基于跨媒体解纠缠表示学习的风格化图像描述生成

蔺泽浩 李国庭 曾祥极 邓 悦 张 寅 庄越挺
(浙江大学计算机科学与技术学院 杭州 310027)

摘 要 风格化图像描述生成的文本不仅被要求在语义上与给定的图像一致,而且还要与给定的语言风格保持一致.随着神经网络在计算机视觉和自然语言生成领域的技术发展,有关这个主题的最新研究取得了显著进步.但是,神经网络模型作为一种黑盒系统,人类仍然很难理解其隐层空间中参数所代表的风格、事实及它们之间的关系.为了提高对隐层空间中包含的事实内容和语言风格属性的理解以及增强对两者的控制能力,提高神经网络的可控性和可解释性,本文提出了一种使用解纠缠技术的新型风格化图像描述生成模型 Disentangled Stylized Image Caption(DSIC).该模型分别从图像和描述文本中非对齐地学习解纠缠表示,具体使用了两个解纠缠表示学习模块——D-Images 和 D-Captions 来分别学习图像和图像描述中解纠缠的事实信息和风格信息.在推理阶段,DSIC 模型利用图像描述生成解码器以及一种特别设计的基于胶囊网络的信息聚合方法来充分利用先前学习的跨媒体信息表示,并通过直接控制隐层向量来生成目标风格的图像描述.本文在 SentiCap 数据集和 FlickrStyle10K 数据集上进行了相关实验.解纠缠表示学习的实验结果证明了模型解纠缠的有效性,而风格化图像描述生成实验结果则证明了聚合的跨媒体解纠缠表示可以带来更好的风格化图像描述生成性能,相对于对比的风格化图像描述生成模型,本文方法在多个指标上的性能提升了 17%至 86%.

关键词 跨媒体;机器学习;解纠缠表示学习;风格化图像描述生成;自然语言生成
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.02510

A Stylized Image Caption Approach Based on Cross-Media Disentangled Representation Learning

LIN Ze-Hao LI Guo-Dun ZENG Xiang-Ji DENG Yue ZHANG Yin ZHUANG Yue-Ting
(*Department of Computer Science and Technology, Zhejiang University, Hangzhou* 310027)

Abstract The task of stylized image caption aims to generate a natural language description that is semantically related to a given image and consistent with a given linguistic style. Both requirements make this task significantly more difficult than the traditional image caption task. However, with the availability of the large-scale image-text corpora and advances in deep learning techniques of computer vision and natural language processing, stylized image caption research has made significant advances in recent years. Widely adopted neural networks have demonstrated their powerful abilities to handle the complexities and challenges of the stylized image caption task. A typical stylized image caption model is usually an encoder-decoder architecture. The model inputs go through many layers of non-linear transformations, e.g. ReLU layer in the Convolutional Neural Networks (CNNs), to yield latent representations. This makes the latent representations and parameters of model lack interpretability and controllability, which can restrict the understanding

收稿日期:2021-05-21;在线发布日期:2022-09-15.本课题得到国家自然科学基金(62072399,61402403,U19B2042)、中国工程科技知识中心、数字图书馆教育部工程研究中心、中国工程科技数据和知识技术研究中心、中央高校基本科研业务费和百度人工智能课题基金资助.**蔺泽浩**,博士研究生,中国计算机学会(CCF)会员,主要研究方向为自然语言处理、对话系统、机器学习、多模态.E-mail:georgelin@zju.edu.cn.**李国庭**,硕士研究生,主要研究方向为自然语言处理、机器学习.**曾祥极**,硕士研究生,主要研究方向为自然语言处理、反事实推理、机器学习.**邓 悦**,博士研究生,主要研究方向为强化学习、机器学习.**张 寅**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究方向为人工智能、智能问答系统、知识计算、数字图书馆.E-mail:zhangyin98@zju.edu.cn.**庄越挺**,博士,教授,中国计算机学会(CCF)会员,教育部长江学者特聘教授,国家杰出青年科学基金入选者,中国人工智能学会(CAAI)会士,主要研究领域为人工智能、跨媒体计算、数字图书馆.

of this task and its further improvement. In this paper, we focus on the problem of understanding and controlling the latent representations of linguistic style and factual content in stylized image caption models by learning disentangled representations. Existing disentanglement methods mainly work on single modal data, such as computer vision or natural language processing. However, in stylized image caption, there are two types of media, images and texts, involved to learn a representation that is faithful to the underlying data structure. How to disentangle the latent space of cross-media data still needs to be explored. Inspired by the successful applications of disentangled representation learning on Computer Vision and Natural Language Processing, we propose a novel approach, Disentangled Stylized Image Caption (DSIC), to learn the disentangled representations on unparallel cross-media data. With the help of the VAE framework, two latent space filter modules, style filter and fact filter, are designed to enhance the disentangling performance. These filters slice the latent representation to different segments. Each filter is going to retain the style-specific or fact-specific information in the image, by minimizing the proposed auxiliary classifier loss, and screen out other irrelevant information by another auxiliary discriminator loss. Concretely, we use two modules, D-Images and D-Captions, to disentangle the stylistic and factual latent information in the images and captions respectively. To fully utilize obtained cross-media disentangled latent information from both images and captions, we adopt an aggregation method using capsule network with routing-by-agreement. This makes it possible for the LSTM based caption generator to generate stylized captions with target linguistic styles by directly controlling the learnt latent vectors. To validate the effectiveness of our approach, we conduct two groups of experiments: the disentanglement performance test and the stylized image caption test, on two popular public image caption datasets, SentiCap and FlickrStyle10K. Experimental results for disentanglement performance show that our model can successfully disentangle the stylistic and factual information and reveal that style information existing in both human beings' experience and images themselves. Experimental results on stylized image caption datasets show that our model significantly outperforms the competitive baseline models and prove that the aggregated cross-media disentangled representations lead to around 17% to 86% improvements in terms of multiple performance metrics for stylized image caption.

Keywords　cross-media; machine learning; disentangled representation learning; stylized image caption; natural language generation

1　引　言

风格化图像描述生成(Stylized Image Caption)[1-4]任务旨在生成与给定图像语义相关且与给定语言风格一致的自然语言描述.这两个需求使得该任务相对于传统的图像描述生成任务困难得多.随着大规模图像-文本跨媒体语料库的出现以及深度学习技术在计算机视觉和自然语言处理等领域的发展,风格化图像描述生成在过去几年取得了很大的进步.在现有的方法中,已经被广泛采用的神经网络显现出其强大的能力来应对风格化的图像描述生成任务中的复杂性和挑战[5-6].

如图 1 上半部分所示,一个典型的风格化图像描述生成模型通常是一个编码器-解码器(Encoder-Decoder)架构.模型首先采用编码器将给定的图像转换为向量化的隐层表示(Vectorized Latent Representation),然后使用解码器将该隐层表示生成带有目标风格的风格化自然语言描述.然而,模型的中间状态是经过许多层非线性变换处理的,例如卷积神经网络(Convolutional Neural Networks,CNNs)中的 ReLU 层.这使得隐层表示和模型参数缺乏可解释性和可控性,可能会限制对该任务的理解和进一步完善.

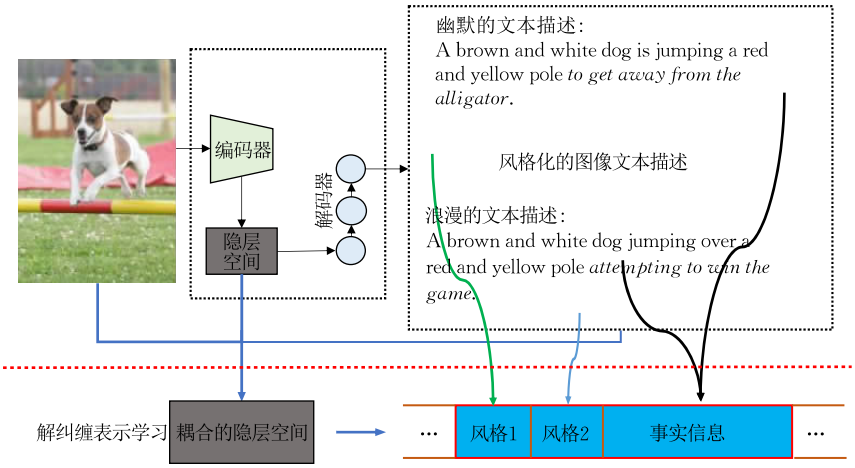


图 1 一个风格化图像文本描述生成的例子(在图像上半部分的端到端模型中,隐层空间的信息的含义几乎无法被人类所理解.而图像下半部分通过解纠缠表示学习能一定程度提升对跨媒体模型的耦合的隐层空间的可控性和可解释性)

本文主要研究了如何利用解纠缠表示学习(Disentangled Representation Learning)技术来理解和控制风格化图像描述生成模型中语言风格和事实内容的隐层表示(如图 1 下半部分所示).现有的解纠缠方法主要工作在单一类型的数据上,如在计算机视觉中[7]或在自然语言处理中[8].然而,在风格化的图像描述生成任务中,为了学习一种忠实于底层数据结构的表示,需要使用图像和文本这两种类型的媒体[9].如何学习多媒体模型的隐层空间的解纠缠表示,并将其运用到下游任务中,还有待探索.

受计算机视觉和自然语言处理中解纠缠表示学习成功应用的启发,本文提出了一种全新的方法,Disentangled Stylized Image Caption(DSIC),以学习在跨媒体数据上的解纠缠表示.在变分自编码器(Variational Autoencoders,VAE)[10]框架的帮助下,本文设计了两个隐层空间过滤器,分别为风格过滤器(Style Filter)和事实过滤器(Fact Filter),以提高解纠缠性能.这些过滤器将隐层表示切片到不同的片段,如图 2 子图(a)所示.每个过滤器将通过最小化辅助分类器损失函数来保留所选择的信息,例如图像中隐含的风格信息和事实信息,并通过另一个辅助鉴别器的损失函数来剔除其他不相关的信息.为了充分利用之前从图像和描述文本中获得的跨媒体解纠缠隐层信息,本文采用了基于协议路由(routing-by-agreement)的胶囊网络[11]聚合方法.胶囊网络是由 Hinton 等人提出的认为未来可以替换传统神经网络的一种新的神经网络,具有更强的可解释性且更符合人类神经元的原理.这使得基于长短期记忆网络(Long Short-Term Memory,LSTM)[12]的图像描述生成器可以通过直接控制学习到的隐层向量生成带有目标语言风格的描述文本.LSTM 是一种循环神经网络(Recurrent Neural Network,RNN),相对于原始的 RNN(Vanilla RNN)[13]更具鲁棒性.

最后,为了验证该模型的有效性,本文在两个流行的公开图像描述数据集上进行了两类实验:解纠缠性能测试和风格化图像描述生成性能测试.实验结果表明,本文的模型能够成功地分离出风格信息和事实信息,揭示了风格信息既存在于人的经验中,也存在于图像本身中.对风格化图像描述生成的实验结果表明,本文的解纠缠表示学习有利于图像的可解释性、可控性,并提高了下游任务的性能.

综上所述,本文的贡献主要为:

(1)本文首次将解纠缠表示学习技术引入了风格化图像描述生成中,以更好地理解隐层空间中包含的风格和事实信息.

(2)本文提出了一种聚合方法,能够更加有效地利用图像和描述文本的解纠缠表示,使生成模型能够充分利用跨媒体数据中的风格和事实信息.

(3)通过模型以及消融实验还发现,描述文本的语言风格不仅仅和文本(即人类的经验和想象)有关,同样和图像本身的视觉信息有关.

(4)多项实验结果显示 DSIC 模型相对于现有的先进方法(state-of-the-art)具有更加优秀的生成效果(在多个指标上的结果超过了目前最佳风格化图像描述生成模型 17%至 86%).这也说明了解纠缠表示学习能够给风格化描述文本生成模型带来更好的性能提升.此外,本文还将模型与代码开放①以供学术界及工业界后续研究.

① 实验模型与代码. https://1drv.ms/u/s!AjA1nukw8Z0WlsRDlhmap1LJTrHpzg?e=IHgkw3

本文在第 2 节介绍风格化图像描述生成和解纠缠表示学习的相关工作；第 3 节介绍本文提出的基于跨媒体解纠缠表示学习的风格化图像描述生成方法；第 4 节和第 5 节说明实验设置和对实验结果的分析以及相应的结论.

2　相关工作

本文研究基于解纠缠表示学习的风格化图像描述生成方法，本节将分别介绍风格化图像描述生成和解纠缠表示学习领域的相关工作.

2.1　风格化图像描述生成

图像描述生成[14-17]是一个非常经典且热门的话题.它处在计算机视觉和自然语言处理技术交叉任务中最基本和最前沿的领域.最近，新颖的图像描述生成模型[18-20]会利用预训练模型从视觉隐层空间或跨媒体(图像和文本)隐层空间中生成自然语言的图像描述的方式.这些方法通常是先利用类似 ResNet[21]这样的图像预训练模型来分析图像的视觉内容，然后将分析得到的隐层表示通过自然语言预训练模型直接生成图像描述[22-24].

风格化图像描述生成任务[25]则更进一步，需要在生成图像描述的过程中，同时附加指定的语言风格：例如幽默或者浪漫的语言风格.一些工作[1,5-6]使用对齐的(一一对应的)图像-描述文本对来训练风格化图像描述生成模型，而有些研究工作[26]则更进一步，采用非对齐的风格化文本训练数据.StyleNet 模型[27]中提出了一种通过分解输入的权重矩阵来抽取包含特定风格的因子的特征矩阵，以生成具有幽默或者浪漫风格的图像描述.这个工作主要关注点为如何捕捉多个语言风格信息.类似的，Chen 等人[5]提出的模型通过学习两组不同的矩阵来分别抽取事实和风格化的语言知识.与前述工作不同，You 等人[28]的工作则采取了另外一种方法.他们将情感信息注入图像描述模型中并通过在半监督模式下提供不同的风格情感标签来控制最终生成模型的目标情感.这种方式首次探索了如何通过非对齐的风格化图像文本数据(即图像和描述文本这两组数据之间并不是对应的)构建风格化图像描述生成模型.生成对抗网络[29]也应用在了该任务上.Nezami 等人[30]提出了一种名为 ATTEND-GAN 的图像描述生成模型.该模型采用注意力机制的图像描述生成器以生成与图像高度语义相关的描述，并采用鉴别器和对抗训练机制使得生成描述具有更类似人类的风格模式.SemStyle[26]则研究了一种不需要任何对应的图像信息，即可从大批量的风格化文本语料库中直接学习和视觉相关的风格特征信息的方法.

2.2　解纠缠表示学习

虽然深度学习的进步显著提高了风格化图像描述生成的表现，但如何理解、解释神经网络的隐层空间含义仍需探索.近年来，许多工作[31-33]试图对神经网络中的因子进行解纠缠，以提高模型的可控性和可解释性.Higgins 等人[9]提出了一种原则性的解决方案，即通过关注世界的转换属性，可以发现解纠缠的特征.Locatello 等人[7]严谨地审视了该领域的最新进展，并提出了对当前领域一些常见假设的挑战.他们从理论上首次证明了，如果模型和数据都没有归纳偏差和先验知识，那么通过完全的无监督学习训练解纠缠表示从根本上是不可能的.Li 等人[34]提出了一种全新的，被称作 Semi-supervised Disentangled VAE(SDVAE)的模型.该模型能够将输入的数据编码为解纠缠的表示形式和不可解释的表示形式，在等式约束下，类别信息等监督信号会被用来规范化训练解纠缠的表示.Litany 等人[35]通过监督学习到的线性变换操作，来鼓励单一变量线性化.

至于解纠缠表示学习在下游任务中的应用，van Steenkiste 等人[36]评估了 360 个最新的无监督解纠缠表示方法的可靠性和性能，结果显示了解纠缠表示学习的确可以在下游任务上带来更好的性能提升.Jain 等人[37]提出了一种用于学习对文本进行解纠缠表示的方法，该方法针对不同和互补的方面进行编码，目的是提供更加有效的模型可迁移性和可解释性.John 等人[8]首次解决了在非平行文本风格迁移任务中风格和内容的隐层表示解纠缠学习的难题.该方法和之前的风格迁移模型相比，在风格迁移的准确率、内容保持和语言的流畅性方面都取得了更好的表现.当前的解纠缠表示研究通常针对单一媒体，但是如何同时对图像和文本等跨媒体数据进行解纠缠表示，仍然亟待探究.

3　方　法

3.1　方法概述

本文提出了一种基于变分自编码器(Variational Autoencoder，VAE)[7,10,38]架构的模型，Disentangled Stylized Image Caption(DSIC)，以学习在跨媒体数据上的解纠缠表示.变分自编码器在这种类型的任务上相对于自编码器具有更好的性能表现[9].如图 2 所示，DSIC 首先对图像和描述文本进行编码，并在其隐层空间上分别进行解纠缠表示学习.

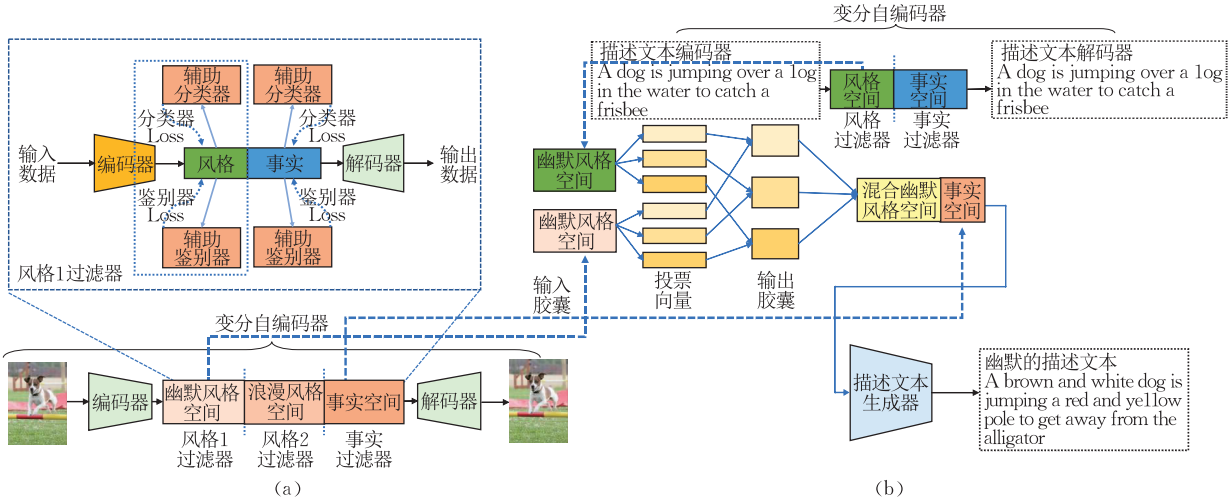


图 2 模型概览((a)对图像和描述文本的隐层空间表示进行解纠缠处理;(b)将之前从图像中和描述文本中获得的跨媒体风格向量进行聚合,并结合事实信息利用描述文本生成器生成风格化的图像描述)

在训练变分自编码器架构的过程中,本文提出了两个隐层空间过滤器:分别为风格过滤器和事实过滤器,来学习隐层空间的解纠缠表示.该部分将会在第 3.3 节中详细描述.风格过滤器的主要目标是在保留隐层表示中的风格相关信息的同时去除掉其他所有无关信息例如事实信息.相对应的,事实过滤器则是要保留所有和事实相关的信息而去除掉其他无关的信息例如风格信息.通过这样的方式,DSIC 模型可以通过将隐层空间分割成不同的片段(如图 2 所示的风格空间和事实空间),学习到一个能够充分反映原数据的结构[9]的表示.在两个过滤器的帮助下,DSIC 模型不仅能够对不同空间包含的信息进行解释,还能够通过人工控制隐层空间,例如替换或组合不同的片段,达到控制目标数据生成的目的.

此外,风格化图像描述生成需要利用事实信息和风格信息.对于事实信息而言,DSIC 以图像中解纠缠出的事实信息作为基础.为了符合实际需求,本文训练和测试采用非对齐的图像和描述数据,因此数据集中描述文本解纠缠出的事实信息在此处并未采用.对于风格信息而言,DSIC 既利用了图像中与风格相关的信息,也引入了大量文本中与人类经验相关的风格信息.

然后,DSIC 还采取了一种利用协议路由(routing-by-agreement)胶囊网络(Capsule Network)的聚合(aggregation)方式,整合之前学习到的图像和描述文本的风格信息解纠缠表示.该部分将会在 3.4 节中详细描述.最后,如 3.5 节中所述,DSIC 使用一个独立训练的描述文本生成器结合之前整合的信息来完成风格化图像描述生成任务.

3.2 变分自编码器

基于变分自编码器的 DSIC 模型旨在学习如何对图像和文本进行编码及信息重构.自编码器(Auto Encoder)是一类在半监督学习和非监督学习中使用的人工神经网络,通过编码器(Encoder)将原先的数据压缩为低维向量,然后通过解码器(Decoder)把低维向量还原为原来的数据.而变分自编码器[10](Variational Autoencoder,VAE)模型则是一种引入隐变量 z 的生成模型,使得编码器的输出结果能对应到目标分布的均值和方差.此处我们主要参考“Auto-Encoding Variational Bayes”一文理论推导部分,建立适用于神经网络的 VAE 模型.

具体而言,在引入隐变量 z 后,VAE 定义了原始数据对数概率密度分布为

$$\log p(x)=\mathbf{E}_{z\sim q(z|x)}[\log p(x)] \tag{1}$$

其中 x 表示输入的文本或者图像,z 表示输入数据的隐层表示,$q(z|x)$ 表示编码器的概率分布,$p(x|z)$ 表示解码器的概率分布.经过贝叶斯法则等转换:

$$\log p(x)=\mathbf{E}_{q(z|x)}[\log p(x|z)]-\mathrm{KL}(q(z|x)|p(z))+\mathrm{KL}(q(z|x)|p(z|x)) \tag{2}$$

其中 KL 指的是 KL 散度(Kullback-Leibler Divergence,KLD),因为 $\mathrm{KL}(q(z|x)|p(z|x))$ 大于等于 0,因此公式的前两项可以作为原始数据概率密度的下界(lower bound),在这里称为证据下界(Evidence Lower Bound,ELBO),即 $\log p(x)\geq\mathbf{E}_{q(z|x)}[\log p(x|z)]-\mathrm{KL}(q(z|x)|p(z))$.最终变分自编码器的训练目标是转化为最大化证据下界,即最小化如下损失函数:

$$\mathcal{L}_{vae}=-\mathbf{E}_{q(z|x)}[\log p(x|z)]+\lambda_{kl}\mathrm{KL}(q(z|x)|p(z)) \tag{3}$$

其中 λ_{kl} 是用于平衡重构损失和 KL 项的超参数. 该式由一个重建损失项和 KL 正则项组成. 式(3)的第一部分表示重建损失，即期望负对数似然损失，其目标是鼓励解码器能够更好地重建原始图像或文本. VAE 模型参数记为 θ_{vae}，$p(z)$ 假设为标准正态分布 $\mathcal{N}(0,1)$，z 的后验分布 $q(z|x)$ 为 $\mathcal{N}(\mu,\mathrm{diag}\sigma^2)$，参数 μ 和 σ^2 由编码器预测得到. 根据编码器输出的 μ 和 σ^2，利用重参数化技巧(Reparameterization Trick)[10]，可以得到隐层空间的向量表示 $\boldsymbol{z}$，即：

$$\boldsymbol{z}=\mu+\sigma\cdot\epsilon,\ \epsilon\sim N(0,I) \tag{4}$$

其中 ϵ 是符合标准正态分布的随机噪声向量.

3.3　风格与事实的解纠缠

对于解纠缠表示学习，本文设计了两种损失函数：辅助分类损失和辅助鉴别损失，并应用到两种过滤器上，来将隐层空间 z 分离为分别包含且仅包含风格和事实信息的空间切片. 每一种过滤器由相应的分类器和鉴别器组成.

通过辅助分类损失函数，期望捕捉到每个隐层空间相应的风格或事实信息. 具体而言，DSIC 在每一个风格空间 $z_{s_*}\in\{z_{s_1},z_{s_2}\}$ 上应用风格过滤器的分类器来预测风格标签 $l(\cdot)$. 具体的，$l(\cdot)$ 代表在样本上预测风格的分布向量. 使用解纠缠表示学习相关方法[8]，将分类器设计为通过最小化预测分布和真实分布之间的交叉熵来进行训练，如式(5)所示.

$$\mathcal{L}_{cls(s)}=-\sum_{s\in\text{labels}}t(s)\log p(l(s)|z_{s_*};\theta_{cls(s)}) \tag{5}$$

其中 $t(\cdot)$ 表示为样本对应的真实风格分布 one-hot 向量. $p(l|z_{s_*})=\mathrm{Softmax}(W_{cls(s)}z_{s_*}+b_{cls(s)})$ 是预测风格的分布向量，分类器参数记为 $\theta_{cls(s)}=[W_{cls(s)};b_{cls(s)}]$. 同样地，在事实空间中 DSIC 应用分类器来预测事实性描述的词袋分布(Bag-of-Words，BoW)，其训练目标为

$$\mathcal{L}_{cls(f)}=-\sum_{w\in\mathcal{V}^f}t_w^f\log p(w|z_f;\theta_{cls(f)}) \tag{6}$$

类似的，式(6)中，t_w^f 是事实性描述的真实词袋的 BoW 分布，$p(w|z_f)$ 是事实性描述的预测词袋的 BoW 分布，$\mathcal{V}^f$ 是去除风格词汇和停用词的事实性词表，分类器参数记为 $\theta_{cls(f)}=[W_{cls(f)};b_{cls(f)}]$.

通过辅助鉴别损失函数，每个隐层空间期望不包含其他空间的信息. 具体而言，本文首先训练一种对抗鉴别器去鉴别其他空间是否包含当前空间信息. 其中，风格过滤器中的鉴别器旨在基于事实空间去预测风格化描述的词袋分布，而事实过滤器中的鉴别器则基于整个风格空间 $z_s=[z_{s_1};z_{s_2}]$ 去预测事实性描述的词袋分布. 鉴别器的训练目标如式(7)、(8)所示.

$$\mathcal{L}_{dis(s)}=-\sum_{w\in\mathcal{V}^s}t_w^s\log p(w|z_f;\theta_{dis(s)}) \tag{7}$$

$$\mathcal{L}_{dis(f)}=-\sum_{w\in\mathcal{V}^f}t_w^f\log p(w|z_s;\theta_{dis(f)}) \tag{8}$$

$\mathcal{V}^s$ 是去除事实性词汇和停用词的风格词表，t_w^f 是事实性描述的真实词袋分布，t_w^s 是风格性描述的真实词袋 BoW 分布. 鉴别器的训练步骤对应算法 1(见后文)中的第 2 到 3 行和第 7 到 8 行.

在上述鉴别器训练好以后，借助对抗训练的思想，VAE 通过最大化信息熵的方式来学习如何“欺骗”两种过滤器中的鉴别器，即最小化如式(9)、(10)所示的损失函数.

$$\mathcal{L}_{adv(s)}=\sum_{w\in\mathcal{V}^s}p(w|z_f)\log p(w|z_f) \tag{9}$$

$$\mathcal{L}_{adv(f)}=\sum_{w\in\mathcal{V}^f}p(w|z_s)\log p(w|z_s) \tag{10}$$

类似式(6)，其中 $p(w|z_s)$ 为风格化描述的预测词袋的 BoW 分布.

在训练 VAE 的过程中，之前训练好的两种鉴别器的参数 $\theta_{dis(s)}$ 和 $\theta_{dis(f)}$ 均不更新. 训练步骤对应算法 1(见后文)中的第 4 行和第 9 行.

3.3.1　在图像上的解纠缠(D-Images)

首先，DSIC 通过图像编码器 $\mathcal{E}$ 来获取输入图像 x 的视觉表示. 而后，通过两个变换 $f(\cdot)$ 和 $g(\cdot)$ 来对视觉表示进行编码和重构，即从视觉表示映射为隐层表示，再重构回原有视觉表示. 重构损失被定义为 $|\mathcal{E}(x)-g(f(\mathcal{E}(x)))|$，其中 $|\cdot|$ 表示 l_2 范数. 因此式(3)被改写为

$$\mathcal{L}_{vae}^{img}=|\mathcal{E}(x)-g(f(\mathcal{E}(x)))|+\lambda_{kl}\mathrm{KL}(q(z|x)|p(z)) \tag{11}$$

具体地，依据先前工作[39]，这里采用 ResNet-152[21] 作为图像编码器 $\mathcal{E}$. $f(\cdot)$ 和 $g(\cdot)$ 分别由两层全连接层组成，并对第一层的输出应用非线性函数 ReLU. 如图 2 所示，为了从图像的隐层空间中解纠缠出风格和事实信息，而输入图像拥有两种风格化描述和一种事实性描述，DSIC 将两个过滤器应用在三个空间切片(z_{s_1}，z_{s_2} 和 z_f)上，其初始时直接从图像隐层空间中切片得到. 每一个过滤器均拥有一个辅助分类损失和一个辅助鉴别损失.

由此，针对图像的解纠缠模块(记为 D-Images)的损失函数 $\mathcal{L}_{ovr}^{img}$ 如式(12)所示.

$$\begin{aligned}\mathcal{L}_{ovr}^{img}=&\mathcal{L}_{vae}^{img}+\lambda_{cls(s)}\mathcal{L}_{cls(s)}^{img}+\lambda_{adv(s)}^{img}\mathcal{L}_{adv(s)}^{img}+\\&\lambda_{cls(f)}^{img}\mathcal{L}_{cls(f)}^{img}+\lambda_{adv(f)}^{img}\mathcal{L}_{adv(f)}^{img}\end{aligned} \tag{12}$$

其中 λ_*^{img} 表示调节每一种损失函数权重的超参数.

3.3.2 在描述文本上的解纠缠(D-Captions)

对于描述文本,其目标是重构风格化的图像描述 $x=(x_1,x_2,\cdots,x_n)$. 具体地,DSIC 利用 LSTM 编码器来获得 x 的隐层表示 z,然后用 LSTM 解码器来重构 x,即在每个时间步 t 预测单词 x_t 的词表概率 $p(x_t|z,x_1,\cdots,x_{t-1})$. 因此,重构损失被定义为 $-\sum_{t=1}^{n}\log p(x_t|z,x_1,\cdots,x_{t-1})$. 式(3)被改写为

$$\mathcal{L}_{vae}^{cap}=-\sum_{t=1}^{n}\log p(x_t|z,x_1,\cdots,x_{t-1})+\lambda_{kl}\mathrm{KL}(q(z|x)|p(z)) \tag{13}$$

与 D-Images 模块类似,DSIC 利用两种过滤器及对应的辅助损失函数,从描述文本的隐层表示中解纠缠得到风格和事实信息. 不同点在于,对于输入的一条风格化图像描述,其仅表示一种特定的风格,故隐层空间被分割为两个空间切片(z_s,z_f),其中 z_s 表示某一种特定风格的空间切片(s_1 或 s_2).

对于风格过滤器,按照式(5)、(7)和式(9),可以得到 $\mathcal{L}_{cls(s)}^{cap}$,$\mathcal{L}_{dis(s)}^{cap}$ 和 $\mathcal{L}_{adv(s)}^{cap}$. 对于 D-Captions 模块,由于一种输入的风格化描述仅含有一种特定的风格,故式(3)中的 z_{s_*} 即代表整个风格空间 z_s.

对于事实过滤器,其辅助分类损失使得事实空间应包含事实信息,而辅助鉴别损失则防止风格空间包含事实信息. 与 D-Images 模块类似,按照式(4)、(6)和式(8),可以得到 $\mathcal{L}_{cls(f)}^{cap}$,$\mathcal{L}_{dis(f)}^{cap}$ 和 $\mathcal{L}_{adv(f)}^{cap}$,从而根据式(10)导出 $\mathcal{L}_{ovr}^{cap}$.

3.4 跨媒体解纠缠信息的聚合

利用上述方式训练好的 D-Images 和 D-Captions 模块,DSIC 可以分别得到图像和描述文本隐层空间的解纠缠表示.

对于 D-Captions 模块,利用重参数化技巧,可以从风格化描述的后验分布 $q(z|x)$ 中采样并取平均,获得一个固定的、通用的风格表示 z_s^{cap}(对于风格 s_1,其为 $\boldsymbol{z}_{s_1}^{cap}$;对于风格 s_2,其为 $\boldsymbol{z}_{s_2}^{cap}$,即目标风格的描述文本经 D-Captions 编码的隐层风格表示的均值). 对于 D-Images 模块,DSIC 同样能获得采样的,特定于图像的风格和事实表示:$z_{s_1}^{img}$,$z_{s_2}^{img}$ 和 z_f^{img}.

为了整合上述解纠缠表示,DSIC 采取了一种利用协议路由胶囊网络的聚合方法 AGGREGATE,其包含两个部分:输入胶囊和输出胶囊(input capsule and output capsule). 输入胶囊包含两个胶囊:$\Omega_1=\boldsymbol{U}z_{s_1}^{img}$ 和 $\Omega_2=\boldsymbol{V}\boldsymbol{z}_{s_1}^{cap}$,其中 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 为可学习的矩阵参数. 输出胶囊中的 n 个胶囊则表示聚合的跨媒体表示的 n 个部分. 每一个输入胶囊 Ω_i 拥有 n 个"投票向量(vote vector)"$\{\boldsymbol{A}_{i1},\boldsymbol{A}_{i2},\cdots,\boldsymbol{A}_{in}\}$,以表示从图像或描述文本中提取的风格信息对输出胶囊的贡献,具体的第 j 个投票向量表示为

$$\boldsymbol{A}_{ij}=\Omega_i W_{ij},\ i=\{1,2\}\ \text{and}\ j=[1,n] \tag{14}$$

每一个输出胶囊 Ω_j^{out} 被定义为

$$\Omega_j^{\text{out}}=\frac{\sum_{i=1}^{2}C_{ij}A_{ij}}{\sum_{i=1}^{2}C_{ij}} \tag{15}$$

其中,耦合系数 C_{ij} $\left(\sum_{j=1}^{n}C_{ij}=1\right)$衡量了 A_{ij} 和 Ω_j^{out} 之间的信息传递量,计算方式如下:

$$C_{ij}=\frac{\exp(B_{ij})}{\sum_{k=1}^{n}\exp(B_{ik})} \tag{16}$$

其中,B_{ij} 衡量了输入胶囊 Ω_i 和输出胶囊 Ω_j^{out} 之间的耦合度,其被初始化为 0 并由式(17)进行更新.

$$B_{ij}\leftarrow B_{ij}+\Omega_j^{\text{out}}\cdot A_{ij} \tag{17}$$

随后,应用非线性压缩函数将输出胶囊的长度映射到 0 和 1 之间,以表征概率.

$$\Omega_j^{\text{out}}=\frac{|\Omega_j^{\text{out}}|^2}{1+|\Omega_j^{\text{out}}|^2}\frac{\Omega_j^{\text{out}}}{|\Omega_j^{\text{out}}|} \tag{18}$$

最后,将 n 个输出胶囊连接(concatenate)起来,从而形成聚合的跨媒体风格表示 z^{s_1},与事实表示 z_f^{img} 连接起来,一起被输入到描述文本生成器中,从而生成带风格 s_1 的风格化描述(风格 s_2 同理).

3.5 风格化图像描述生成

基于上述模块,本文提出了一个针对聚合方式特别设计的解码器来生成目标风格的描述.

具体而言,针对一张输入图像 x^*,令 $q(z_{s_1}|x^*)$,$q(z_{s_2}|x^*)$ 和 $q(z_f|x^*)$ 分别表示风格 z_{s_1},风格 z_{s_2} 和事实 z_f 三个空间切片上的后验分布. 利用重参数化技巧,DSIC 可以从 D-Images 中采样得到如下向量:

$$\begin{aligned}\boldsymbol{z}_{s_1}^*&\sim q(z_{s_1}|x^*),\\ \boldsymbol{z}_{s_2}^*&\sim q(z_{s_2}|x^*),\\ \boldsymbol{z}_f^*&\sim q(z_f|x^*)\end{aligned} \tag{19}$$

随后,将来源于图像的三个向量和从描述文本中学到的固定风格向量 $\boldsymbol{z}_{s_1}^{cap}$,$\boldsymbol{z}_{s_2}^{cap}$ 输入 AGGREGATE 模块,获得聚合的跨媒体风格表示 z^{s_1} 和 z^{s_2}.

在训练阶段,如图 2 子图(b)所示,z^{s_1} 和 z^{s_2} 分别和特定于图像的事实表示 $\boldsymbol{z}_f^*$ 连接起来,输入到描述文本生成器 $\mathcal{D}$ 中(这里是 LSTM)执行贪心解码,分别生成带有风格 s_1 和风格 s_2 的图像描述. 在测试阶

段，通过上述步骤得到的 z^{s_1} 和 z^{s_2} 分别和特定于图像的事实表示 z_f^* 连接起来，输入训练好的描述文本生成器中执行大小为 5 的集束搜索.

3.6　训练和测试步骤

DSIC 模型由 VAE 架构的 D-Images 模块和 D-Captions 模块，以及 LSTM 架构的描述文本生成器 $\mathcal{D}$ 组成. 该模型的详细训练步骤如算法 1 所示，而详细的测试步骤如算法 2 所示.

算法 1.　模型训练过程.

输入：训练数据集 $T=\{(x^i,y_{s_1}^i,y_{s_2}^i,y_f^i)|_{i=1}^N\}$，其中 x^i 表示第 i 张图像，$y_{s_1}^i$ 表示对应的风格为 s_1 的描述文本，$y_{s_2}^i$ 表示对应的风格为 s_2 的描述文本，y_f^i 表示对应的事实性描述文本，用于构建事实性描述的词袋分布. 数据集大小为 N

输出：训练好的 DSIC 模型

BEGIN

//图像上的解纠缠表示学习(D-Images)

1. FOR 小批量数据 DO
2. 　最小化损失 $\mathcal{L}_{dis(s)}^{img}$ 来优化 D-Images 的参数 $\theta_{dis(s)}^{img}$；
3. 　最小化损失 $\mathcal{L}_{dis(f)}^{img}$ 来优化 D-Images 的参数 $\theta_{dis(f)}^{img}$；
4. 　最小化损失 $\mathcal{L}_{ovr}^{img}$ 来优化 D-Images 的参数 θ_{vae}^{img}，$\theta_{adv(s)}^{img}$ 和 $\theta_{adv(f)}^{img}$；
5. DONE

//描述文本上的解纠缠表示学习(D-Captions)

6. FOR 小批量数据 DO
7. 　最小化损失 $\mathcal{L}_{dis(s)}^{cap}$ 来优化 D-Captions 的参数 $\theta_{dis(s)}^{cap}$；
8. 　最小化损失 $\mathcal{L}_{dis(f)}^{cap}$ 来优化 D-Captions 的参数 $\theta_{dis(f)}^{cap}$；
9. 　最小化损失 $\mathcal{L}_{ovr}^{cap}$ 来优化 D-Captions 的参数 θ_{vae}^{cap}，$\theta_{adv(s)}^{cap}$ 和 $\theta_{adv(f)}^{cap}$；
10. DONE

//利用重参数化技巧，可以从风格化描述的隐层空间中采样并取平均，获取固定的，通用的风格表示 $\boldsymbol{z}_{s_1}^{cap}$ 和 $\boldsymbol{z}_{s_2}^{cap}$

11. FOR $(x^i,y_{s_1}^i,y_{s_2}^i,y_f^i)$ IN T
12. DO
13. 　$z_{s_1}^i \leftarrow$D-Captions$(y_{s_1}^i)$；//重参数化采样
14. 　$z_{s_2}^i \leftarrow$D-Captions$(y_{s_2}^i)$；//重参数化采样
15. DONE
16. $\boldsymbol{z}_{s_1}^{cap}=\frac{1}{N}\sum_{i=1}^{N}z_{s_1}^i$ //在整个训练集上求平均
17. $\boldsymbol{z}_{s_2}^{cap}=\frac{1}{N}\sum_{i=1}^{N}z_{s_2}^i$ //在整个训练集上求平均
18. FOR $(x^i,y_{s_1}^i,y_{s_2}^i,y_f^i)$ IN T
19. DO
20. 　$z_{s_1}^{img},z_{s_2}^{img},z_{s_f}^{img} \leftarrow$D-Images$(x^i)$；//重参数化采样
21. 　$z^{s_1} \leftarrow$AGGREGATE$(z_{s_1}^{img},\boldsymbol{z}_{s_1}^{cap})$；//跨媒体风格信息聚合表示
22. 　$z^{s_2} \leftarrow$AGGREGATE$(z_{s_2}^{img},\boldsymbol{z}_{s_2}^{cap})$；//跨媒体风格信息聚合表示
23. 　$z^{s_1}=[z^{s_1};z_{s_f}^{img}]$ //与事实表示相连接
24. 　$z^{s_2}=[z^{s_2};z_{s_f}^{img}]$ //与事实表示相连接
25. 　$\hat{y}_{s_1} \leftarrow \mathcal{D}(z^{s_1})$；//输入生成器 $\mathcal{D}$ 执行贪心解码
26. 　$\hat{y}_{s_2} \leftarrow \mathcal{D}(z^{s_2})$；//输入生成器 $\mathcal{D}$ 执行贪心解码
27. 利用交叉熵损失函数来优化描述生成器 $\mathcal{D}$
28. DONE

END

算法 2.　模型测试过程.

输入：任意图像 x

输出：针对该图像的带有风格 s_1 和风格 s_2 的描述 y_{s_1} 和 y_{s_2}

BEGIN

1. $z_{s_1}^{img},z_{s_2}^{img},z_{s_f}^{img} \leftarrow$D-Images$(x)$；//重参数化采样
2. $z^{s_1} \leftarrow$AGGREGATE$(z_{s_1}^{img},\boldsymbol{z}_{s_1}^{cap})$；//跨媒体风格信息聚合
3. $z^{s_2} \leftarrow$AGGREGATE$(z_{s_2}^{img},\boldsymbol{z}_{s_2}^{cap})$；//跨媒体风格信息聚合
4. $z^{s_1}=[z^{s_1};z_{s_f}^{img}]$ //与事实表示相连接
5. $z^{s_2}=[z^{s_2};z_{s_f}^{img}]$ //与事实表示相连接
6. $y_{s_1} \leftarrow \mathcal{D}(z^{s_1})$；//输入生成器 $\mathcal{D}$ 执行大小为 5 的集束搜索
7. $y_{s_2} \leftarrow \mathcal{D}(z^{s_2})$；//输入生成器 $\mathcal{D}$ 执行大小为 5 的集束搜索

END

在训练时，首先对图像和文本数据分别进行解纠缠表示学习，即按照小批量训练方式分别优化 D-Images 和 D-Captions 模块. 而后，为了获取固定的，通用的风格表示 $\boldsymbol{z}_{s_1}^{cap}$ 和 $\boldsymbol{z}_{s_2}^{cap}$，DSIC 将训练集中的所有描述文本经 D-Captions 编码的隐层风格表示取平均. 对于训练集的每张图像，将其输入 D-Images，获得其相应的隐层风格表示和事实表示. 利用上文所述 AGGREGATE 聚合方法对跨媒体风格信息聚合后，与事实表示相连接，输入到描述文本生成器 $\mathcal{D}$ 中执行贪心解码，并采用交叉熵损失函数对 $\mathcal{D}$ 进行优化.

在测试时，对于任意输入的一张图像，将其输入训练好的 D-Images，获得其相应的隐层风格表示和事实表示. 再将该隐层风格表示与从训练过程中获得的通用的风格表示，即训练集中目标风格的描述文本经 D-Captions 编码的隐层风格表示的均值，进行聚合，并与事实表示进行连接，输入到训练好的描述生成器 $\mathcal{D}$ 中，执行大小为 5 的集束搜索，从而生成对应风格的图像描述.

4　实　验

本节将主要介绍数据集、两类实验任务的实验

设置、评价指标以及模型参数.

4.1 数据集

本文进行了两类实验,第 4.2 节中的解纠缠表示学习性能测试和第 4.3 节中的风格化图像描述生成性能测试.上述实验在两个流行的风格化图像描述生成数据集 FlickrStyle10K[27] 和 SentiCap[40] 上进行了详细的测试和对比.FlickrStyle10K 数据集是基于 Flickr 30K[41] 图像描述生成数据集上收集和构建而来.原始的 FlickrStyle10K① 数据集包含了 10 000 对图像和风格化描述的文本数据.然而,只有其中的 7000 对数据是对外开放的.因此,本文采取了和 MSCap[42] 以及 MemCap[43] 中相同的做法,即在这 7000 对中抽取 6000 对图像文本作为训练集,并将其余的样本对作为测试集.为了能够得到最佳的超参数设定,本文还从上述 6000 对训练集中随机分割了 100 对作为实验的验证集.本文的另一个数据集为 SentiCap.SentiCap 从 MSCOCO 的验证集部分的图像上额外构建了一个带有情感倾向的描述文本集.其中积极的和消极的子集包含了 1753/236 个图像用来训练/测试.此外还有 236 张图像作为验证.每一个测试/验证的图像都包含了三条积极的和三条消极的描述文本.此外,上述两个数据集的每个图像还拥有对应的五条事实性描述文本,用于构建事实性描述的词袋分布.

4.2 解纠缠表示学习性能测试

4.2.1 实验设置

本文在图像和描述文本上分别进行了解纠缠表示学习的性能测试.这个实验的结果将能说明:(1)图像和描述文本的解纠缠模块是否能够成功学习到解纠缠的隐层空间;(2)图像和文本中的风格和事实信息之间的关系是怎么样的.在这个实验中,本文采用了以下评价指标:

4.2.2 评价指标

(1)Disentangled Space Accuracy(DSA).本文在模型之外独立训练了一个基于逻辑回归的隐层空间风格分类器,然后将这个分类器应用在:①解纠缠的风格隐层空间;②解纠缠的事实隐层空间;③完整的隐层空间.最后,通过计算分类器分类正确的比例,可以得到自动评测的隐层空间风格准确率.值得注意的是,对于 D-Captions 模块而言,每次输入的是某一种风格的描述文本,其完整的隐层空间包含了一种风格和事实.本文将其输入到隐层空间风格分类器中,统计其预测正确风格的比例来判定分类准确度.对于 D-Images 模块而言,虽然其完整的隐层空间包含了两种风格和事实,但是本文实际测试输入图像时,分别将两种风格空间和事实空间结合,输入到隐层空间风格分类器中,统计两者预测为正确风格的比例的均值来判定分类准确度.

(2)t-SNE[44].该实验使用了一个二维 Disentangled Space Accuracy 的 t-SNE 图像来可视化地展示本文提出的模型在不同的隐层空间上的解纠缠效果.

4.3 风格化图像描述生成性能测试

4.3.1 实验设置

为了展示如何将解纠缠的隐层表示应用到下游任务——风格化的图像描述生成任务中,并检验解纠缠表示学习如何影响图像描述生成的效果,本文将该模型和如下几个基线模型、先进(state-of-the-art)的模型以及消融实验模型进行对比:

4.3.1.1 基线模型

(1)CNN+LSTM.一个基于经典的 encoder-decoder 架构的图像描述生成模型,结构简单且具有不错的性能和鲁棒性.该模型利用 CNN 来编码图像,然后用 LSTM 和固定的风格向量作为指示生成目标风格的图像描述.

(2)StyleNet[27].该模型提出了一个名为 factored LSTM 的模型组件.利用该组件,模型能够自动蒸馏出单语言语料库中的风格因子.这个模型能够显式地控制描述文本生成过程中的风格因素,以此产生更具吸引力的具有目标风格的视觉描述文本.本文采用开源代码②复现了该工作.

(3)Style-Factual LSTM(SF-LSTM)[5].该模型提出了一个基于参考事实模型的自适应学习方法,可以在从风格化图像描述中学习时向模型提供事实知识,并可以自适应地计算每个时间步长要提供多少信息.当模型从风格化描述文本中学习时,能够向模型提供事实知识,并且可以自适应地计算每个时间步长要提供多少信息.本文采用作者提供的源代码复现了该工作.

4.3.1.2 消融实验模型

(1)D-Images 和 D-Captions.为本文提出的模型的消融结构.他们分别为仅采用 D-Images 结合 LSTM 解码器或者 D-Captions 模块结合 CNN 编

① FlickrStyle10K 数据集.https://github.com/kacky24/stylenet

② https://github.com/kacky24/stylenet

码器结构，详细结构请见附录.

(2) DSIC-concat 和 DSIC-attention. 在跨媒体解纠缠信息的聚合过程中，为了研究不同聚合方式对最终效果的影响，本文设计了上述两种消融结构. DSIC-concat 是指将来自图像和文本的风格表示 $z_{s_1}^{img}$ 和 $\boldsymbol{z}_{s_1}^{cap}$ 直接拼接起来，然后输入到描述文本生成器中. DSIC-attention 则采用了注意力(attention)机制来聚合跨媒体信息，即通过学习两个媒体之间的注意力权重从而将其聚合起来. 以生成风格 s_1 的跨媒体聚合表示为例，其过程如下所示，其中 W，b 为对应的参数，g 为注意力权重，S 表示 Softmax 函数：

$$\begin{aligned}
&h_{img}=W_{img}z_{s_1}^{img}+b_{img},\\
&h_{cap}=W_{cap}\boldsymbol{z}_{s_1}^{cap}+b_{cap},\\
&g=S(W_{attn}[h_{img};h_{cap}]+b_{attn}),\\
&z^{s_1}=g\cdot h_{img}+(1-g)\cdot h_{cap}
\end{aligned}\tag{20}$$

4.3.2　评价指标

本节将从两个角度评价 DSIC 生成流畅且精确的描述文本的能力，以及生成的描述文本是否具有目标的风格. 本文参照了相关工作[42-43]，利用 BLEU[45]、METEOR[46]、CIDEr[47]、ROUGE-L[48] 和 SPICE[49] 等相关指标测试了生成的句子的语言流畅性和相关性. 具体为：

(1) BLEU[45]. 用于计算 n-gram 精度，n-gram 精度是候选文本中存在于任何参考文本中的 n-gram 的分数. 本文采用 BLEU 来显示生成的描述文本与正确标注的文本的相似性.

(2) METEOR[46]. 是用于机器翻译评估的自动度量，它基于机器翻译和人工翻译参考之间的字母组合匹配的一般概念.

(3) CIDEr[47]. 是一种度量标准，它比各种来源生成的句子中的现有度量标准更好地捕获了人类对共识的判断.

(4) ROUGE-L[48]. 是用于评估机器翻译和文本摘要的一组指标，其中最常用的指标之一是 ROUGE-L，其通过计算最长公共子序列，确定计算机生成的描述文本与人类创建的理想描述文本之间的最长共现序列.

(5) SPICE[49]. 是在创造的场景图上定义的自动描述文本生成评估指标，其基于图的语义表示来编码描述中的对象，属性和关系.

(6) ICSA. 为了测试生成的描述文本是否具备符合预期的风格，本文采用图像描述风格精度(Image Caption Style Accuracy，ICSA)这一指标. 在这个指标中，本文在原始的数据集上独立训练了一个 TextCNN[50] 分类器用以预测输入的图像描述的风格，然后利用这个分类器来测试生成的描述文本是否具备正确的风格并计算准确度. 根据图像描述生成[43] 以及文本风格迁移[8] 等方面的工作，这样的方式可以很好地为风格迁移的效果提供一个可量化的评价.

4.4　实验参数

本文采用深度学习框架 PyTorch 对相关模型进行编码实现，并在 Ubuntu 20.04 系统上用 GPU (NVIDIA GTX 1080Ti) 进行模型的训练和调试. 考虑到 D-Images 模块和 D-Captions 模块结构的不同、图像和文本数据本身的差异性，结合相关论文[8] 和具体实践，本文分别为两个模块设置了不同的超参数.

具体而言，针对 DSIC 模型中的 D-Images 模块，本文采用 ResNet-152 作为图像编码器 $\mathcal{E}$ 来提取 2048 维的视觉表示，并利用两个由两层全连接层组成的变换来编码和重构视觉表示. 第一层的输出还应用了非线性函数 ReLU. 该模块的隐层空间的维度被设置为 1088，其中包括两个大小为 64 维的风格空间和一个大小为 960 维的事实空间. 本文分别使用 RMSProp 优化器和 Adam 优化器来训练鉴别器和其他部件，直到验证集上的效果收敛，初始学习率均为 0.001，其他优化器参数均采用 PyTorch 框架提供的默认参数. 损失函数中的超参数被设置为 $\lambda_{cls(s)}^{img}=10$，$\lambda_{adv(s)}^{img}=0.05$，$\lambda_{cls(f)}^{img}=5$，$\lambda_{adv(f)}^{img}=0.05$ 和 $\lambda_{kl}^{img}=0.01$.

针对 DSIC 模型中的 D-Captions 模块，本文使用 LSTM 作为编码器和解码器. 在模块的隐层空间上，维度为 256 的隐层向量被线性映射到 8 维的风格空间和 128 维的事实空间. 损失函数中的超参数 $\lambda_{cls(s)}^{cap}$，$\lambda_{adv(s)}^{cap}$，$\lambda_{cls(f)}^{cap}$，$\lambda_{adv(f)}^{cap}$ 和 λ_{kl}^{cap} 分别为 10，1，3，0.03 和 0.03. 此外，本文还使用 Word2Vec 来初始化 300 维的词向量并在训练集上对其进行训练.

在聚合方法中，本文使用两个输入胶囊和三个输出胶囊，输入胶囊和投票向量的维度均为 64，路由迭代算法的迭代数(iteration)为 3.

最后，LSTM 被作为描述文本生成器在两个数据集上训练和测试. 在 FlickrStyle10K 数据集上，生成器的词向量维度和隐藏层维度分别为 300 和 500，采用初始学习率为 5e-4 的 Adam 优化器进行训练，批量大小为 64，$dropout$ 值为 0.3. 在 SentiCap 数据

集上，生成器的词向量维度和隐藏层维度分别为400和500，采用初始学习率为1e-4的Adam优化器进行训练，批量大小为32，*dropout*值为0.3.

5 实验结果和分析

为了更好地分析模型，本节将根据实验结果回答以下几个问题：

(1)本文的模型能否学习到图像和描述文本的解纠缠表示？

(2)我们能够从解纠缠表示学习的性能中得到什么？

(3)解纠缠表示学习是如何影响下游任务——风格化图像描述生成的？

(4)跨媒体信息的聚合方式是如何影响风格化图像描述生成性能的？

5.1 本文的模型能否学习到图像和描述文本的解纠缠表示？

如表1所示，本文分别收集了D-Images和D-Captions模块在解纠缠隐层空间上的风格分类效果.可以看到分类器能够非常轻松地预测到D-Images和D-Captions模块学习到的风格隐层空间所代表的风格(93%和94%的分类准确度)，相对应的，同样的分类器几乎完全无法分类事实隐层空间所代表的风格信息(只有52%和53%的准确性，近乎随机).与此同时，该分类器同样能够在完整的隐层空间上获得相近的风格分类效果(94%和98%的分类准确性).从以上结果中，可以观察到仅仅从解纠缠表示的事实隐层空间中分辨出原始的风格信息是几乎不可能的，因为其可靠性仅略微高于随机猜测.然而，在风格隐层空间中的风格分类效果要显著的更高甚至接近100%，同时完整的隐层空间上的分类精度则没有进一步的提升或下降.这些结果表明了模型的解纠缠表示学习的有效性，即风格隐层空间包含风格信息，而事实隐层空间中则不包含任何风格信息.

表1 由Disentangled Space Accuracy(DSA)指标测量的风格分类器在不同的解纠缠空间切片的分类结果

解纠缠空间切片	D-Images	D-Captions
事实空间	0.52	0.53
风格空间	0.93	0.94
完整隐层空间	0.94	0.98

利用t-SNE可视化，本文在图3中展示了在D-Captions和D-Images模块上的解纠缠表示学习效果.如可视化的数据所展示的，具有不同风格的描述文本非常明显且有序地在风格空间中分离开来.与此相反的是，事实空间上的风格信息则完全混合在一起且无法分辨.接着，通过对D-Captions和D-Images上的t-SNE可视化信息的对比，同样可以得到一个结论：D-Captions模块上对文本的风格解纠缠效果相对于D-Images对图像的风格解纠缠的效果更加明显清晰一些.尽管在表1中两个模块在DSA指标上的结果非常接近(0.93和0.94)，但是结合t-SNE可视化结果和表2和表3中具体生成效果，显示了D-Captions模块和D-Images模块都能够成功地对隐层空间进行解纠缠，然而文本中包含了相对于图像中更加显著的可分离的风格信息.

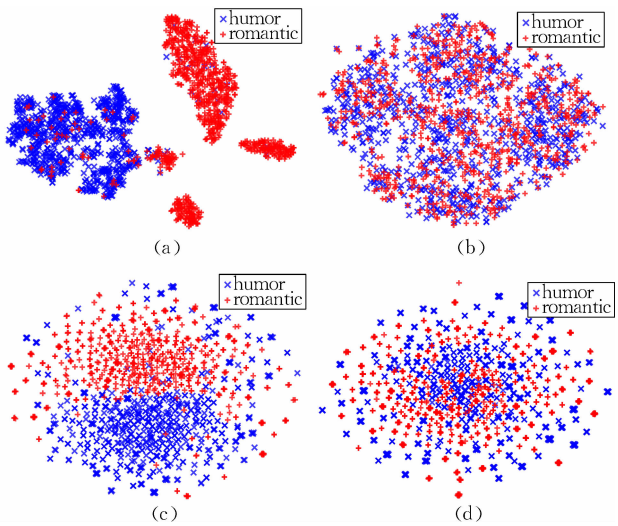


图3 t-SNE图((a)D-Captions模块解纠缠的风格空间、(b)D-Captions模块解纠缠的事实空间、(c)D-Images模块解纠缠的风格空间以及(d)D-Images模块解纠缠的事实空间)

表 2　在 SentiCap 数据集上的风格化图像描述生成结果(数据集上的最佳效果表示为粗体)

模型	SentiCap							
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN+LSTM	36.58	17.29	8.92	4.57	9.04	29.45	15.50	5.39
StyleNet	36.45	17.32	9.00	4.81	10.21	28.26	19.37	6.23
SF-LSTM	36.87	19.13	10.24	5.66	10.86	28.58	22.67	7.21
DSIC	**43.07**	**25.40**	**15.66**	**9.52**	**13.94**	**34.18**	**42.21**	**11.39**
D-Images	41.68	24.65	15.18	9.08	13.64	33.39	42.12	11.35
D-Captions	40.82	23.20	13.28	7.65	13.48	31.92	38.51	10.62
模型	积极风格							
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN+LSTM	36.38	16.93	8.43	4.50	8.85	29.98	14.68	5.25
StyleNet	37.48	18.57	9.37	5.11	10.50	29.11	19.12	6.35
SF-LSTM	37.44	19.45	10.67	5.72	11.40	28.42	23.37	6.66
DSIC	**45.31**	**27.04**	**16.31**	**9.70**	**14.87**	**36.20**	**47.19**	**11.87**
D-Images	44.32	26.28	15.99	9.75	14.66	34.97	45.19	11.65
D-Captions	41.07	23.15	12.67	7.04	13.55	31.89	39.11	10.31
模型	消极风格							
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN+LSTM	36.78	17.65	9.42	4.64	9.22	28.92	16.32	5.53
StyleNet	35.42	16.07	8.62	4.51	9.92	27.40	19.63	6.11
SF-LSTM	36.30	18.81	9.82	5.59	10.32	28.74	21.97	7.77
DSIC	**40.84**	**23.77**	**15.01**	**9.35**	**13.00**	**32.17**	**37.23**	**10.90**
D-Images	39.05	23.03	14.38	8.42	12.62	31.81	39.04	11.05
D-Captions	40.56	23.26	13.90	8.26	13.40	31.95	37.91	10.93

表 3　在 FlickrStyle10K 数据集上的风格化图像描述生成结果(数据集上的最佳效果表示为粗体)

模型	FlickrStyle10K							
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN+LSTM	24.64	11.75	6.47	3.77	8.97	22.78	26.38	9.94
StyleNet	23.55	11.49	6.29	3.63	9.01	22.28	25.92	10.11
SF-LSTM	22.83	11.01	5.94	3.36	8.72	21.96	26.29	11.00
DSIC	**28.79**	**16.03**	**9.17**	**5.34**	**11.38**	**27.51**	**35.09**	**13.54**
D-Images	25.29	12.48	6.87	4.02	9.91	23.46	31.06	11.71
D-Captions	26.09	13.03	7.21	4.24	10.21	23.79	33.48	12.44
模型	浪漫风格							
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN+LSTM	24.50	11.74	6.52	3.85	8.90	22.86	26.40	9.58
StyleNet	23.56	11.79	6.52	3.78	9.12	22.53	28.27	10.86
SF-LSTM	25.37	12.19	6.64	3.90	9.20	23.38	29.28	11.14
DSIC	**29.14**	**16.43**	**9.50**	**5.52**	**11.39**	**27.84**	**35.72**	**13.84**
D-Images	26.07	12.92	7.25	4.41	10.12	23.84	32.86	11.62
D-Captions	26.82	13.40	7.56	4.55	10.41	24.33	35.00	12.45
模型	幽默风格							
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN+LSTM	24.77	11.77	6.41	3.68	9.05	22.69	26.36	10.30
StyleNet	23.54	11.19	6.06	3.49	8.90	22.02	23.57	9.37
SF-LSTM	20.30	9.83	5.24	2.81	8.24	20.54	23.30	10.85
DSIC	**28.45**	**15.62**	**8.84**	**5.17**	**11.37**	**27.19**	**34.46**	**13.24**
D-Images	24.50	12.04	6.50	3.63	9.71	23.08	29.25	11.81
D-Captions	25.38	12.66	6.85	3.93	10.00	23.26	31.96	12.43

5.2　我们能够从解纠缠表示学习的性能中得到什么?

从上文的解纠缠表示学习的实验结果中,还可以发现一个有趣的现象.在以往的风格化图像描述生成任务中,通常认为语言风格是存在的且仅存在于文本中,这意味着应当仅从描述文本中对风格信息进行解纠缠表示学习.

然而,从 D-Images 的解纠缠表示学习的实验结果中观察发现,DSIC 同样可以成功地从图像的整个隐层空间中分离出语言风格相关的隐层信息.从这个现象以及 D-Images 和 D-Captions 两个模块的实验结果对比中可以得出一个结论:描述文本的语言风格不仅仅和文本,即人类的经验和想象有关,同样

和图像本身的视觉信息有关.但 D-Captions 的更佳性能也说明来自于人类语言经验占生成描述文本的风格信息相对于图像视觉信息占有更大的比重.

5.3 解纠缠表示学习是如何影响下游任务——风格化图像描述生成的?

基于解纠缠表示学习的风格化图像描述生成模型和当前最佳性能工作生成的描述文本的评测结果收集在表 2 和表 3 中.

关于生成的描述文本的流畅性和准确性,可以从表 2 和表 3 中看到,DSIC 模型在两个数据集的几乎所有的评价指标中都得到了最佳成绩.例如在完整的 SentiCap 数据集上,相比于该数据集上表现最好的基线方法 SF-LSTM,DSIC 模型的 BLEU-1 提高了 17%(从 36.87 到 43.07),而在 CIDEr 上达到了所有结果的最大的提升幅度 86%(从 22.67 到 42.21).在 FlickrStyle10K 数据集上,DSIC 模型的 BLEU-4 相比表现最好的基线方法 StyleNet 提高了 47%(从 3.63 到 5.34),ROUGE-L 提高了 23%(从 22.28 到 27.51).这充分表明了解纠缠表示学习对下游任务的表现带来了显著的提升.此外,本文展示了一些由 DSIC 模型生成的风格化图像描述的样例于图 4 中.这些实验结果充分展示了本文的解纠缠表示学习如何给风格化图像描述生成模型带来更好的生成质量.

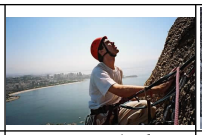
原始图片				
生成的幽默风格描述文本	a dog is jumping over a log in the water to catch a frisbee	a person is climbing a rock wall to find roses	a snowboarder jumps over a barrier in the air to catch a fish	two people riding on a rock face to reach outer space
生成的浪漫风格描述文本	a dog is jumping over a log attempting to win the game	a person is climbing a rock face to impresses his girlfriend	the skateboarder jumps over the waves to impress his lover	two people stand on a snowy mountain to conquer the high

图 4 DSIC 模型生成的风格化图像描述样例(每列均包含原始图像以及幽默和浪漫的风格化描述文本(或正面和负面,根据数据集而定),其中反映语言风格的词或短语带有颜色)

此外,本文还训练了一个 TextCNN 模型并将其用于预测生成的文本的风格倾向,实验结果收集于表 4 中.从这个结果中可以看到,DSIC 模型在两个数据集上分别取得了 96.13%和 100%的分类正确率,其性能显著优于其他对比模型.这个结果表明了解纠缠表示学习在控制解纠缠的风格因子方面的巨大优势.其中的一个主要原因是:相对于其他模型,本文的模型不仅仅抽取了风格信息用来控制文本的生成过程,同时还过滤了其他和风格不相关的事实信息.这样的操作能够让模型更加容易地生成风格化的图像描述并减少不相关信息的噪音干扰.与此同时,还观察到完整的 DSIC 模型相对于消融实验的模型(D-Images 和D-Captions)明显表现更好.例如,在 FlickrStyle10K 的 BLEU-4 指标上,D-Images 和 D-Captions 分别取得了 4.02 和 4.24 的分数,均优于 CNN+LSTM、StyleNet 和 SF-LSTM 这三种基线模型(最高 3.77),说明了 D-Captions 模块和 D-Images 模块均能够提升模型在任务上的表现.而完整的 DSIC 取得了所有模型中最高的得分(5.34),说明了 D-Images 和 D-Captions 模块能够相互补充,互相依赖.这进一步体现了无论从图像描述还是图像本身中学习到的解纠缠表示信息对生成模型的性能都有明显的帮助.综上所述,可以得出结论:DSIC 的解纠缠表示学习方法不仅能够更好地帮助人们理解风格化图像描述生成模型中的隐层空间的含义,还能够显著地提升风格化图像描述生成的性能.

表 4 由图像描述风格精度测量(ICSA)的生成效果(数据集上的最佳效果表示为粗体)

模型	FlickrStyle10K	SentiCap
CNN+LSTM	50.65	50.56
StyleNet	69.31	83.40
SF-LSTM	69.85	82.22
DSIC	**96.13**	**100.00**
D-Images	91.77	**100.00**
D-Captions	90.28	99.37

5.4 跨媒体信息的聚合方式是如何影响风格化图像描述生成性能的?

在跨媒体解纠缠信息的聚合过程中,为了研究不同聚合方式对最终效果的影响,本文设计 DSIC-concat 和 DSIC-attention 两种消融结构,并与 DSIC 模型在风格化图像描述生成任务中的表现进行了对比.实验结果被收集至表 5 中.从结果中可以发现,在两个数据集的所有评价指标中,DSIC 模型,即基于协议路由的胶囊网络聚合方式,都获得了最佳的结果.而另外两种聚合方式则相对低一些.

表 5　不同聚合方式在 FlickrStyle10K 和 SentiCap 数据集上的风格化图像描述生成结果(最佳效果表示为粗体)

模型	FlickrStyle10K							
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
DSIC	**28.79**	**16.03**	**9.17**	**5.34**	**11.38**	**27.51**	**35.09**	**13.54**
DSIC-concat	25.68	13.01	7.40	4.40	10.27	24.02	35.03	12.60
DSIC-attention	25.64	13.09	7.41	4.37	10.24	23.88	34.51	12.79
模型	SentiCap							
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
DSIC	**43.07**	**25.40**	**15.66**	**9.52**	**13.94**	**34.18**	**42.21**	**11.39**
DSIC-concat	41.55	23.87	14.41	8.43	13.40	33.30	40.53	10.99
DSIC-attention	41.42	23.95	14.26	8.05	13.09	32.93	37.09	10.18

例如在SentiCap数据集上，DSIC模型在BLEU-1和CIDEr上获得了43.07和42.21的表现，而DSIC-concat和DSIC-attention则为(41.55，41.42)和(40.53，37.09)的成绩.该现象说明了基于协议路由的胶囊网络能够更好地利用并整合图像和文本两种媒体的解纠缠信息.因为相比简单的线性变换而言，协议路由机制能够有效地聚集分布在各个媒体中的丰富信息，同时减少冗余的信息传递.此外，即使是DSIC-concat和DSIC-attention表现稍差，其最终的效果仍然显著好于其他基线模型.例如FlickrStyle10K数据集上METEOR和SPICE表现较好的StyleNet模型获得了9.01和10.11的成绩，如表2所示.相对而言，DSIC-concat和DSIC-attention在FlickrStyle10K上，METEOR成绩为10.27和10.24，SPICE成绩则为12.60和12.79，有着13.65%和24.63%的提升.这样的现象说明了无论哪种聚合方式，D-Captions和D-Images模块学习到的跨媒体的解纠缠表示都能够显著提升风格化图像描述生成的性能.

此外，本文还发现，DSIC-concat和DSIC-attention两种结构的性能相对来说非常接近，在不同的数据集的大部分评价指标上未有显著差距.例如在SentiCap上，利用attention聚合方式的模型BLEU-2获得了23.95的成绩，略好于DSIC-concat的23.87，而在FlickrStyle10K上，CIDEr的评价指标则是DSIC-concat的35.03略好于DSIC-attention的34.51.虽然在许多模型和任务中，注意力机制都有着相对于简单拼接更好的效果，然而在DSIC模型上，由于隐层空间类型较少(只有图像和文本两类)，而各个空间切片本身的信息却复杂(参数量较多)，因此注意力机制在这里并不能发挥相较于简单拼接更加明显的优势.而基于协议路由的胶囊网络则不然，其特殊的协议路由机制能够动态地更新每个媒体传递给跨媒体聚合表示的信息比例，从而最大程度捕捉分布在各个媒体的丰富信息，减少冗余信息传递.

综上所述可以得出结论，采用基于协议路由的胶囊网络聚合方式能够更好地聚合跨媒体的解纠缠信息，从而显著提升风格化图像描述生成的性能.

6　结　论

本文主要研究了风格化图像描述生成模型中的隐层空间上风格信息和事实信息的可解释性问题.本文首次将解纠缠表示学习技术引入到了风格化图像描述生成模型中来，并提出了一种先聚合后生成的方式充分利用同时来自图像和文本的解纠缠表示信息.在两个数据集上的多组实验结果表明，本文提出的模型性能显著超越了当前性能最佳的模型，并且证明了解纠缠表示学习技术在风格化图像描述生成任务中具有更好的可控性、可解释性和生成性能.此外，通过实验，本文通过模型学习到的解纠缠表示隐层空间还发现风格信息不仅仅存在于语言和人类的经验中，还存在于图像本身的视觉信息中.

参　考　文　献

[1] Miao Yi, Zhao Zeng-Shun, Yang Yu-Lu, et al. Survey of image captioning methods. Computer Science, 2020, 47(12): 149-160(in Chinese)
(苗益，赵增顺，杨雨露等. 图像描述技术综述. 计算机科学，2020，47(12)：149-160)

[2] Huang Yuan, Bai Cong, Li Hong-Kai, et al. Image captioning based on conditional generative adversarial nets. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(6): 911-918(in Chinese)
(黄远，白琮，李宏凯等. 基于条件生成对抗网络的图像描述生成方法. 计算机辅助设计与图形学学报，2020，32(6)：911-918)

[3] Zhao Jia-Qi, Wang Han-Zheng, Zhou Yong, et al. Remote sensing image description generation method based on attention and multi-scale feature enhancement. Computer Science, 2021, 48(1): 190-196(in Chinese)

(赵佳琦，王瀚正，周勇等.基于多尺度与注意力特征增强的遥感图像描述生成方法. 计算机科学，2021，48(1)：190-196)

[4] Wei Zhong-Yu, Fan Zhi-Hao, Wang Rui-Ze, et al. From vision to text: A brief survey for image captioning. Journal of Chinese Information Processing, 2020, 34(7): 19-29(in Chinese)
(魏忠钰，范智昊，王瑞泽等.从视觉到文本：图像描述生成的研究进展综述. 中文信息学报，2020，34(7)：19-29)

[5] Chen T, Zhang Z, You Q, et al. "Factual" or "Emotional": Stylized image captioning with adaptive learning and attention// Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 519-535

[6] Shuster K, Humeau S, Hu H, et al. Engaging image captioning via personality//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 12516-12526

[7] Locatello F, Bauer S, Lucic M, et al. Challenging common assumptions in the unsupervised learning of disentangled representations//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 4114-4124

[8] John V, Mou L, Bahuleyan H, et al. Disentangled representation learning for non-parallel text style transfer//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 424-434

[9] Higgins I, Amos D, Pfau D, et al. Towards a definition of disentangled representations. arXiv preprint arXiv: 1812.02230, 2018

[10] Kingma D P, Welling M. Auto-encoding variational Bayes// Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada, 2014: 1-14

[11] Li J, Yang B, Dou Z Y, et al. Information aggregation for multi-head attention with routing-by-agreement//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1(Long and Short Papers). Minneapolis, USA, 2019: 3566-3575

[12] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[13] Rumelhart D, Hinton G, Williams R. Learning internal representations by error propagation//Rumelhart D E, McClelland J L, eds. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations. Cambridge, USA: MIT Press, 1986: 318-362

[14] Hossain M Z, Sohel F, Shiratuddin M F, et al. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR), 2019, 51(6): 1-36

[15] Sheng S, Moens M F. Generating captions for images of ancient artworks//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France, 2019: 2478-2486

[16] Song Y, Chen S, Zhao Y, et al. Unpaired cross-lingual image caption generation with self-supervised rewards//Proceedings of the 27th ACM International Conference on Multimedia. New York, USA, 2019: 784-792

[17] Guo L, Liu J, Tang J, et al. Aligning linguistic words and visual semantic units for image captioning//Proceedings of the 27th ACM International Conference on Multimedia. New York, USA, 2019: 765-773

[18] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3156-3164

[19] Yao X, She D, Zhao S, et al. Attention-aware polarity sensitive embedding for affective image retrieval//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 1140-1150

[20] Yin G, Sheng L, Liu B, et al. Context and attribute grounded dense captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6234-6243

[21] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 4278-4284

[22] Xu K, Ba J L, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention//Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 2048-2057

[23] Yao T, Pan Y, Li Y, et al. Boosting image captioning with attributes//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 4894-4902

[24] You Q, Jin H, Wang Z, et al. Image captioning with semantic attention//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4651-4659

[25] Chen C K, Pan Z, Liu M Y, et al. Unsupervised stylish image description generation via domain layer norm//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 8151-8158

[26] Mathews A, Xie L, He X. SemStyle: Learning to generate stylised image captions using unaligned text//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, United States, 2018: 8591-8600

[27] Gan C, Gan Z, He X, et al. StyleNet: Generating attractive visual captions with styles//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3137-3146

[28] You Q, Jin H, Luo J. Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. arXiv preprint arXiv:1801.10121, 2018

[29] Zhao Zeng-Shun, Gao Han-Xu, Sun Qian, et al. Latest development of the theory framework, derivative model and application of generative adversarial nets. Journal of Chinese Computer Systems, 2018, 39(12): 2602-2606(in Chinese)
(赵增顺，高寒旭，孙骞等. 生成对抗网络理论框架、衍生模型与应用最新进展. 小型微型计算机系统，2018，39(12)：2602-2606)

[30] Nezami O M, Dras M, Wan S, et al. Towards generating stylized image captions via adversarial training//Proceedings of the 2019 Pacific Rim International Conference on Artificial Intelligence. Cuvu, Fiji, 2019: 270-284

[31] Locatello F, Tschannen M, Bauer S, et al. Disentangling factors of variation using few labels. arXiv preprint arXiv: 1905.01258, 2019

[32] Ye R, Shi W, Zhou H, et al. Variational template machine for data-to-text generation. arXiv preprint arXiv: 2002.01127, 2020

[33] Achille A, Eccles T, Matthey L, et al. Life-long disentangled representation learning with cross-domain latent homologies//Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018. Montreal, Canada, 2018: 9873-9883

[34] Li Y, Pan Q, Wang S, et al. Disentangled variational auto-encoder for semi-supervised learning. Information Science, 2019, 482: 73-85

[35] Litany O, Morcos A, Sridhar S, et al. Representation learning through latent canonicalizations. arXiv preprint arXiv: 2002.11829, 2020

[36] van Steenkiste S, Locatello F, Schmidhuber J, et al. Are disentangled representations helpful for abstract visual reasoning?//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019. Vancouver, Canada, 2019: 14222-14235

[37] Jain S, Banner E, van de Meent J W. Learning disentangled representations of texts with application to biomedical abstracts//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 4683-4693

[38] Kingma D P, Welling M. An introduction to variational autoencoders. Foundations and Trends in Machine Learning, 2019, 12: 307-392

[39] Mahajan S, Botschen T, Gurevych I, et al. Joint wasserstein autoencoders for aligning multimodal embeddings//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea, 2019: 4561-4570

[40] Mathews A P, Xie L, He X. SentiCap: Generating image descriptions with sentiments//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 3574-3580

[41] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013, 47: 853-899

[42] Guo L, Liu J, Yao P, et al. MSCap: Multi-style image captioning with unpaired stylized text//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4204-4213

[43] Zhao W, Wu X, Zhang X. MemCap: Memorizing style knowledge for image captioning//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 12984-12992

[44] van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(86): 2579-2605

[45] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318

[46] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Beijing, China, 2005: 65-72

[47] Vedantam R, Lawrence Zitnick C, Parikh D. CIDEr: Consensus-based image description evaluation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4566-4575

[48] Lin C Y. ROUGE: A package for automatic evaluation of summaries//Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. Spain, 2004: 74-81

[49] Anderson P, Fernando B, Johnson M, et al. SPICE: Semantic propositional image caption evaluation//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 382-398

[50] Kim Y. Convolutional neural networks for sentence classification//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1746-1751

附　录. 消融实验模型.

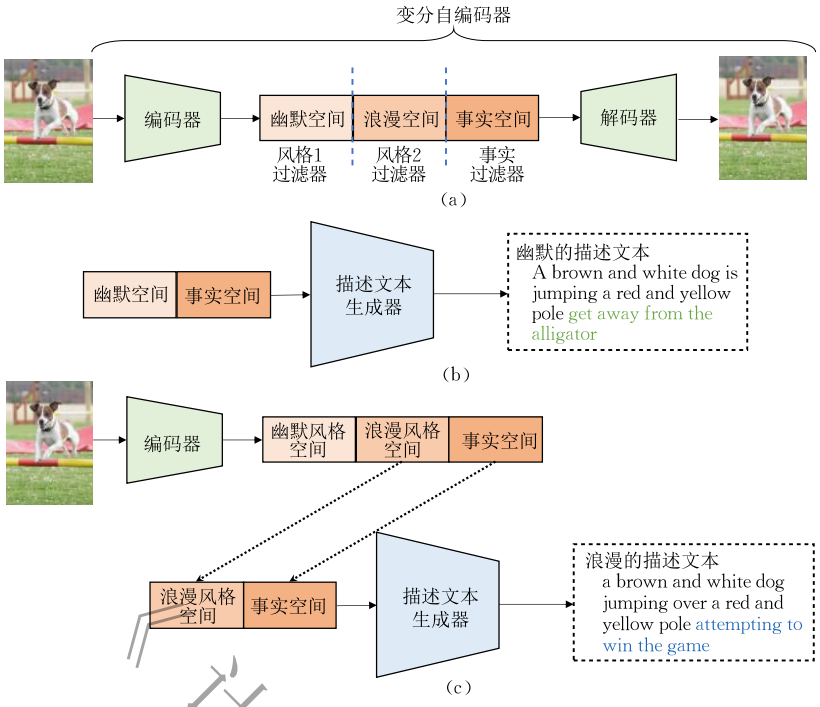
D-Images 模型

如附图 1 所示，在训练时，(1) 本文首先利用 $\mathcal{L}_{dis(s)}$，$\mathcal{L}_{dis(f)}$ 和 $\mathcal{L}_{ovr}$ 交替训练基于 VAE 结构的 D-Images 模型；(2) 对于训练集的每张图像 x^*，将其输入 D-Images 模型，并从中采样出风格表示 $\boldsymbol{z}_{s_1}^*$，$\boldsymbol{z}_{s_2}^*$ 和事实表示 $\boldsymbol{z}_f^*$. 本文分别将两种风格表示和事实表示相连接，输入到基于 LSTM 结构的描述生成器 $\mathcal{D}$ 中执行贪心解码，并采用交叉熵损失函数训练 $\mathcal{D}$；(3) 最后，本文对变分自编码器的编码端和描述文本生成器进行联合微调，即将图像输入编码端，然后在生成器中解码出相应的风格化描述.

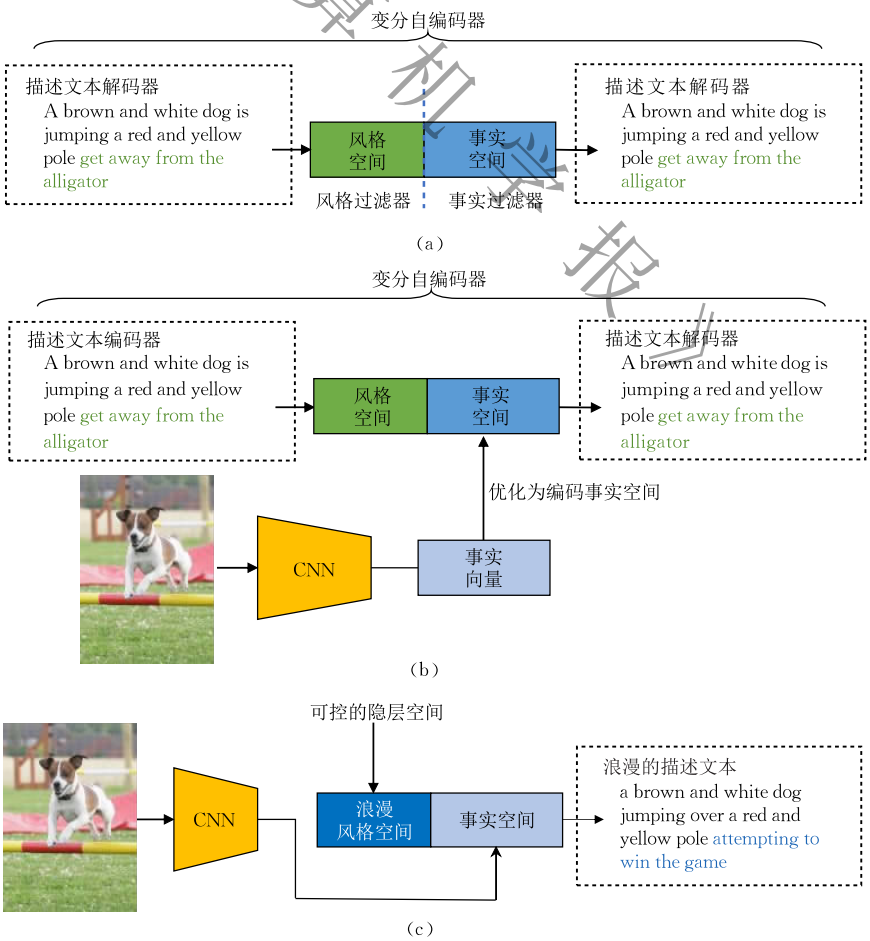
在测试时，给定任意一张图像，将其输入训练好的 D-Images 中，采样得到两种风格表示和一个事实表示. 随后，本文分别将两种风格表示和事实表示相连接，输入到训练好的描述生成器中执行大小为 5 的集束搜索，从而生成两种对应风格的图像描述.

D-Captions 模型

如附图 2 所示，在训练时，(1) 本文首先利用 $\mathcal{L}_{dis(s)}$，$\mathcal{L}_{dis(f)}$ 和 $\mathcal{L}_{ovr}$ 交替训练基于 VAE 结构的 D-Captions 模型. 而后，本文



附图 1 D-Images 模型((a)利用过滤器学习图像隐层空间的解纠缠表示;(b)使用对应风格的描述文本和风格及事实表示训练描述文本生成器;(c)对变分自编码器的编码器和描述文本生成器进行联合微调)



附图 2 D-Captions 模型((a)利用过滤器来学习描述文本隐层空间的解纠缠表示;(b)训练图像编码器编码事实空间;(c)对图像编码器和变分编码器中的解码器进行联合微调. 在图示例子中,模型利用从变分自编码器中得到的浪漫风格表示和从图像编码器中得到的事实表示来生成浪漫风格的描述文本)

在整个训练集上采样并取平均，获得通用的，固定的风格表示 z_s^*；(2) 对一张输入图像 x^*，本文采用之前提到的图像编码器 $\mathcal{E}$ 来获得视觉表示 v^*. 直觉上，本文将 v^* 视作图像 x^* 中的事实信息，并在图像编码器上应用 l_2 损失函数，使得视觉表示 v^* 能够尽可能接近 D-Captions 中得到的隐层事实表示 z_f^*；(3) 最后，本文对图像编码器和变分自编码器的解码端进行联合微调，即将图像输入图像编码器，然后与通用的风格表示 z_s^* 进行连接，输入到变分自编码器的解码端执行贪心解码，并采用交叉熵损失函数进行微调.

在测试时，给定任意一张图像，将其输入训练好的图像编码器，得到其事实表示，然后与通用的风格表示进行连接，输入到变分自编码器的解码端执行大小为 5 的集束搜索，从而生成对应风格的图像描述.

LIN Ze-Hao, Ph. D. candidate. His research interests include natural language processing, dialogue systems, machine learning and multimodal.

LI Guo-Dun, M. S. candidate. His research interests include natural language processing and machine learning.

ZENG Xiang-Ji, M. S. candidate. His research interests include natural language processing, counterfactual reasoning, and machine learning.

DENG Yue, Ph. D. candidate. His research interests include deep reinforcement learning and machine learning.

ZHANG Yin, Ph. D., associate professor. His research interests include artificial intelligence, intelligent question answering system, knowledge computing, and digital library.

ZHUANG Yue-Ting, Ph. D., professor. His research interests lie in artificial intelligence, cross-media computing, digital library.

Background

In this paper, we focus on the problem of understanding and controlling the latent representations of linguistic style and factual content in stylized image caption models by using the techniques of disentangled representation learning, which is a totally newly proposed task in cross-media domain.

Stylized image captioning is one of the cutting-edge topics in cross-media researches and applications. At the same time, how to appropriately interpret meaning behind parameters of neural networks is also a most critical problem in the machine learning community.

From perspective of representation learning, existing disentangling methods mainly focus on single modal data, such as images or texts. How to disentangle the latent space of cross-media model and apply it to downstream task still needs to be explored. From perspective of image captioning, stylized image captioning takes a further step, aims at generating captions with a target style. Existing studies mainly concentrate on how to adapt neural architecture to better incorporate linguistic styles into natural language generation, without consideration of improving the interpretability, generalization to unseen scenarios and faster learning on downstream tasks.

In this paper, the authors propose a novel approach, Disentangled Stylized Image Caption (DSIC), to learn the disentangled representations on unparallel cross-media data. Experimental results for disentanglement performance show that our models can successfully disentangle the stylistic and factual information and reveal that stylistic information exists in both human beings' experience and images themselves. Experimental results for stylized image caption show that our disentangled representation learning approach benefits the interpretability, controllability, and boosts the downstream stylized image captioning performance.

The authors of the paper have conducted research on image captioning, natural language generation and disentangled representation learning in recent years. Especially, they proposed several neural network models for multimodal representation learning to boost the performance on downstream tasks, which have been published in EMNLP-2019/2020, AAAI-2020/2021, ACM MM-2021 conferences.

This work was supported by the National Natural Science Foundation of China (No. 62072399, No. 61402403, No. U19B2042), the Chinese Knowledge Center for Engineering Sciences and Technology, MoE Engineering Research Center of Digital Library, the Fundamental Research Funds for the Central Universities and the Artificial Intelligence Research Foundation of Baidu Inc.