中可以看出,在引入内省对抗学习策略之后,所有模型的准确性和多样性指标都得到了提升. 使用SCST强化学习方法进行训练时,模型的多样性指标急剧下降,这是由于SCST方法仅关注准确性的提升,其在强化学习过程中倾向于为包含常见短语的句子分配较高的奖励分值. 本文模型通过内省对抗学习可以在训练过程中自评价生成的描述,在不引入预先定义的评价指标的条件下,仍然能够在准确性和多样性之间取得较好的平衡.

3.4　实验结果定量分析

为了证明本文模型的有效性,将提出的模型与主流多样化图像描述模型进行了定量对比. 这些方法包括Div-BS(Diverse Beam Search)[23]、PoS(Part-of-Speech)[24]、AG-CVAE(Conditional Variational Auto-Encoder with Additive Gaussian)[10]、Seq-CVAE(Sequential Conditional Variational Auto-Encoder)[11]、COS-CVAE(Context-Object Split Conditional Variational AutoEncoder)[12]、UDA(Unified Diversity and Accuracy)[13]和DCL-CVAE(Conditional Variational AutoEncoder with Dual Contrastive Learning)[14].

在MSCOCO数据集M-RNN划分条件下,使用Oracle重新排序后对各方法进行了准确性对比,实验结果如表2所示. 其中,Oracle重排使用每个指标的最大得分对不同的描述进行排序并选择最好的描述,每个Oracle评估分数都是测试集的平均值. 与对比方法常用实验设置相同,通过采样20和100个隐变量,然后输入生成器解码出多样化的描述语句. 本文方法在所有准确性指标上的得分均优于所对比的方法.

表2　各方法在MSCOCO数据集M-RNN测试集上使用Oracle重新排序后准确性的性能对比

采样数量	方法	B-4↑	B-3↑	B-2↑	B-1↑	C↑	R↑	M↑	S↑
20	Div-BS[23]	38.3	53.8	68.7	83.7	140.5	65.3	35.7	26.9
	PoS[24]	44.9	59.3	73.7	87.4	146.8	67.8	36.5	27.7
	AG-CVAE[10]	47.1	57.3	69.8	83.4	130.8	63.8	30.9	24.4
	Seq-CVAE[11]	44.5	59.1	72.7	87.0	144.8	67.1	35.6	27.9
	DCL-CVAE[14]	45.9	59.8	73.5	88.5	150.2	67.8	35.8	29.4
	COS-CVAE[12]	50.0	64.0	77.1	90.3	162.4	70.6	38.7	29.5
	UDA[13]	52.1	—	—	—	168.1	71.4	40.0	31.1
	本文方法	**54.9**	**67.2**	**79.2**	**92.1**	**176.0**	**72.7**	**41.5**	**33.5**
100	Div-BS[23]	40.2	55.5	69.8	84.6	144.8	66.6	37.2	29.0
	PoS[24]	55.0	67.2	78.7	90.9	166.1	72.5	40.9	31.1
	AG-CVAE[10]	55.7	65.4	76.7	88.3	151.7	69.0	34.5	27.7
	Seq-CVAE[11]	57.5	69.1	80.3	92.2	169.5	73.3	41.0	32.0
	DCL-CVAE[14]	61.1	71.8	82.8	94.4	182.3	75.2	42.8	33.7
	COS-CVAE[12]	63.3	73.9	84.2	94.2	189.3	77.0	45.0	33.9
	UDA[13]	61.7	—	—	—	193.2	76.5	45.4	35.2
	本文方法	**70.2**	**78.5**	**87.5**	**96.5**	**207.8**	**80.8**	**50.9**	**37.8**

值得注意的是,所提出的模型在生成过程中没有引入其他的额外信息,而AG-CVAE、POS、DCL-CVAE和COS-CVAE分别在推理过程中利用目标对象信息、PoS标签、预训练模型和增强的上下文信息.这些实验结果有效验证了本文模型能够生成更加准确的图像描述语句.

表3进一步评估了不同模型在Consensus重排条件下的多样性指标得分.从表3可以看出,本文方法的综合多样性表现更好.在分别采样20和100个描述语句的条件下,本文方法的mBLEU指标得分更有优势,这表明它生成的不同描述之间有着显著的差异.在Div-1和Div-2指标上,本文显著优于其他方法,其在20和100个采样中分别获得0.73和0.61的Div-2指标得分.

表3 各方法在MSCOCO数据集M-RNN测试集上使用Consensus重新排序后多样性的性能对比

采样数量	方法	Uniqueness↑	Novel↑	mBLEU↓	Div-1↑	Div-2↑
20	Div-BS[23]	**100**	3 106	0.81	0.20	0.26
	PoS[24]	96.3	3 394	0.64	0.24	0.35
	AG-CVAE[10]	69.8	3 189	0.66	0.24	0.34
	Seq-CVAE[11]	94.0	4 266	0.52	0.25	0.54
	DCL-CVAE[14]	97.9	4 899	0.54	0.39	0.65
	COS-CVAE[12]	96.0	4 249	0.52	0.33	0.52
	UDA[13]	98.5	4 862	0.40	0.43	0.70
	本文方法	99.1	**4 955**	**0.43**	**0.46**	**0.73**
100	Div-BS[23]	**100**	3 421	0.82	0.20	0.25
	PoS[24]	91.5	3 446	0.67	0.23	0.33
	AG-CVAE[10]	47.4	3 069	0.70	0.23	0.32
	Seq-CVAE[11]	84.2	4 215	0.64	0.33	0.48
	DCL-CVAE[14]	92.1	4 607	0.66	0.35	0.55
	COS-CVAE[12]	96.3	4 404	0.53	0.39	0.57
	UDA[13]	96.8	4 586	0.53	0.39	0.59
	本文方法	97.8	**4 643**	**0.52**	**0.40**	**0.61**

综上,在准确性和多样性指标上,本文方法较现有的方法均获得了显著的性能提升,说明提出的方法能够生成准确而多样的描述,进一步验证了所提模型隐空间表征能力和自评价机制的有效性.

3.5 实验结果定性分析

图2(*b*)和图2(*d*)对比了所提方法和传统多样化描述生成方法采样得到的语句.其中,红色表示错误单词,绿色表示重复短语.可以看出,与其他方法相比,本文方法可以准确识别出图像中的鸟的数量,而其他方法则会生成不准确的量词,以及一些不正确的单词.此外,其他方法倾向于生成重复单词或短语,而本文方法则可以有效缓解模式坍塌问题,生成更自然和多样的描述,例如,生成了不常见的单词"long legged"和"long necked"等.

为了进一步对比所提出方法和流行的多模态大模型生成多样化描述的能力,实验中进一步对BLIP2[15](参数量:2.7×10^9)、LLaVA-V1.6(参数量:34×10^9)[16]和GPT4V(参数量:$>1\times10^{12}$)[17]三种典型多模态大模型进行了实验测试.其中,对LLaVA-V1.6和GPT4V分别输入了"Generate 3 diverse COCO-style captions"提示词,生成三个不同的描述语句.如图2(*e*)和图2(*f*)所示,BLIP2不具备多样化描述生成能力,仅能生成单一描述语句.尽管LLaVA-V1.6和GPT4V根据提示词生成了测试图像对应的不同的描述,且描述内容更加丰富,然而它们在描述生成过程中存在幻觉现象,即存在生成不真实或虚构描述的倾向,红色标记的内容表示因大模型幻觉问题生成的无法进行验证的虚构描述.由于所提出的方法引入了生成描述的内省评价机制,在较小参数量(77.9 M)的条件下仍然能够生成准确而多样的描述.因此,若生成与图像语义更加相符的陈述性短描述语句,本文提出的方法更加适用.

(*a*) 测试图像 1

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

本文方法:
- a white and brown cat rests on a suitcase on floor
- a cat that is sitting on a piece of red and black luggage
- a fluffy cat laying down on top of a suitcase

(*b*) 本文与传统方法对测试图像 1 生成描述对比

(*c*) 测试图像 2

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

本文方法:
- there are many small birds walking in the water
- a group of long legged birds wading in a body of water
- several long necked birds standing in the pond

(*d*) 本文与传统方法对测试图像 2 生成描述对比

LLaVA-V1.6:
- A curious cat sits atop a vibrant red suitcase, its eyes gleaming with interest.
- A curious feline lounging on a piece of luggage, seemingly intrigued by its surroundings.
- A fluffy white and brown cat is the unexpected centerpiece on a red suitcase, adding a touch of whimsy to the scene.

GPT4V:
- A cat perched comfortably upon a suitcase, possibly hoping to tag along on a trip.
- The watchful eyes of a cat guard the contents of a closed suitcase.
- A pet cat sits on a suitcase, blending its fur with the suitcase's intricate pattern.

BLIP2:
- a cat laying on top of a suitcase

本文方法:
- a white and brown cat rests on a suitcase on floor
- a cat that is sitting on a piece of red and black luggage
- a fluffy cat laying down on top of a suitcase

(*e*) 本文与典型多模态大模型对测试图像 1 生成描述对比

LLaVA-V1.6:
- Three elegant birds with striking red accents, captured in a moment of tranquility on a wet surface.
- Three elegant birds with striking red accents, poised on a wet ground
- A group of three white birds with red beaks and legs, captured in a moment of tranquility on a reflective surface.

GPT4V:
- Feathered trio braving the elements in their watery domain.
- Ungainly grace demonstrated by water-loving birds in the rain.
- Trio of aquatic birds, likely ibises, gathered on a water-covered habitat.

BLIP2:
- a cat laying on top of a suitcase

本文方法:
- there are many small birds walking in the water
- a group of long legged birds wading in a body of water
- several long necked birds standing in the pond

(*f*) 本文与典型多模态大模型对测试图像 2 生成描述对比

图 2　本文方法与多样化描述方法定性对比

4　结论

本文将条件变分推断和内省对抗学习引入图像描述生成,提出了一种新的多样化图像描述生成方法. 通过优化条件概率模型的变分证据下界,将序列化条件变分自编码与图像描述生成模型无缝融合,构建基于 Transformer 的图像多样化描述模型,以提升模型对单词级多样性的建模能力. 为克服预定义评价指标的局限性,引入内省对抗学习的思想,将编码网络作为判别器进行内省对抗学习. 在无需额外判别器的情况下,使模型具备对生成语句质量的自评价能力,从而进一步提升多样化描述的准确性和多样性. 定性和定量实验结果表明,在准确性和多样性指标上,所提出的方法均显著优于传统模型. 与多模态大模型相比,本文方法更适用于生成多样化的陈述性描述语句. 下一步工作中,我们将引入扩散模型进行语言建模,以进一步提高多样化图像描述生成的性能.

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] 刘浩阳, 林耀进, 刘景华, 等. 由粗到细的分层特征选择[J]. 电子学报, 2022, 50(11): 2778-2789.
LIU H Y, LIN Y J, LIU J H, et al. Hierarchical feature selection from coarse to fine[J]. Acta Electronica Sinica, 2022, 50(11): 2778-2789. (in Chinese)

[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] 卓亚琦, 魏家辉, 李志欣. 基于双注意模型的图像描述生成方法研究[J]. 电子学报, 2022, 50(5): 1123-1130.
ZHUO Y Q, WEI J H, LI Z X. Research on image caption-

ing based on double attention model[J]. Acta Electronica Sinica, 2022, 50(5): 1123-1130. (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] 魏博文, 全红艳. 基于语义与形态特征融合的语义分割网络[J]. 电子学报, 2022, 50(11): 2688-2697.
WEI B W, QUAN H Y. Semantic segmentation network based on semantic and morphological feature fusion[J]. Acta Electronica Sinica, 2022, 50(11): 2688-2697. (in Chinese)

[7] 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)

[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] 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.

[11] 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.

[12] 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. New York: ACM, 2020: 3613-3624.

[13] WANG Q, WAN J, CHAN A B. On diversity in image captioning: Metrics and methods[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44 (2): 1035-1049.

[14] 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.

[15] LI J N, LI D X, SAVARESE S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of the 40th International Conference on Machine Learning. New York: ML Research Press, 2023: 19730-19742.

[16] LIU H T, LI C Y, WU Q Y, et al. Visual instruction tuning[C]//Proceedings of the 37st International Conference on Neural Information Processing Systems. New York: ACM, 2023: 4566-4575.

[17] CHEN L, LI J, DONG X, et al. Sharegpt4v: Improving large multi-modal models with better captions[EB/OL]. (2023-11-28)[2024-05-21]. https://arxiv.org/abs/2311.12793.

[18] HUANG H B, LI Z H, HE R, et al. IntroVAE: Introspective variational autoencoders for photographic image synthesis[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 52-63.

[19] 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.

[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] 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.

[22] WANG J N, XU W J, WANG Q Z, et al. On distinctive image captioning via comparing and reweighting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2088-2103.

[23] 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. Menlo Park: AAAI Press, 2018:

7371-7379.

[24] 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.

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