计算机科学
COMPUTER SCIENCE

基于解耦-检索-生成的图像风格化描述生成模型

陈章辉，熊赟

引用本文

陈章辉，熊赟. 基于解耦-检索-生成的图像风格化描述生成模型[J]. 计算机科学, 2022, 49(6): 180-186.

CHEN Zhang-hui, XIONG Yun. Stylized Image Captioning Model Based on Disentangle-Retrieve-Generate[J]. Computer Science, 2022, 49(6): 180-186.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于特征注意力融合网络的遥感变化检测研究

Remote Sensing Change Detection Based on Feature Fusion and Attention Network

计算机科学, 2022, 49(6): 193-198. https://doi.org/10.11896/jsjkx.210500058

基于样本分布损失的图像多标签分类研究

Study on Multi-label Image Classification Based on Sample Distribution Loss

计算机科学, 2022, 49(6): 210-216. https://doi.org/10.11896/jsjkx.210300267

多分支 RA 胶囊网络及在图像分类中的应用

Multi-branch RA Capsule Network and Its Application in Image Classification

计算机科学, 2022, 49(6): 224-230. https://doi.org/10.11896/jsjkx.210400087

机器学习在金融资产定价中的应用研究综述

Application of Machine Learning in Financial Asset Pricing:A Review

计算机科学, 2022, 49(6): 276-286. https://doi.org/10.11896/jsjkx.210900127

基于注意力门控图神经网络的文本分类

Text Classification Based on Attention Gated Graph Neural Network

计算机科学, 2022, 49(6): 326-334. https://doi.org/10.11896/jsjkx.210400218

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100129

基于解耦-检索-生成的图像风格化描述生成模型

陈章辉　熊　赟
复旦大学计算机科学技术学院　上海 200433
上海市数据科学重点实验室　上海 200433
(18210240004@fudan.edu.cn)

摘　要　图像描述旨在为输入的图像生成描述文本以准确描述图像内容,而图像的风格化描述在此基础上引入了对语言风格的考虑,恰当表达出特定的语言风格,使得模型生成的描述文本更具多样性。为了更好地在生成的描述文本中融入风格元素,提出了基于解耦-检索-生成的图像风格化描述生成模型。该模型首先将风格化语料中的句子拆分成内容词汇和风格词汇,并构建了一个内容-风格词汇的记忆模块;然后根据图像的事实描述从记忆模块中检索出与之相匹配的风格词汇;最后将图像的事实描述和检索出的风格词汇输入语言模型中生成风格描述。在真实数据集上的实验结果表明,相比已有方法,所提模型在各项评价指标上都有着更好的性能表现,可以在描述图像内容的同时表达出特定的风格。

关键词:图像描述;深度学习;文本生成;语言风格;编码器-解码器
中图法分类号　TP181

Stylized Image Captioning Model Based on Disentangle-Retrieve-Generate

CHEN Zhang-hui and XIONG Yun
School of Computer Science, Fudan University, Shanghai 200433, China
Shanghai Key Laboratory of Data Science, Shanghai 200433, China

Abstract　Image captioning aims to generate a description text for the input image to accurately describe the image content. The stylized image captioning goes a step further on the basis of image captioning and introduces the consideration of language style. It also needs appropriately express the specific language style, which makes the generated text more diverse. In order to better incorporate style factors to the description text, a stylized image captioning model based on disentangle-retrieve-generate framework is proposed. The model first splits the sentences in the stylized corpus into content and style parts, and constructs a content-style memory module, then retrieves appropriate style from the memory module according to the factual caption of the image. Finally, the factual caption and retrieved style part are input into the language model for stylized caption generation. Experimental results on real datasets show that, compared to existing methods, the proposed model has better performance in various evaluation metrics, and can accurately describe the image content while expressing a specific style.

Keywords　Image captioning, Deep learning, Text generation, Language style, Encoder-decoder

1　引言

图像描述(Image Captioning)可以为输入的图像生成自然语言描述,它不仅需要对图像内容有高层次、全面的理解,而且需要生成人们可以理解的描述来表达图像中的视觉信息。由于深度学习方法研究的兴起以及大规模高质量的图像-文本描述数据集(MS COCO[1]和 Flickr 30K[2]等)的公开,当前的图像描述模型已经取得了很大进展。然而,目前大多图像描述模型聚焦于对图像内容进行客观描述[3],目的是对图像中的对象、属性、动作和对象之间的关系进行描述,而忽略了语言层面的风格元素,如积极(Positive)、消极(Negative)、幽默(Humorous)和浪漫(Romantic)等,此类模型生成的描述文本被称为事实描述。

近年来,生成描述的语言风格也开始受到研究者的关注。语言风格[4]是人类语言的一个重要因素,它可以揭示一个人的性格、情感、社交关系和心理状态[5]。以情感(Sentiment)为例,它是日常交流中的一种常见现象,以至于超过一半的在线图像附带的文本都包含了表情符号[6]。Zhao 等[7]指出:"一个完善的图像描述模型不仅要准确描述图像的内容,而且要将语言风格恰当地融入到生成的描述中",这类模型被称为图像的风格化描述模型(Stylized Image Captioning),该模型所生成的包含特定风格的图像描述被称为风格化描述。

风格化描述可以丰富图像描述的表达并使其更具有吸引力[8],这对于许多实际应用而言相当有价值,例如增加用户在社交媒体上分享图像和视频的参与度[8],生成有吸引力的图像或新闻标题,以获得更多的用户点击率[7];以及儿童早教中

到稿日期:2021-11-12　返修日期:2022-02-23
通信作者:熊赟(yunx@fudan.edu.cn)

为图像编写有趣的描述[7]和在聊天机器人中的应用[3]等。

当前有关风格化描述生成的工作主要有：Mathews 等[6]提出使用 SentiCap 生成积极或者消极的图像描述，该模型由卷积神经网络提取视觉特征，由循环神经网络（Recurrent Neural Network，RNN）生成事实性词汇，由微调的 RNN 网络生成带有风格的词汇。Chen 等[9]在长短期记忆网络（Long Short-Term Memory，LSTM）中引入了风格对应的门控单元和参数，并采用适应性学习控制模型从风格化描述和事实性参考模型中的学习强度来生成风格化描述。上述方法[6,9]都依赖成对的图像-风格化描述进行模型训练。为了解决这一问题，Gan 等[8]通过分解 LSTM 网络中的参数矩阵来为图像生成幽默或浪漫的描述，Chen 等[10]为 LSTM 网络的门控单元添加了层标准化操作，Guo 等[3]采用了对抗训练的方法。然而，这些工作[3,8,10]都依赖文本生成器来获取风格知识，缺乏对风格知识显式地学习过程。Zhao 等[7]通过记忆机制对风格知识进行显式学习，但该机制需要为描述句子构建场景图。此外，上述工作都是基于图像的视觉特征，利用 RNN 或 LSTM 网络直接生成风格化描述。

本文提出了基于解耦-检索-生成的图像风格化描述生成模型（Disentangle-Retrieve-Generate Captioning Model，DRGCap），该模型在解耦时构建了内容-风格词汇的记忆模块，显式提取了风格知识；检索时，借助图像描述模型，从生成的事实描述出发，将图像到文本的生成转化为文本到文本的生成，在消除视觉和文本之间语义鸿沟的同时，很好地保留了输入图像中的语义信息；生成过程中，文本编码器和解码器基于 Transformer 模型[11]进行实现，并利用风格化语料上训练的分类器来提供风格损失。实验结果证实了本文提出的 DRGCap 模型的有效性。

2 模型

2.1 问题描述

对于图像的风格化描述生成任务，具体而言，给定一幅输入图像 x 和风格标签 s，需要学习出一个模型 F 用于生成描述 $\hat{y}^s$，其中 $\hat{y}^s$ 既要正确描述图像 x 的视觉内容，又要满足标签 s 所指定的语言风格[3]，表示如下：

$$\hat{y}^s = F(x, s) \tag{1}$$

其中，$s \in \{Positive, Negative, Humorous, Romantic\}$。该任务中用来训练图像风格化描述模型的数据有成对的图像-事实描述数据集和非成对的风格化语料（即只有风格化描述而没有与之对应的图像）。

2.2 模型概述

本文提出的 DRGCap 模型示意图如图 1 所示，整个过程分为 3 步。

（1）解耦。这一步利用在风格化语料上预训练的分类器将风格化描述拆分成内容和风格词汇两部分，例如，对于句子“the basketball player is jumping to take a beautiful shot”，经过拆分，内容部分包含“the basketball player is jumping to take a shot”，风格词汇包含“beautiful”；然后，将内容和风格词汇组成二元组保存在记忆模块中。

（2）检索。对于输入图像，利用预训练的图像描述模型为其生成描述，然后根据生成的描述从目标风格的记忆模块中检索出与之最匹配的风格词汇。

（3）生成。基于步骤（2）生成的描述和检索出的风格词汇进行风格化描述的生成。

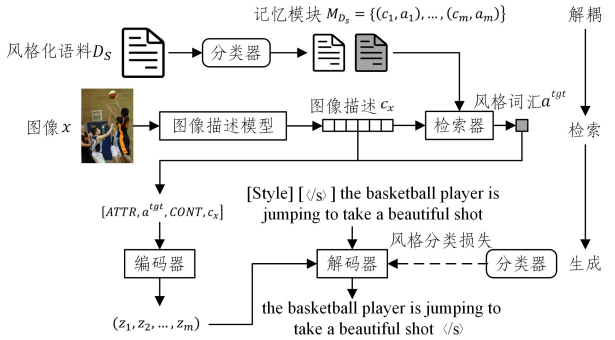


图 1 DRGCap 模型示意图

Fig. 1 Schematic diagram of DRGCap

2.3 解耦

解耦过程的目的是将风格化语料句子中的内容和风格词汇拆分开，该过程通过在风格化语料上预训练的分类器来完成。采用预训练分类器的好处在于，它能直接对输入的句子进行操作，还可以通过设置超参数控制分解的程度。给定一个输入句子 $y^s=(w_1, w_2, \cdots, w_n)$，其风格概率计算式如下：

$$p_{y^s} = P(s \mid y^s, \theta_C) \tag{2}$$

其中，p_{y^s} 是分类器预测出的风格概率值，C 和 θ_C 分别表示分类器及其参数。分类器基于 Kim[12]的工作进行实现，首先利用词嵌入层将单词编码为 512 维的向量，然后使用大小为 2～5 的 4 个卷积核来捕获元组特征，在卷积操作后依次经 ReLU 激活函数层和最大池化层，最后拼接成一个向量，经过全连接层预测结果输出。如果将 y^s 中的单词 w_i 删除，i 表示单词 w_i 在句子 y^s 中的索引，得到句子 $y^s_{-w_i}$，其风格概率计算式如下：

$$p_{y^s_{-w_i}} = P(s \mid y^s_{-w_i}, \theta_C) \tag{3}$$

定义式（2）和式（3）的差值为单词 w_i 的重要性分数（Importance Score，IS），其表达式如下：

$$IS_{w_i} = p_{y^s} - p_{y^s_{-w_i}} \tag{4}$$

每一轮迭代时，从句子 y^s 中删除重要性分数最高的单词 w_i，当满足以下条件之一时，迭代过程结束。1）最终句子的风格概率小于超参数 α，即 $p_{y^{s,t}} < \alpha$；2）内容词汇的比例小于超参数 β，即 $\frac{|y^{s,t}|}{|y^s|} < \beta$，其中，$y^{s,t}$ 表示第 t 轮迭代，即删除 t 个单词后的句子，$|y^{s,t}|$ 和 $|y^s|$ 分别表示 $y^{s,t}$ 和 y^s 的长度。迭代过程中删除的单词将被添加到集合 a 中，即每个句子经拆分后可以得到二元组 (c, a)，其中 c 为迭代过程结束后的句子。针对风格化语料集 D_s，经解耦这一步后得到包含内容和风格词汇二元组的记忆模块 $M_{D_s}=\{(c_1, a_1), \cdots, (c_m, a_m)\}$，其中 m 为 D_s 中句子的数量。

2.4 检索

对于给定图像 x，先利用图像描述模型[13]为其生成事实描述 c_x，再从记忆模块 M_{D_s} 中检索出与之最匹配的风格词汇，其表达式如下：

$$a^{tgt}=\arg\max_{(c_i,a_i)\in M_{D_s}} Sim(c_x,c_i) \tag{5}$$

其中，Sim 为相似度的度量，本文采用的是余弦相似度，通过计算事实描述c_x和内容c_i的相似度，检索出与c_x最匹配的风格词汇a^{tgt}。具体过程为，将事实描述c_x和风格句子内容c_i分别编码成向量表示，计算两者的余弦相似度值，值越大，c_x与c_i的相似度越高。为了得到文本的向量表示，可以采用单词级别的 TF-IDF 向量，句子中所有单词 GloVe[14] 向量表示的平均值以及句子级别的通用句子编码器（Universal Sentence Encoder）[15] 和 Sentence-Bert[16] 等。我们认为，当事实描述c_x和内容c_i相似度越高时，与c_i同时出现的风格词汇a_i也更适合与c_x组合在一起，得到风格化的描述。对于风格词汇的检索，也可以不对输入图像生成描述，而是直接使用图像来计算相似度，即利用图文匹配模型[17]来进行相似度计算。考虑到利用风格化语料进行训练时，没有对应的图像数据，只能使用文本，因此便没有直接使用图像进行检索。

2.5 生成

相比基于 RNN 的编码器-解码器模型，基于 Transformer 的模型在许多文本生成任务[11,18]上表现出了较大优势，且捕获单词间长距离依赖关系的能力更强[11,19]。本文采用 Transformer 编码器和解码器来生成风格化描述，编码器对检索到的风格词汇和图像的事实描述进行编码，解码器以编码器得到的向量表示和模型当前生成的单词为输入，来预测下一步要生成的单词，生成过程在模型生成结束标志或达到设定的最大长度时结束。

2.5.1 Transformer 模型[11]

Transformer 模型由一系列基本块（Block）组成，这些基本块主要包含多头注意力和前馈神经网络。多头注意力的计算式如下：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{6}$$

$$MHA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(\boldsymbol{head}_1,\cdots,\boldsymbol{head}_h)\boldsymbol{W}^O \tag{7}$$

其中，$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_Q$，$\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}_K$，$\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_V$，$\boldsymbol{head}_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V)$，$\boldsymbol{W}_Q$，$\boldsymbol{W}_K$，$\boldsymbol{W}_V$，$\boldsymbol{W}_i^Q$，$\boldsymbol{W}_i^K$，$\boldsymbol{W}_i^V$ 均为参数矩阵，h 表示注意力头的数量，d_k为向量 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的维度，$\boldsymbol{X}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)$表示输入的 n 个编码向量（模型输入的原始向量由单词嵌入向量和位置向量相加得到）。前馈神经网络的计算式如下：

$$FFN(\boldsymbol{X})=\max(0,\boldsymbol{X}\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{8}$$

其中，$\boldsymbol{W}_1$和$\boldsymbol{W}_2$为参数矩阵，$\boldsymbol{b}_1$，$\boldsymbol{b}_2$为偏置项。在多头注意力和前馈神经网络之后均会进行残差连接和层规范化操作。

2.5.2 编码器

编码器的基本块包含一个多头注意力和一个前馈网络。在 2.4 节中，我们得到了图像的事实描述c_x，并检索到了风格词汇a^{tgt}，需要对这两项进行编码。参考 Devlin 等[20] 和 Sudhakar 等[21] 的方法，在风格词汇a^{tgt}和描述c_x前面分别添加特殊标记 $ATTR$ 和 $CONT$，两者分别示意其后为风格词汇和内容文本，然后将这两部分进行拼接，即编码器的输入为$[ATTR,a^{tgt},CONT,c_x]$，对其进行编码得到向量表示（$\boldsymbol{z}_1$，$\boldsymbol{z}_2$，…，$\boldsymbol{z}_m$），整个编码过程可表示为：

$$(\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_m)=TransformerEncoder([ATTR,a^{tgt},CONT,c_x];\theta_E) \tag{9}$$

其中，θ_E表示编码器中的参数。

2.5.3 解码器

解码器的基本块包含两个多头注意力和一个前馈神经网络，其中对第一个多头注意力进行了遮掩（Mask）操作，以防止在解码单词时使用到后面单词的信息。本文在生成风格化描述时，在输入句子的前面添加了两个特殊标记[Style]和[〈s〉]，[Style]指定了生成描述的风格，[〈s〉]为生成描述的开始标记。解码器以编码器编码得到的向量表示（$\boldsymbol{z}_1$，$\boldsymbol{z}_2$，…，$\boldsymbol{z}_m$）和模型当前生成的单词$y^s_{1,t-1}$为输入，来预测下一个时间步 t 应当生成的单词y^s_t，可表示为：

$$y^s_t=TransformerDecoder([\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_m],y^s_{1,t-1};\theta_D) \tag{10}$$

其中，θ_D表示解码器中的参数。当模型生成结束标记[〈/s〉]或达到设定的最大长度时，整个生成过程结束，得到最终的风格化描述。

2.6 模型训练

模型训练包括预训练和微调。预训练时，利用图像-事实描述数据集D_f来训练文本生成器，对于给定图像 x，图像描述模型为其生成事实描述c_x，然后基于事实描述c_x生成数据集D_f中的描述y^f。文本生成器采用交叉熵进行训练，损失函数公式如下：

$$\mathscr{L}_{xe}=-\log p_{\theta_E,\theta_D}(y^f\mid c_x) \tag{11}$$

风格化描述数量相对较少，而训练一个语言模型需要较多的训练数据，预训练可以帮助文本生成器更好地进行初始化。在微调阶段，利用风格化语料D_s来训练文本生成器，对于语料中的描述y^s，先将其拆分成内容c_{y^s}和风格词汇a_{y^s}，然后从 2.3 节中得到的M_{D_s}中检索出与之匹配的风格词汇 a，将a_{y^s}和 a 随机混合，得到最终的风格词汇a^{tgt}，基于内容c_{y^s}和风格词汇a^{tgt}重新生成描述y^s，损失函数为交叉熵损失，可表示为：

$$\mathscr{L}_{\mathrm{rec}}=-\log p_{\theta_E,\theta_D}(y^s\mid c_{y^s},a^{tgt}) \tag{12}$$

交叉熵损失在一定程度上可以保证生成文本的内容相关性，而对于风格准确性，还需要引入风格损失来帮助文本生成器进行学习。风格损失由预训练的风格分类器 C 提供，文本生成器基于内容c_{y^s}和风格词汇a^{tgt}生成风格描述$\hat{y}^s$，分类器 C 判断生成的文本是否具有[Style]所指定的风格 s，损失函数可表示为：

$$\mathscr{L}_{\mathrm{style}}=-\log p_{\theta_c}(C(\hat{y}^s)=s\mid c_x,a^{tgt}) \tag{13}$$

由于生成风格描述文本$\hat{y}^s$是离散的，因此我们将解码器的输出向量表示作为分类器的输入，而不是以最终生成的离散文本为输入。

3 实验及结果

3.1 数据集

本文采用的图像-事实描述数据集为 MS COCO 和 Flickr 30K。风格化语料包括 Mathews 等[6] 公开的 SentiCap 和 Gan 等[8] 公开的 FlickrStyle 10K。SentiCap 包含从 MS COCO 中选取的 2360 幅图像、5013 个积极描述的句子和 4500 个消极描述的句子。依照数据集的原始划分，对于积极风格，分别

使用 2994 和 2019 个句子进行训练和测试；对于消极风格，分别使用 2991 和 1509 个句子进行训练和测试。FlickrStyle 10K 包含 10000 幅图像，每幅图像包含一个幽默风格和一个浪漫风格的描述句子，但是原作者只公开了训练集中的 7000 幅图像和对应的描述句子，参考 Zhao 等[7]和 Guo 等[3]的划分，将其中 6000 幅图像作为训练数据，剩余的 1000 幅图像用于模型测试。另外，对于每个数据集，从训练集中选取 100 幅图像来组成验证集。

3.2 评价指标

与 Zhao 等[7]和 Guo 等[3]的工作相同，本文主要从两个方面对模型进行评价。1)生成描述与图像的内容相关性，即是否正确描述了输入图像的内容，采用 BLEU-1[22]，BLEU-3[22]，METEOR[23]和 CIDEr[24]这 4 个指标，由公开的标准评价代码来计算；2)生成描述的风格准确性，即是否符合风格标签所指定的风格，使用的指标为风格分类准确率(Style Classification Accuracy，CLS)和困惑度(Perplexity，PPL)。CLS 通过逻辑回归分类器计算某种风格下生成的描述文本中，正确反映所需风格句子的数量和所有生成描述文本数量的比值[7]。对于 PPL，本文使用 SRILM 工具包[25]训练基于三元组的统计语言模型，计算所有生成描述的困惑度，然后取平均值[7]。所有的评价指标，除 PPL 值越小代表模型效果越好之外，其他指标都是取值越大，模型效果越好。

3.3 实验结果分析

本节主要从与基准算法的对比实验、消融实验、参数敏感性分析以及生成的风格化描述结果示例等几个方面展开介绍。

3.3.1 与基准算法的对比实验

为了验证 DRGCap 模型的有效性，本文将其与目前在图像的风格化描述任务上表现较好的模型进行了比较。

(1)NIC[26]：基于编码器-解码器的图像描述模型，该模型使用 MS COCO 数据集中成对的图像-事实描述进行训练。

(2)NIC-FT：使用 4 种风格的成对数据分别对训练完成的 NIC 模型进行微调而得到的模型。

(3)StyleNet[8]：在训练时，首先用成对的图像-事实描述对风格无关的参数进行学习，然后用 4 种风格化的语料对风格相关的参数进行学习。

(4)SF-LSTM[9]：对 LSTM 模块进行了改进，针对每种风格使用成对的数据进行有监督学习。

(5)MSCap[3]：采用对抗训练的思想，使用成对的图像-事实描述和非成对的风格语料进行训练。模型只需训练一次，便可以生成多种风格的描述。

(6)MemCap[7]：单独设计了风格记忆模块，并采用注意力机制对此模块进行风格知识的提取和更新，也支持一次训练生成多种风格的描述。

单风格的图像风格化描述实验结果如表 1 所列，可以发现，本文提出的模型在内容相关性的指标 BLEU-1(B-1)，BLEU-3(B-3)和 METEOR(M)，CIDEr(C)上都优于其他对比方法，这表明使用图像描述模型生成的事实描述对风格化描述的内容相关性有较大帮助。虽然在积极(Pos)、浪漫(Roman)和幽默(Humor)风格上的 CLS 略差于基准算法 MemCap，但均优于其他对比算法，而 PPL 指标也优于其他对比算法，说明 DRGCap 模型可以保证风格化描述的风格准确性。

表 1 单风格的图像风格化描述实验结果

Table 1 Results of single-style image captioning

Method	Style	B-1	B-3	M	C	PPL	CLS
NIC	Pos	47.6	16.3	14.9	55.1	25.6	22.4
	Neg	46.9	16.1	14.8	54.0	25.4	23.2
	Roman	25.1	7.0	10.6	33.0	61.6	24.3
	Humor	25.5	7.2	9.7	33.5	57.1	25.5
NIC-FT	Pos	48.2	17.3	16.6	54.3	20.4	91.3
	Neg	47.3	17.8	16.1	55.4	21.5	89.5
	Roman	26.9	7.5	11.0	35.4	27.7	82.6
	Humor	26.3	7.4	10.2	35.1	31.8	80.1
SF-LSTM	Pos	50.5	19.1	16.6	60.0	—	—
	Neg	50.3	20.1	16.2	59.7	—	—
	Roman	27.8	8.2	11.2	37.5	—	—
	Humor	27.4	8.5	11.0	39.5	—	—
StyleNet	Pos	45.3	12.1	12.1	36.3	24.8	45.2
	Neg	43.7	10.6	10.9	36.6	25.0	56.6
	Roman	13.3	1.5	4.5	7.2	52.9	37.8
	Humor	13.4	0.9	4.3	11.3	48.1	41.9
MemCap	Pos	50.8	17.1	16.6	54.4	13.0	**99.8**
	Neg	48.7	19.6	15.8	60.6	14.6	93.1
	Roman	21.2	4.8	8.4	22.4	14.4	**98.7**
	Humor	19.9	4.3	7.4	19.4	16.4	**98.9**
DRGCap	Pos	**51.6**	**18.8**	**16.9**	**54.8**	**12.7**	97.6
	Neg	**49.3**	**20.1**	**16.4**	**61.2**	**14.1**	**95.4**
	Roman	**22.4**	**5.3**	**8.6**	**25.5**	**14.0**	98.6
	Humor	**20.7**	**4.9**	**7.8**	**24.3**	**15.8**	98.7

多风格的图像风格化描述实验结果与基准算法的比较如表 2 所列，DRGCap 模型 4 种风格上的实验指标 BLEU-1，BLEU-3，METEOR 和 CIDEr 以及 PPL 上均优于基准算法，而在风格准确率(CLS)上，积极(Pos)、消极(Neg)和幽默(Humor) 3 种风格的指标略差于 MemCap 模型，这与单风格的实验结果是一致的，说明 DRGCap 模型不仅可以支持单风格描述生成，也适用于多风格描述生成。

表 2 多风格的图像风格化描述实验结果

Table 2 Results of multi-style image captioning

Method	Style	B-1	B-3	M	C	PPL	CLS
MSCap	Pos	46.9	16.2	16.8	55.3	19.6	92.5
	Neg	45.5	15.4	16.2	51.6	19.2	93.4
	Roman	17.0	2.0	5.4	10.1	20.4	88.7
	Humor	16.3	1.9	5.3	15.2	22.7	91.3
MemCap	Pos	51.1	17.0	16.6	52.8	18.1	**96.1**
	Neg	49.2	18.1	15.7	59.4	18.9	**98.9**
	Roman	19.7	4.0	7.7	19.7	19.7	91.7
	Humor	19.8	4.0	7.2	18.5	17.0	**97.1**
DRGCap	Pos	**51.4**	**18.4**	**17.6**	**56.4**	**17.8**	95.2
	Neg	**49.9**	**19.1**	**16.9**	**59.7**	**18.5**	98.0
	Roman	**21.6**	**4.7**	**8.3**	**21.3**	**19.3**	**94.3**
	Humor	**20.4**	**4.5**	**7.5**	**20.8**	**16.5**	96.5

3.3.2 消融实验

首先对检索阶段的句子向量进行了不同的设置，尝试了以下的变体模型。

(1)TF-IDF：文本向量利用 TF-IDF 向量进行表示。

(2)GloVe：将句子中所有单词用 GloVe 向量进行表示，然后取平均值作为句子的向量表示。

(3)U-S-E：采用 Universal Sentence Encoder 模型对句子进行编码得到向量表示。

(4)Sent-Bert:采用 Sentence-Bert 来计算句子之间的相似度。

(5)w/o R:去掉检索模块进行实验,即对于给定的内容文本,不检索与之相匹配的风格词汇,在生成时,Transformer 编码器的输入只有内容文本。

检索方式的消融实验结果如表 3 所列,在进行风格词汇的检索时,使用 Sentence-Bert 效果最好,使用 Glove 词向量效果最差;去掉检索模块(w/o Retrieve)时,PPL 和 CLS 指标变差,证明了检索模块对风格化描述风格准确性的重要性。

表 3 单风格图像描述检索方式消融实验结果

Table 3 Results of ablation studies for retrieval methods on single-style image captioning

Method	Style	B-3	M	C	PPL	CLS
TF-IDF	Pos	18.5	16.7	56.2	13.4	96.1
	Humor	4.7	7.5	24.0	16.3	97.5
Glove	Pos	18.2	15.8	55.4	14.1	95.7
	Humor	4.3	6.9	23.1	17.2	95.6
U-S-E	Pos	18.1	16.2	55.7	13.8	96.0
	Humor	4.4	7.1	23.5	17.0	95.8
Sent-Bert	Pos	18.8	16.9	56.7	12.7	97.6
	Humor	4.9	7.8	24.3	15.8	98.7
w/o R	Pos	16.3	16.0	49.9	25.9	72.4
	Humor	4.1	6.8	21.5	25.3	59.8
DRGCap	Pos	18.8	16.9	56.7	12.7	97.6
	Humor	4.9	7.8	24.3	15.8	98.7

对于训练时采用的损失函数,我们尝试了以下几种设置。

(1)w/o xe L:不使用图像-事实描述对文本生成器进行预训练,只利用风格化语料来训练。

(2)w/o rec L:去掉了在风格化语料上重新生成原始描述的过程,只使用图像-事实描述对进行训练。

(3)w/o cls L:去掉了在风格化语料上预训练的分类器提供的分类损失。

训练损失消融实验结果如表 4 所列,可以看出,去掉交叉熵损失函数(w/o xe L)时,模型生成文本的内容相关性和风格准确性都受到了影响;去掉重构损失函数(w/o rec L)时,模型生成文本的内容相关性受到的影响较小,PPL 和 CLS 则急剧下降;去掉风格分类器损失(w/o cls L)时,模型的内容相关性有较好的表现,甚至超过了完整的 DRGCap 模型,但 PPL 变差,而 CLS 则急剧下降。

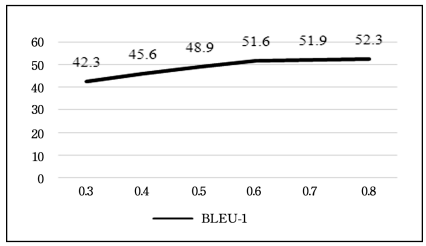
表 4 单风格图像描述训练损失消融实验结果

Table 4 Results of ablation studies for training loss on single-style image captioning

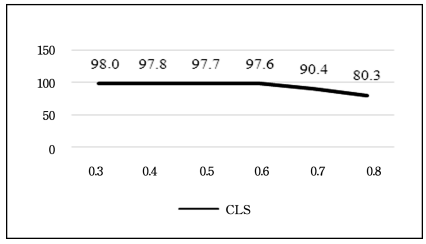
Method	Style	B-3	M	C	PPL	CLS
w/o xe L	Pos	18.3	16.2	56.0	14.2	96.5
	Humor	4.2	7.6	24.5	18.6	96.9
w/o rec L	Pos	18.4	16.5	56.3	22.7	92.5
	Humor	4.3	7.4	25.0	24.2	89.7
w/o cls L	Pos	19.5	17.3	57.4	14.5	69.8
	Humor	4.6	8.1	25.2	17.8	52.3
DRGCap	Pos	18.8	16.9	56.7	12.7	97.6
	Humor	4.9	7.8	24.3	15.8	98.7

3.3.3 参数敏感性分析

本文模型在积极风格上,BLEU-1 和 CLS 指标随风格概率阈值 α 的变化曲线如图 2 所示,可以看出,BLEU-1 值随阈值 α 的增大有所提升,CLS 值则随阈值 α 的增大而减小。从图中的变化可以看出,当 α 取值为 0.8 时,BLEU-1 指标最高,达到了 52.3,此时 CLS 值却下降到了 80.3。通过分析可知,当 α 取值变大时,经过拆分这一步,得到的风格词汇占比很少,即从记忆模块中检索时,可能无法检索出合适的风格词汇,导致生成的风格化描述无法有效融入合适的风格,使 CLS 值下降;而当 α 取值减小时,内容词汇会被识别为风格词汇,即在检索出的风格词汇中可能会包含内容词汇,由于在风格词汇中混入了内容词汇,导致生成的风格化描述在内容相关性上受到影响,BLEU-1 指标值下降。本文结合文本的内容相关性和风格准确性这两方面的考虑,在积极风格语料上进行解耦时,将超参数 α 的取值设置为 0.6。



(a)BLEU-1 随阈值 α 的变化曲线

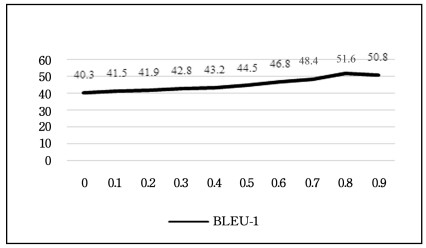


(b)CLS 随阈值 α 的变化曲线

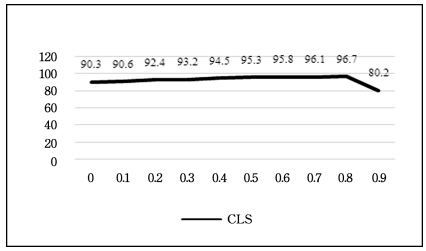
图 2 阈值 α 的参数敏感性分析

Fig. 2 Parametric sensitivity analysis of α

拆分结束时内容词汇所占比例 β 的参数敏感性分析如图 3 所示。



(a)BLEU-1 随阈值 α 的变化曲线



(b)CLS 随阈值 α 的变化曲线

图 3 阈值 β 的参数敏感性分析

Fig. 3 Parametric sensitivity analysis of β

当 β 取值为 0 时,我们将整个风格描述句子作为风格

知识，与内容文本一起输入生成器中生成描述；当 β 取值增大时，风格词汇的个数减少。从图 3 可以看出，当阈值 β 增大时，BLEU-1 和 CLS 值都是先增大，然后减小，β 取 0.8，即内容词汇个数占整个句子长度的比例为 0.8 时，拆分的效果最好。当内容词汇比例占比较低时，此时拆分得到的风格词汇中会混入大量的内容词汇，使模型在描述文本内容相关性和风格准确性上都受到了影响，BLEU-1 和 CLS 值都有所减小；而当内容词汇比例占比很高时（如 0.9），会导致风格词汇缺失，以致风格准确性急剧下降。

3.3.4 结果示例

DRGCap 模型生成的风格描述结果如图 4 所示，对于第一幅图像，指定 Positive 标签时，模型生成的描述为“beautiful cars are sitting in front of interesting buildings”，其中“beautiful”和“interesting”体现了积极风格，“car”“in front of”和“buildings”则体现了内容相关性；指定 Negative 标签时，生成描述中包含了“ugly”，融入了消极风格。在第二幅图像生成的描述中，“sweet”和“sad”分别体现了积极和消极的风格。第三幅和第四幅图像生成的描述中“into snowdogs”和“waiting for airship”体现了幽默的语言风格，“in love”和“show his courage”则体现了浪漫的风格。可以发现，这些描述不仅恰当地融入了风格元素，还比较准确地描述出了图像的视觉元素。

图 4 DRGCap 模型生成的风格化描述文本示例

Fig. 4 Examples of stylized captioning text generated by DRGCap

结束语 本文提出了基于解耦-检索-生成的图像风格化描述生成方法，通过图像描述模型生成的事实描述保证了风格化描述的内容相关性，从风格化语料得到的记忆模块和分类器提供的风格损失保证了生成描述的风格准确性。与目前效果较好的算法的对比实验表明了本文模型的有效性。但使用预训练的图像描述模型不可避免会引入一些噪声，且没有充分利用原始的视觉特征，同时文本风格迁移中的一些研究[19,27]表明解耦操作有时并不是必要的，下一步将尝试使用多模态特征，以非解耦的方式生成风格化描述。

参 考 文 献

[1] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// European Conference on Computer Vision. Cham: Springer, 2014: 740-755.

[2] PLUMMER B A, WANG L, CERVANTES C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 2641-2649.

[3] GUO L, LIU J, YAO P, et al. Mscap: Multi-style image captioning with unpaired stylized text[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4204-4213.

[4] BELL A. Language style as audience design[J]. Language in society, 1984, 13(2): 145-204.

[5] PENNEBAKER J W. The secret life of pronouns [J]. New Scientist, 2011, 211(2828): 42-45.

[6] MATHEWS A, XIE L, HE X. Senticap: Generating image descriptions with sentiments[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2016: 3574-3580.

[7] ZHAO W, WU X, ZHANG X. Memcap: Memorizing style knowledge for image captioning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 12984-12992.

[8] GAN C, GAN Z, HE X, et al. Stylenet: Generating attractive visual captions with styles[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3137-3146.

[9] CHEN T, ZHANG Z, YOU Q, et al. “Factual” or “Emotional”: Stylized Image Captioning with Adaptive Learning and Attention[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 519-535.

[10] CHEN C K, PAN Z, LIU M Y, et al. Unsupervised stylish image description generation via domain layer norm[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 8151-8158.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[12] KIM Y. Convolutional Neural Networks for Sentence Classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1746-1751.

[13] HUANG L, WANG W, CHEN J, et al. Attention on attention

for image captioning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:4634-4643.
[14] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2014:1532-1543.
[15] CER D, YANG Y, KONG S, et al. Universal sentence encoder [J]. arXiv:1803.11175, 2018.
[16] REIMERS N, GUREVYCH I. Sentence-bert: Sentence embeddings using siamese bert-networks [J]. arXiv: 1908.10084, 2019.
[17] LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018:201-216.
[18] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. https://cdn. openai. com/research-covers/language-unsupervised/language_understanding_paper. pdf.
[19] DAI N, LIANG J, QIU X, et al. Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:5997-6007.
[20] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810.04805, 2018.
[21] SUDHAKAR A, UPADHYAY B, MAHESWARAN A. Transforming Delete, Retrieve, Generate Approach for Controlled Text Style Transfer[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2019:3269-3279.
[22] 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.
[23] 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.
[24] VEDANTAM R, LAWRENCE Z C, PARIKH D. Cider: Consensus-based image description evaluation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:4566-4575.
[25] STOLCKE A. SRILM—an extensible language modeling toolkit [C]// Seventh International Conference on Spoken Language Processing. 2002:901-904.
[26] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:3156-3164.
[27] LAMPLE G, SUBRAMANIAN S, SMITH E, et al. Multiple-attribute text rewriting[J]. arXiv:1811.00552, 2018.

CHEN Zhang-hui, born in 1994, postgraduate. His main research interests include big data and data mining.

XIONG Yun, born in 1980, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include data science and data mining.

(责任编辑:李亚辉)