硕士学位论文

基于隐变量生成模型的多样化视频描述

TOWARDS HUMAN-LIKE DIVERSE VIDEO CAPTIONING VIA A LATENT GENERATIVE MODEL

研　　究　　生：刘柱
指　导　教　师：郑锋副教授

南方科技大学

二〇二二年六月

国内图书分类号：XXxxx.x　　学校代码：14325
国际图书分类号：xx-x　　密级：公开

工学硕士学位论文

基于隐变量生成模型的多样化视频描述

学 位 申 请 人：刘柱
指 导 教 师：郑锋副教授
学 科 名 称：电子科学与技术
答 辩 日 期：2022 年 5 月
培 养 单 位：计算机科学与工程系
学位授予单位：南方科技大学

Classified Index: XXxxx.x
U.D.C: xx-x

Thesis for the degree of Master of Engineering

TOWARDS HUMAN-LIKE DIVERSE VIDEO CAPTIONING VIA A LATENT GENERATIVE MODEL

Candidate:	Liu Zhu
Supervisor:	Associate Prof. Zheng Feng
Speciality:	Electronic Science and Technology
Date of Defence:	May, 2022
Affiliation:	Department of Computer Science and Engineering
Degree-Conferring-Institution:	Southern University of Science and Technology

学位论文公开评阅人和答辩委员会名单

公开评阅人名单

刘 XX	教授	南方科技大学
陈 XX	副教授	XXXX 大学
杨 XX	研究员	中国 XXXX 科学院 XXXXXXX 研究所

答辩委员会名单

主席	赵 XX	教授	南方科技大学
委员	刘 XX	教授	南方科技大学
	杨 XX	研究员	中国 XXXX 科学院 XXXXXXX 研究所
	黄 XX	教授	XXXX 大学
	周 XX	副教授	XXXX 大学
秘书	吴 XX	助理研究员	南方科技大学

南方科技大学学位论文原创性声明和使用授权说明

南方科技大学学位论文原创性声明

本人郑重声明：所提交的学位论文是本人在导师指导下独立进行研究工作所取得的成果。除了特别加以标注和致谢的内容外，论文中不包含他人已发表或撰写过的研究成果。对本人的研究做出重要贡献的个人和集体，均已在文中作了明确的说明。本声明的法律结果由本人承担。

作者签名：　　日期：

南方科技大学学位论文使用授权书

本人完全了解南方科技大学有关收集、保留、使用学位论文的规定，即：

1. 按学校规定提交学位论文的电子版本。

2. 学校有权保留并向国家有关部门或机构送交学位论文的电子版，允许论文被查阅。

3. 在以教学与科研服务为目的前提下，学校可以将学位论文的全部或部分内容存储在有关数据库提供检索，并可采用数字化、云存储或其他存储手段保存本学位论文。

（1）在本论文提交当年，同意在校园网内提供查询及前十六页浏览服务。

（2）在本论文提交 □ 当年/ □ 一年以后，同意向全社会公开论文全文的在线浏览和下载。

4. 保密的学位论文在解密后适用本授权书。

作者签名：　　日期：

指导教师签名：　　日期：

摘 要

本文主要研究视频描述任务，旨在为视频生成多样化的描述内容。该任务涉及计算机视觉和自然语言处理，在短视频描述、新闻摘要、人机辅助、智能助理等重要领域都有着广泛的前景和应用。然而，由于视频场景中复杂的交互和丰富多样的语言表达等特点，单句描述往往无法像人类一样全面刻画视频内容。而人类倾向于提供一个具有多样性的描述集合，用于反映视频中不同层级的细节。因此视频描述本质上是一个“一对多”的映射任务。

当前视频描述模型在现有的评价指标上达到了比人类更高的水平，但是这些模型过度追求准确性，难以满足描述的多样化需求问题。除此之外，现有的指标也无法完全刻画这种多样性。这些问题使得现有方法难以应用于不确定性较大的现实场景并提供类似于人类标注的描述，如自动驾驶、语言教育等领域。

因此，本文开展了面向多样化视频描述生成任务的研究，并基于 VAE 生成模型，提出一系列训练方式和模型架构以增强描述的多样性。本方法首先设计了一个动作和上下文分离的结构化隐空间（ATVAE），用于捕捉视频场景下动作以及语言表达的多样性。同时利用对比学习进一步提高句子间的差异性，能够有效缓解 VAE 框架常见的后验坍塌问题。在 ATVAE 的基础上，本方法进一步设计并实现了双阶段渐进训练方式（STR），具体的训练过程包括：（1）第一阶段，模型在区分度较大的话题句上进行训练。（2）第二阶段立足于前一阶段，将模型用于整个数据集上进行训练，从而得到更多样化的表达。本文通过大量实验从定性和定量两个方面论证了两类方法的有效性：它们可以在不损害准确性的前提下，有效提高生成的描述的多样性。为了能够衡量生成字幕集合的整体表现性能，本方法提出两个新的指标，命名为 hau 和 o2o。与现有的其他指标相比，本文提出的新指标能够从多维角度同时考虑描述的准确性和多样性。实验结果表明，两个指标都具有与人类评估更高的相关性，这对于模型评估和模型选择都有重要意义。

关键词：多样化视频描述；变分自编码器；评价指标；结构化隐空间

Abstract

Video captioning aims to generate a sentence to describe a short video clip. As a typical task under artificial intelligence, a suitable description is expected to be fluent and coherent in language and accurate and relevant to the salient objective, actions, and background, which is therefore challenging. The task belongs to the overlapped area between computer vision and natural language processing. It embraces various applications, including short-video captioning, news summarization, human-aided systems, and intelligent agents. Due to complicated interactions in a video and the ambiguousness of natural language, a single sentence cannot cover the video thoroughly, while human annotators tend to provide multiple proper candidate sentences (i.e., one-to-many mapping). However, most current models are accuracy-based and generate one single sentence, ignoring such multimodal distribution. Thus it fails to solve the areas which demand high security or encourage diverse descriptions, such as auto-driving and education.

The thesis deals with the task of human-like diversity video captioning and proposes a series of training mechanisms and models based on VAE, a type of latent generative model. In the first part, we design a structured latent space, dubbed ATVAE, splitting action and template variables and attempt to capture the diversity hidden in the interactions. Meanwhile, a contrastive learning term is added to the objective function to further improve diversity by alleviating the posterior collapse, which typically appears in the framework of VAE. In the second part, we employ a two-stage learning strategy (STR) with each stage based on the model proposed in the first part. The design utilizes the relationship among different captions predicted, i.e., topic and expression. Specifically, first, we train our model in a subset consisting of several topic sentences, then we rephrase the captions based on the second stage. Experiments in both parts demonstrate the effectiveness and efficiency of our methods quantitatively and qualitatively. The last part introduces two new metrics dubbed hau and o2o that combine accuracy and diversity. We conduct a well-designed human evaluation on results from different models. Finally, our proposed metrics prove a stronger correlation to human measurement, which has a significant role in model evaluation and selection.

Keywords: diversity video captioning; variational autoencoders; structured latent space

目 录

摘 要.....I
Abstract.....II
符号和缩略语说明.....V
第 1 章 绪论.....1
1.1 研究工作的背景与意义.....1
1.2 国内外研究现状与发展趋势.....4
1.2.1 单句视频描述.....4
1.2.2 多句视频描述.....7
1.3 本文的主要工作.....10
1.4 本文的结构安排.....11
第 2 章 背景知识.....12
2.1 生成模型.....12
2.1.1 自回归模型.....12
2.1.2 隐变量生成模型.....13
2.2 变分自编码器.....14
2.3 条件变分自编码器.....16
2.3.1 后验坍塌.....16
2.4 不确定性.....17
2.5 本章小结.....18
第 3 章 基于动作-模板分离的隐变量模型建模.....19
3.1 引言.....19
3.2 方法.....21
3.2.1 问题正式化描述.....21
3.2.2 分离空间的条件变分自编码器.....21
3.2.3 带有对比学习正则化的解码器.....22
3.2.4 模型训练和推断.....23
3.3 实验验证与分析.....23
3.3.1 数据集.....23
3.3.2 执行细节.....25

3.3.3 性能比较.....26
3.4 消融实验.....27
3.4.1 分离隐空间.....27
3.4.2 对比学习.....28
3.4.3 正则化系数.....28
3.5 本章小节.....30
第 4 章 基于渐进式训练的多样视频描述.....31
4.1 引言.....31
4.2 问题正式化描述.....33
4.3 方法.....34
4.3.1 话题聚类.....34
4.3.2 渐进式训练.....35
4.3.3 模型训练.....36
4.4 实验.....37
4.4.1 实验细节.....37
4.4.2 定量比较.....39
4.4.3 定性分析.....43
4.5 讨论：与单阶段模型的联系和区别.....44
4.6 本章小结.....45
第 5 章 集合水平的评估指标设计.....46
5.1 已有指标.....46
5.1.1 语言任务相关.....46
5.1.2 视觉任务相关.....48
5.2 指标设计.....49
5.2.1 融合召回率的集合距离.....49
5.2.2 一对一匹配的集合距离.....50
5.3 评测相关性.....51
5.3.1 问卷评测设计.....51
5.3.2 问卷评测结果分析.....53
结 论.....56
参考文献.....58
致 谢.....67
个人简历、在学期间完成的相关学术成果.....68

符号和缩略语说明

$\mathbf{A}$	矩阵
$\mathbf{a}$	向量
a	标量
W	神经网络参数矩阵
$\mathcal{D}$	数据集
$\mathcal{D}'$	数据集的子集
$\mathcal{R}$	参考集合
$\mathcal{H}$	预测集合或者假设集合
$\mathcal{X}$	整个输入空间（句子）
x	输入或者重构的句子
z	隐变量
$\mathbb{E}$	词嵌入映射矩阵
c	视频特征（条件输入）
$\mathcal{L}$	损失函数
p	真实分布
q	近似分布
$\mathcal{N}$	高斯分布
f	映射函数
ϕ	变分参数
$\mathbb{R}$	实数集
VAE	变分自编码器
ELBO	证据下界
NN	神经网络
CNN	卷积神经网络
RNN	循环神经网络
MCMC	马尔可夫蒙特卡洛采样
VI	变分推断
i.i.d.	独立同分布

第 1 章　绪论

1.1　研究工作的背景与意义

视频描述是指给定一段视频，计算机自动得出可以恰当描述视频内容的语句。该任务主要考虑短视频场景下的单句描述生成。视频描述同时涉及计算机视觉领域和自然语言领域：在视觉方面，计算机需要正确识别、检测视频中出现的主要物体以及它们之间的互动，正确理解视频中正常发生的事件；在自然语言方面，该任务要求计算机生成像人类描述一样地，语法正确、通顺流畅、信息丰富的语句。因此，视频描述一直是人工智能（Artificial Intelligence，简称 AI）领域中一项富有挑战性的任务。

视频描述在日常生活中有很多应用。新闻、电影、短视频等视频网站都需要简短的摘要，以便更快地引起观众的注意；以抖音为代表的海量短视频创作平台需要大量的语言标签提示，而自动化的标注可以节约大量的人力物力；盲人等视觉有障碍的群体，可以通过智能助手自动获得眼前事件的描述，从而给生活带来许多便利；另一方面，视频描述还可以促进手语视频的翻译，让更多人也可以通过计算机读懂手语。同时，手语作为一门“视觉语言”，其场景中的视觉信号更加丰富多样，可解码出的语义信息也更加丰富；未来机器人、终端助理等智能体的出现则非常依赖理解并描述眼前场景的视频描述技术，尤其涉及到人机对话、人机交互等场景。图 1-1总结了这些应用场景。

除了应该保证生成的句子尽可能准确地描绘视频内容外，视频描述还具有“多样性”的要求：即生成相互独立但彼此间有差异的多条描述。这个过程应与人类标注的情况一致（例如训练数据往往是一个视频对应多个人类标注者）。这些描述除了自身足够准确外，彼此之间在描述内容、表达用语、语言结构等等各方面都会存在差异。多方面的因素导致了这种多样化的结果。

从视觉方面来看，视频场景本身就存在比较复杂和不确定性的因素，而人类心理的注意力机制使得不同的标注者会根据自己的偏好寻找不同的侧重点。根据神经心理学的研究[1]，人类的注意力可以分为两个不同的类型：（1）自底向上（bottom-up）的注意力，指的是纯粹由外部突出因素对主体产生显著的刺激而导致。例如平静湖面中突然飞来的白鹤闯入了观者的视野，使他意外地注意到这一现象。（2）自上而下（top-down）的注意力，指的是出于先验知识、有意计划并且往往带有特定目标的引导。比如为了生物研究而刻意去观察某棵树木等。同时，这

(a) 手语视频需要翻译　　(b) 未来智能助理

(c) 新闻视频摘要　　(d) 盲人导航

图 1-1　视频描述在不同场景下的应用

与心智哲学（philosophy of mind）中的强调主体体验过程的现象意识（phenomenal consciousness）和强调主体主观能动性的自觉意识（access consciousness）相互对应。可以想象，面对一段连续的视觉场景，标注者心理可能会同时存在这两种注意力，从而自觉（自上而下）或不自觉（自底向上）地偏重视频中的某些元素，最终产生丰富多变的描述。

从语言的角度观察，自然语言具有歧义性、模糊性，对于同一个句子内容本身就可以使用多种方式去表述。语言学的观点认为，自然语言的歧义性包括（1）词汇（lexical）。同一个单词有不同的意思，例如词典中某个单词对应的不同义项。（2）句法（syntax）。同一句话根据不同的句法成分分析可能导致不同的意思。（3）语义（semantic）。同一句话有不同的理解方式，因而导致语义差异。（4）篇章（discourse）。这往往由于不同词语的指代模糊导致的。（5）语用（pragmatic）往往指说话者的背景不同可能导致不同的理解，这往往和意图、情感、信念等相关。而不同的视频标注者之间由于个人的教育经历、说话方式、价值选择等差异，“解码”出内心意图的自然语言因此受到上述各个方面的影响而趋异，这也是导致多样化描述的另一个重要方面。

同时不能忽视的是，视频描述涉及到人工智能的认知、理解甚至创造等方面。如果可以将它看作是一个试图模仿人类行为的实体（entity），它应该可以说出像人一样、甚至可以混淆人类的描述。而一个事实是，对于人类而言，这本就是一个富

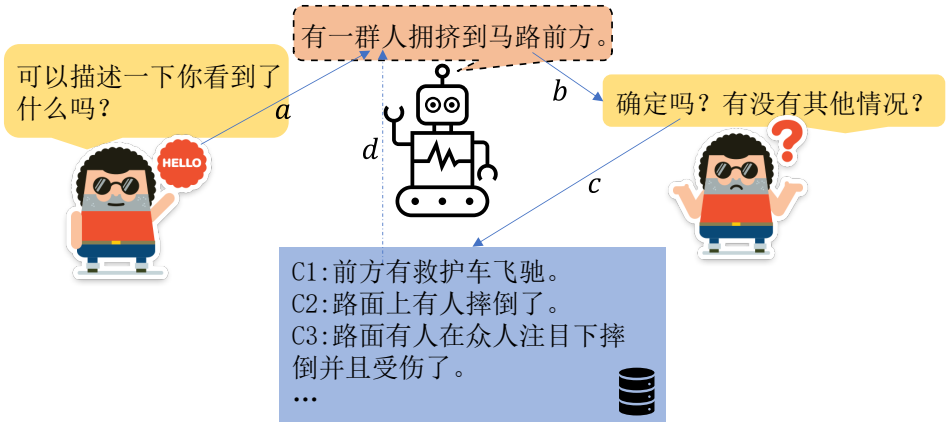


图 1-2　多样化描述在自动驾驶领域可能出现的场景示意图

有表达性（expressiveness）的认知过程：不同教育、经历、理解与表达方式的标注者都会有自己倾向的意图（intention），从而表现为形态各异的描述。以神经网络为基础的实体仍期望重建这种表达性，达到与人类标注可比的效果。

多样化描述仍富有应用场景，尤其是面对一些不确定性较大、鲁棒性和多样性要求较高的场景。（1）教育应用需要指导学生“看图说话”，而多样的描述可以帮助他们从不同的角度描述图像或者视频中的内容或者使用不同的表达方式去描述，从而让他们更加熟练应用语言能力，锻炼语言智力的发展；（2）而当智能助理对当前场景提示的描述不符合常识或者不够全面时，用户可能会进一步询问“还有其它可能吗？”。机器需要储备多条候选的描述，从而帮助用户更好地做出决策等等，例如图 1-2中展示了自动驾驶领域可能出现的一个需要多样化描述的场景：面对智能系统给出的不完整的描述，用户可能由于好奇可能要求给出更多同时发生的表达，以寻求对于当前场景更加全面的了解；（3）另一方面，生成的多样描述也可以作为其它跨模态任务的数据集的补充，例如视觉问答、视觉语言预训练等。由于可以生成多条描述，因而可以快速扩增训练数据的容量。它也可以作为一种数据增强的方式，帮助数据更好的训练。这样做可以减轻人工数据标注的困难、减少标注成本，也可以促进半监督、无监督训练方法的发展。总的来说，多样化描述生成是为了应对更加模糊、不确定的一些场景，而这与日渐受到关注的 AI 安全、AI 鲁棒性等都有密切联系。

目前深度学习在（基于和参考描述库匹配的）准确性和句法正确性上都可以起到不错的效果，然而在生成描述的多样性上却很薄弱。本文试图解决这一问题，旨在生成像人类一样的多样化但却不失准确性的描述句子集合。同时针对评测指标无法统一准确性和多样性两方面的缺陷，本文也提出了两个可以刻画描述集合整体性能的统一指标。

1.2　国内外研究现状与发展趋势

本节介绍与本文提出的方法相关的研究工作。分别从单句视频描述、多句视频描述、基于变分自编码器（Variational Auto-encoder, 简称 VAE）的图像描述三个方面进行说明。

1.2.1　单句视频描述

视频描述文献可以分为三个主要阶段：1）传统方法阶段：它们使用经典的计算机视觉和自然语言处理的模型，首先检测到视频中的实体（对象、动作、场景），然后将它们拟合到提前预处理好的句子模板中。2）统计方法阶段：该阶段采用统计方法来处理相对较大的数据集。这一阶段持续了相对较短的时间。3）深度学习阶段，这是当前最先进的阶段，被认为具有解决开放域自动视频描述问题的潜力。

1.2.1.1　传统方法

基于 SVO（主语、宾语、动词）三元组的方法是专门应用于视频描述的首批成功方法之一。然而，人工智能相关的研究工作早在很久以前就已尝试将视觉内容描述为自然语言，尽管这时的自然语言并未明确用于描述。第一次尝试可以追溯到 Koller 等人[2]，他们在 1991 年开发了一个能够使用动词来表示车辆在真实交通场景中的运动。1997 年，布兰德等人[3]将这一问题称为“逆好莱坞问题”（因为电影需要将好莱坞剧本（描述）中转换为视频；现在刚好需要反过来，即视频转化为文字），并将一系列动作描述为语义标签摘要，对一些教学视频进行语言描述，形成文字上的“故事线”。他们还开发了一个名为“video gister”的系统，该系统能够启发式地将视频解析为一系列关键动作，并生成一个描述视频中检测到的动作的脚本。他们还生成了描述检测到的因果事件的关键帧，并将一系列事件定义为语义表示，例如，使用“进入”、“运动”、“分离”表示添加和使用“移动”、“离开”等表示移除。但 gister 仅限于一只人类手臂与刚体物体交互，并且只能理解五种动作（触摸、放置、获取、添加、删除）。因而它们对于视频内容的理解非常局限。

基于 SVO 三元组的方法将视频描述生成任务分解为两个阶段。第一阶段称为内容识别，它侧重于视频片段中主要对象的视觉识别和分类。这些通常包括动作的发出者（actor）、动作以及该动作的承受者。第二阶段则涉及句子生成，它将第一阶段识别的对象映射到主语（S）、动词（V）和宾语（O）（因此得名 SVO），并添入到一个语法正确、提前人工设计好的一个语言模板中。这些模板是使用语法或基于规则的系统创建的，因而这些系统仅在非常受限的环境中有效果，即对象和动作数量有限的视频。

第一阶段的主要对象识别主要应用当时计算机视觉中的检测算法。1）目标识别包括通过边缘检测或颜色匹配进行的形状匹配、HAAR 特征匹配[4]、基于上下文的目标识别[5]、尺度不变特征变换[6]、基于 part 的判别式模型[7]和可变形零件模型[8-9]。2）施动者（往往是人类）和运动检测：人类检测方法采用了直方图定向梯度[10]等特征，运动检测则涉及到时空兴趣点等特征，例如定向光流直方图[11]、贝叶斯网络[12]、动态贝叶斯网络[13]、隐马尔可夫模型[14]、有限状态机[2]等。3）有些则采用融合的方式，例如随机属性图像语法[15] 和随机上下文无关语法[16]。它们利用物体时空之间的关系，将图像内容解析为不同权重的实体。

第二阶段主要关注自然语言的部分，即如何生成一个完整、符合语法规则、通顺正确的句子。已经有多种方法用于解决这一问题：包括 HALogen 表示①，头驱动的短语结构化语法（Head-driven Phrase Structure Grammar，简称 HPSG）[17]等。这些方法的主要常见任务是定义一个语言模板。模板是包含占位符的由用户定义的语言结构，它通常由词汇、语法和模板规则三部分组成。在基于模板的方法中，通过将最重要的实体拟合到模板所需的每个类别（例如，主语、动词、宾语和地点）来生成句子。这一类方法又可以根据分为侧重主体者（主语）的[18-21]、侧重动作和物体的[17,22-23]以及适用于开放域视频的 SVO 方法[24-25]。值得注意的是，本文之后采取的基于模板和动作分离的空间设计，仍与此处的先检测再填充模板有相似之处，只是本文的分离（在测试阶段）是一种隐式的、是通过神经网络自动感知填充动词的位置以及内容。

1.2.1.2　统计方法

仅仅基于 SVO 元组规则的工程方法不足以描述开放域视频和大型数据集，例如 YouTubeClips[26]、TACoS-MultiLevel[27]、MPII-MD[28]、M-VAD[29]、MSRVTT[30]、VATEX[31]等。这些数据集包含非常大的词典以及数十小时的视频。这些开放域数据集与以前的数据集之间存在三个重要区别。首先，开放域视频包含不可预见的多样化主题、对象、活动和地点。其次，由于人类语言的复杂性，此类数据集通常带有多种可行的有意义的描述。第三，要描述的视频通常很长，可能会长达数小时。多句甚至多段视频的描述变得更可取。

为了避免基于规则的工程方法所需的繁琐工作，Rohrbach 等人[32]提出了一种将视觉内容转换为自然语言的机器学习方法。他们使用平行的视频语料库和相关的描述。他们的方法遵循两步：首先，它学习使用最大后验估计（Maximum a Posteriori，简称 MAP）将视频表示为中间语义标签。然后使用统计机器翻译（Statistical Machine Translation，简称 SMT）[33]技术将语义标签翻译成自然语言句子。在这种

① http://www.isi.edu/publications/licensed-sw/halogen/interlingua.html

机器翻译方法中，中间语义标签表示是源语言，而预期的描述被视为目标语言。

对于目标和动作识别阶段，早期研究从基于阈值的检测[34] 转移到手动特征工程和传统分类器[24,35-36]。对于句子生成阶段，近年来可以观察到机器学习方法的采用，以解决词汇量大的问题。最近的方法趋势也证明了这一点，这些方法使用以弱监督[27,32,37-38] 或全监督[24,36,39-40] 方式学习的模型。然而，这两个阶段的分离使得这类方法无法捕捉到视觉特征和语言模式的相互作用，更不用说学习视觉表示和语言表示之间的公共的表示空间。但这类方法可以视作是深度学习方法的一个过渡期。

1.2.1.3　深度学习

深度学习在计算机视觉和自然语言处理的几乎所有子领域的颠覆式的成功也彻底改变了视频描述的方法。其中，卷积神经网络（Convolutional Neural Network，简称 CNN）[41] 是用于对视觉数据进行建模的常用方法，并且擅长目标识别等任务[41-43]。同时，长短期记忆（Long Short-Term Memory，简称 LSTM）[44] 和更通用的深度循环神经网络（Recurrent Neural Networks, 简称 RNN）是处理序列数据常用到的方法，如机器翻译[45-46]、语音识别[47] 以及与图像字幕密切相关的任务[48-49]。另外，随着基于 transformer 架构模型的预训练-微调范式的兴起[50] 和非常惊人的表现，很多序列生成模型会优先考虑 transformer 模型，尤其在数据量较大的情形下。

视频描述主要采用编码-解码（encoder-decoder）的方法：编码器针对时序视频，主要使用 CNNs 提取包含动作、物体等的视觉信息，解码器用于语言生成，其常常采用 RNN 或者 transformer 进行时序建模。注意到，传统的 SVO 三元组的方式将第一阶段视频内容编码阶段的结果用于直接生成**离散**的单词，再去“硬”匹配一个同样是单词构成的句子，而深度学习的编码策略是将视频特征表示为一个固定的或者动态的**连续**向量。从离散到连续，尽管牺牲了部分可解释性，但可表示的语义空间得到了极大的扩展、匹配更加灵活抽象。

RMN[51] 区分了句子中不同类别的词汇生成，如名词、动词、功能词，并使用注意力机制关注到不同方面的视觉性息。SAAT[52] 发掘动作对于视频描述生成的重要性，在生成整个句子之前先显式地推断出动词。STG-KG[53] 则进一步发掘了物体间地时空关系，并且建模为了图卷积网络。由于视频场景比起图像来说更加复杂，其中物体与物体之间、物体与环境之间的互动关系更加多样，因而这些方法都显式地建模这种复杂的互动关系。SemSynAN[54] 将视觉特征和句法特征[55] 映射到一个共有空间，学习到它们的对齐特征，同时应用一些视觉概念检测器，检测到视频中出现的视觉词汇，取得了较优的结果。

1.2.2　多句视频描述

由于描述的复杂和不确定，对于单个的视频片段，可以对应多句话的输出。本节概括了在视频描述领域中常见的生成多句话的任务，它们的任务特点可以参考表 1-1。

任务名称	短视频	句子间独立	额外辅助信号	标注为一对多
密集视频描述	否	否	否	否
视频段落描述	否	否	否	否
可控视频描述	是	是	是	是
多言视频描述	是	是	否	是

表 1-1　各类多句视频描述任务对比

1.2.2.1　密集视频描述

密集视频描述（dense video captioning）[56]针对视频中出现的不同事件产生一个描述集合，并预测它们的时间位置定位。该任务脱胎于密集图像描述[57]，但基于如下两个观察作者认为仍有单独研究该任务的必要性：第一个观察是视频中的不同事件会在不同时间尺度上有所不同，甚至会有所交叠。例如，演唱会（长期事件）中途伴随着的掌声（短期事件），而过去单句视频描述的场景将一个视频当作整体对待（例如对于整个视频的不同采样帧特征进行平均池化），因而无法应对这类场景；第二个观察是视频中出现的事件往往彼此关联，它们往往具有部分的因果关系（观众鼓掌是由于演唱会事件的发生）。如何发掘事件彼此的关联也需要额外关注。其中，前一种观察主要针对定位模块（localization module），即定位出突出语义的视频边界；后一种观察则针对标注模块（captioning module），主要生成一致的、准确描述的、流畅的语言描述。此外，作者又开放了一个名为 ActivityNet 的密集视频描述的数据集①。

多数方法重点针对标注模块，即如何利用不同事件标注之间的上下文关系。例如工作[58]采取了利用注意力融合机制将过去和现在的上下文融合在一起。同时也有文章[59]利用单样本检测（Single Shot Detector，简称 SSD）的方式去融合事件的时间间隔。同时还有工作通过上下文建模[60-61]，事件水平的关系[62]以及多模态特征融合[63-64]等方式产生更加准确的标注。还有工作探究定位模块如何对标注模块进行辅助，即探究两个模块之间的融合。文章[65] 引入了一个代理任务，即预测生

① http://activity-net.org/index.html

成句子的语言奖励，作为本地化模块的附加优化目标。另外一篇文章[66]提出了一种差分掩蔽机制，将描述损失和时间间隔一起算作损失的一部分，从而实现两个任务的联合优化。另一个模型 PDVC[67]则将这一任务定义为一个并行的集合预测的问题，并将定位模块和标注模块放在一起考虑，从而二者可以有效互补。

1.2.2.2　视频段落描述

与上述的密集视频描述不同的是视频段落描述任务，它是给定一个较长的视频，输出多句话组成的段落。该任务可以给出或者不给出显式的时间间隔，然后得出每个间隔的单句描述，因而最终可以拼接成一个段落。

克劳斯等人[68]提出了带有分层 RNN 的图像段落描述，它首先生成主题向量，然后将主题转换为句子以形成段落。然而，最近的工作[69-70]表明，当通过多样性驱动的训练和推理方法进行增强时，直接将段落生成为长句优于分层方式。不同于类似密集视频描述方法的两阶段的方法[71]，方法[72]也探讨直接从单阶段输出一个段落的方法：文章提出了一种关键帧感知视频编码器来提高编码效率，并提出了一种具有动态视频记忆的注意力机制来学习更多样化和连贯的视觉注意力。此外，文章提出了一种具有高频标记和短语惩罚的多样性驱动训练目标，从而提高语言多样性。

1.2.2.3　可控视频描述

可控视频描述（controllable video captioning）则是给定一个待描述视频和一个控制信号（controllable signal），模型需要得到符合该控制信号的准确描述。与之任务类似研究自然语言任务上的可控语言生成的一篇文章[73]将该控制信号概括为控制类型（aspect），包含：话题、情感等。而每个类型下面又分出具体的属性（attribute），例如情感中的积极、消极。可控视频描述任务可以视作是有条件的单句视频描述任务。由于可控视频描述任务与可控图像描述任务本质类似，这里将二者都列入其中。

根据文章[74]，一个理想的可控信号应该包含两个方面的特征：1）事件兼容性（event-compatible），即可控信号应该是描述的完整事件的一部分；2）样本兼容性（sample-compatible），即可控信号应该可以用来描述相应的样本。同时，文章将可控信号分为基于（视觉）内容的和基于（语言）结构的。1）基于内容的信号：例如视觉关系[75]、对象区域[76-77]、场景图[78-79]和鼠标轨迹[80]。2）基于结构的控制信号：控制信号是关于句子的语义结构。例如，句子的长度级别[81]、词性标签[55]、语义角色[74]、情感[73]或属性[74]。

值得注意的是，控制信号与本文重点使用的隐变量生成模型中的隐变量有相

似之处：控制信号可以视作是显式的隐变量，即把句子生成过程中想要通过隐变量捕捉到的的生成因素显式地表现了出来。例如情感因素可能是标注者要生成一句对应视频的描述时所隐含的因素（常常是无意识的，因而被称作“隐变量”），比如当他看到一个暴力视频时候，内心会产生消极的情感，这也可能体现在他的标注中的消极概念的出现。因此控制信号显式的表达了出来，给了隐变量模型一个“可解释性”。

1.2.2.4　多样视频描述

本文研究的重点是多样视频描述（diverse video captioning），即输入一段视频，同时生成很多**独立**的描述它的单句。与密集视频描述和视频段落描述不同，这些单句之间不存在明显的时间顺序、因果联系；与可控视频描述任务不同，每个单句的产生没有显式的控制信号。这一任务的诞生主要是为了模拟输入端的“一对多”的分布，即每一个标注的受试都会根据自己对于目标的理解、以自己的认知和语言能力独立地给出各自的描述。其必要性可以参考引言部分。

多样视频描述的方法大致可以划分为两类：第一类是利用（隐变量）生成模型，通过从隐空间中多次采样，每一次采样可以视作是一句描述的生成；第二类则没有利用隐变量，而是在原有模型基础上（如，损失函数）增加了一些多样性约束条件。第二类方法往往需要使用集束搜索（Beam Search，简称 BS）在组成句子的单词的联合概率空间中进行贪心搜索，从而产生富有变化的描述。正如后文中所提到的，集束搜索受到集束宽度的影响很大，因而在产生大量的描述的场景下效率很低。注意到，同样地这里主要以图像描述作为综述的部分。

第一类又可以区分为：1）基于对抗生成网络（Generative Adversarial Network，简称 GAN）的。文章[82]利用条件对抗生成网络联合学习到一个生成器和评估描述质量的评估器（判别器），由于序列采样的过程是不可微的，文章采用了强化学习的方式训练模型。另一个基于 GAN 的文章[83]通过 Gumbel softmax 近似来解决梯度不能回传的问题。另一个方法[84]采用了比较对抗学习的架构，利用了同一个视觉信息对应的多个描述，和跨图像的不同描述间的关系，从而生成更有区分性和更加多样性的描述。2）基于 VAE 的。条件 VAE 广泛用于多样的图像描述领域，本文将它们分为两类。一类是聚焦在构建新的有依赖关系的结构化隐空间。例如 Seq-CVAE[85]利用了一个序列隐空间，COS-CVAE[86]设计了一个目标物体和上下文分离的隐空间，而 VSSI-Cap[87]则构造了词汇相关的和句法相关的隐空间。通过设计结构化隐空间可以增大隐变量之间的复杂关系，从而更容易捕捉出原始的多峰分布[88]，提高神经网络的可表达性。另一类方法专注于设计新的先验和后验分布假设。例如 AG(GMM)-CVAE[89]探索了可加高斯分布和混合高斯分布。LFNMM[90]则

采用正则化流（normalizing flow）去建模先验分布。本文所提出的方法正属于这一范畴，同时利用的是时序化可分离的隐变量空间来建模。

第二类方法没有隐空间假设。DBS（Diverse Beam Search）[91]在普通的集束搜索中引入了新的多样性约束，即搜索空间分为若干组别，组间生成的描述词如果不同，则可以获得一个多样性奖励。GroupTalk[92]将训练样本分为不同组，直接学习到一对多的这种多峰分布。GroupCap[93]则提出一个视觉树解析器去发掘不同视觉信息间的关系，之后将这些关系建模为多样性的约束。

1.3　本文的主要工作

本文基于 VAE 对于隐空间以及训练方式提出了一系列的改进方式，并设计了两个可以综合反映多样性和准确性的指标，以下是本文的主要工作介绍：

（1）本文在 VAE 基础上提出了一个动作和上下文（模板）的分离隐空间，即 ATVAE。通过分离出动作空间，模型更加专注于场景之间的互动关系，同时模板部分可以模拟多样的语言表达。同时额外增加了一个对比学习机制，可以进一步拉开句子间的可区分性。

（2）本文的第二部分在 ATVAE 基础上，进一步挖掘了同一组描述之间的关系，通过将输入的全体句子构成的全集分割为不同子集，解耦出主题和表达两个输入空间。之后文章提出了一种类似“训练-微调”的，双阶段训练机制即，STR，有效地捕捉不同的视觉信息（主题）和丰富多样的语言表达。

上述两部分都设计了精确的实验和详尽的分析，包括定性分析和定量分析：其中定量分析主要比较了本文提出的方法和一些先进的多样视频描述以及生成单句的视频描述在准确性和多样性两个方面的结果，同时消融实验可以证明不同模块的重要性。定性分析可视化了一些样例，用来进一步直观地展示视频描述的结果。

（3）针对现有指标单独衡量准确性和多样性，同时准确性仅仅考虑了精度一侧，因而无法从集合层面衡量与真值的匹配问题，本文在第三部分提出了两个可以融合这两方面的指标：hau 和 o2o。它们将整体的准确性对应于集合之间的距离问题，不仅仅考虑到从预测集合到真值集合的精度一侧，也充分考虑到另一方向的召回率一侧，因而可以全面反映生成的描述集合的性能。

本文通过问卷评测的方式去验证现有指标与人类评估的相关性。文章科学地设计了不同视频在不同模型下的预测结果比较，并通过收集不同的受试者对于评测问卷进行人工比较打分，再最终与本文提出的指标进行相关度分析，从而验证提出的指标确实可以一定程度反映人类的评估表现。

1.4　本文的结构安排

本论文围绕隐变量生成模型理论和多样化的视频描述自动生成任务展开介绍和讨论，以章节形式进行叙述，以下是各章节安排：

第 1 章：绪论部分，主要介绍视频描述任务的研究工作的背景与意义，重点阐述了本文研究的课题与众多学科如语言学等的交叉背景。国内外研究现状和发展趋势，这包含单句描述生成和众多多句描述生成任务；之后则介绍了本文的主要工作以及本文的结构安排。

第 2 章：背景知识部分，主要是对于本文用到的主要背景知识和技术进行了详细介绍，包含：生成模型、变分自编码器和条件变分自编码器。重点介绍了不同模型的结构设计、计算原理、对应的目标函数的推导等。主要使用概率图模型对各个生成模型进行了可视化展示。最后介绍了与本文任务相关的不确定性理论。

第 3 章：本章详细介绍了本文提出的第一个改进模型，即基于动作-模板分离的隐变量模型建模。详细介绍了问题正式化描述、对空间进行分离的方式以及模型的训练和推断过程。实验验证与分析从定性和定量两个角度对模型结果进行了说明。

第 4 章：本章建立在第 3 章的基础上，提出了一种基于渐进式训练方式。其行文结构与第 3 章类似，区别主要在于方法部分对话题聚类、渐进式训练和模型训练三部分做了更详细的介绍。实验验证和分析类似第 3 章。

第 5 章：本章详细介绍了本文所提出的集合水平的评估指标设计和评测。第一部分介绍了现有的评测指标，之后则详细介绍了本文提出的 hau 和 o2o，最后一部分详细介绍了本文设计的问卷评测，并分析了本文提出的方法与人类评测的相关性，从而来验证指标的有效性。

最后一章是结论部分，针对本文工作进行了总结，并对未来工作进行了展望。

第 2 章　背景知识

这一章介绍一些论文所采用的技术背景。包含生成模型，变分自编码器（VAE）、条件变分自编码器（CVAE）以及分离空间的 CVAE。由于 VAE 本身属于概率生成模型，有向图模型（directed graphical model）是一种常用的可视化和建模方法。

2.1　生成模型

本文所针对的是一个生成式任务——（视频条件下的）描述文字生成任务，同时本文所采用的模型包含用于生成文本的自回归模型以及隐变量生成模型，与诸如图片分类的判别式模型不同，生成模型试图学习所有变量之间的联合概率分布，并试图模拟真实世界中数据是如何产生的。为了表示不同生成模型的依赖关系，在下图中本文采用有向图模型这种常用的可视化和建模方式。在概率图模型中，随机变量用圆圈表示，其中被观测到的变量用阴影表示（如表示数据的 x），未被观测到的变量无阴影（如表示隐变量的 z），为矩形框表示重复的 N 组样本，它们往往作为局部变量出现，与之相对的矩形外部的变量（如参数变量 θ, ϕ）则因作用于所有样本，实为全局变量。箭头表示依赖关系，即箭头末端变量依赖箭头始端变量。[94]

2.1.1　自回归模型

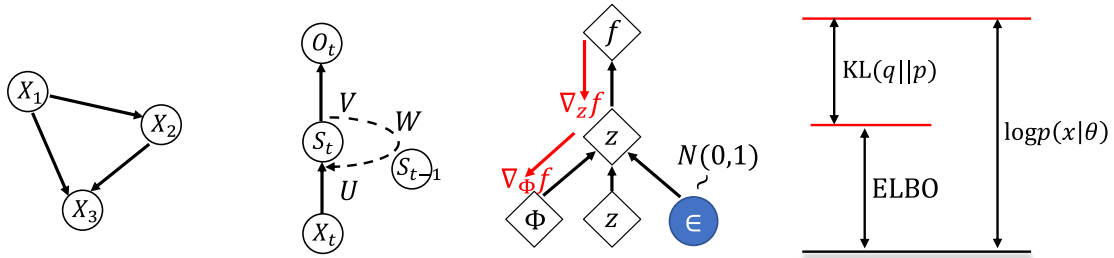
常见对于文本序列（或者高维数据）的建模都采用自回归的方式进行建模，即后面的序列依赖于它之前的序列。它将联合概率分布分解为如下，

$$p_\theta(x) = p_\theta(x_1, \cdots, x_D) = p_\theta(x_1) \prod_{j=2}^{T} p_\theta\left(x_j \mid Pa\left(x_j\right)\right) \tag{2-1}$$

其中，D 表示数据的维度（时间维度上的），$Pa(x_j)$ 表示 x_j 节点的所有有向图中的父节点的联合概率分布。图 2-1(a) 展示了 T=3 时候的自回归模型的图结构。自回归模型常常使用神经网络进行参数化，即：

$$p_\theta\left(x_j \mid x_{<j}\right) = p_\theta\left(x_j \mid \text{NeuralNet}_\theta^j\left(Pa\left(x_j\right)\right)\right) \tag{2-2}$$

这里常用到的神经网络为 RNN[95] 或者 transformer[50]。本文主要使用 RNN 对其进行建模，故这里简要介绍 RNN。RNN 的示意图如图 2-1(b) 所示。X，S 和 O 分别表示输入层、隐藏层和输出层，相应的层与层之间的权重使用 U，V 和 W 来



（a）N=3自回归模型　（b）RNN模型示意　（c）重参数化　（d）似然函数分解

图 2-1　生成模型、重参数化以及似然函数分解示意图

表示，下述公式可以计算相应的输出：

$$O_t = g\left(V \cdot S_t\right)$$

$$S_t = f\left(U \cdot X_t + W \cdot S_{t-1}\right)$$

由于自回归模型之后的序列需要以前之前的序列产生，即称作一种祖先采样（ancestrial sampling），因此它的计算复杂度是 $\mathcal{O}(D)$，即与序列的长度呈正相关。对于较长的序列生成任务来说，计算效率会降低很多。

2.1.2　隐变量生成模型

有向图中可被观测的变量常常来自数据 x，如图像、文本、音视频、标注、监测信号等，这些数据往往以各自的形式（连续或者离散的信号）在训练集或者测试集中出现。隐变量生成模型假设数据的生成过程是由一些不可观测到的变量导致的，它们常常用 z 来表示。隐变量属于模型的一部分，但由于不可观测，故不属于数据集的一部分。也因此隐变量的含义往往较为抽象且任意，例如手写数字的生成中“6”和“9”都会出现的弯曲特征[96]；图像生成中光照、亮度、角度、尺寸甚至风格[97-98]等。尽管有专门研究隐变量的可解释性[99]或者解耦性（disentanglement）[100]，很多工作并没有显式表明隐变量的含义。

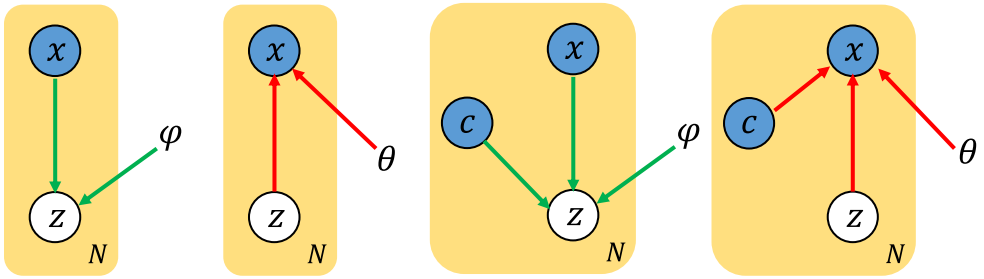
模型中 z 和 x 构成的有向概率图可以使用它们的联合概率 $p_\theta(x,z)$ 来表示，然而模型的训练往往针对可观测数据做最大似然估计，因此有必要求出它的边缘概率分布：

$$p_\theta(x) = \int p_\theta(x,z)dz \tag{2-3}$$

这一分布又被称为 θ 的（边缘）似然函数或者模型证据①

使用最大化似然函数，即公式 (2-3)进行优化参数时，最大的问题是该积分往往难以显式表达（intractability），由于当 z 为连续变量时。注意到由于似然函数的

① 即 evidence，属于统计学中贝叶斯理论的重要术语，用来表示确定模型做出推断决策的依据，常常用（观测到的）数据作为这一依据（如最大似然估计）。



(a) VAE+识别模型　(b) VAE+生成模型　(c) CVAE+识别模型　(d) CVAE+生成模型

图 2-2　VAE 和 CVAE 对应的概率图模型示意图

不可积，导致后验分布 $p_\theta(z|x)$ 同样不可积，这是由贝叶斯公式所知：

$$p_\theta(z \mid x) = \frac{p_\theta(x,z)}{p_\theta(x)} \tag{2-4}$$

需要注意的是联合概率分布 $p_\theta(x,z)$ 由于不涉及到求和操作，往往是容易求出的。

因此有很多方法去近似求出 $p_\theta(z|x)$ 以及 $p_\theta(x)$。一般的近似求出后验方法大致可以分为两类[101]: 基于采样的方式，如 MCMC 等，这类方法求出的解较为精确，但是往往需要对每一个样本点都推断出一个后验分布，导致效率低下，无法有效扩充到大规模的数据集；另外一种方式是利用变分的方式，通过估计一个近似的后验分布族再通过优化的方式寻找最优的后验。本文采用的是后一种方式，即变分自编码器模型（下一部分），它可以充分利用深度模型，并将其作用到大规模的数据上。

2.2　变分自编码器

VAE[102]假设数据的生成过程有一个低维的隐空间 $\mathcal{Z}$，从 $\mathcal{Z}$ 中采样得出的隐变量（latent variables）z 是数据 x 产生的条件，这一生成过程（图 2-2(b)）可以使用如下的公式表示：

$$p_\theta(x,z) = p(z)p_\theta(x|z) \tag{2-5}$$

其中：$p(z)$ 为（隐变量的）先验分布，$p_\theta(x|z)$ 为（数据的）条件似然分布，它由神经网络 θ 进行参数化。在训练过程中，由于数据 x 是已知的，因此可以推断出 z 的后验分布 $p(z|x)$（图 2-2(a)）。对于一般的数据集而言，$p(x)$ 的显式分布无法得出，因而无法求出后验分布的闭式解，一般只能采用近似的方法。不同于如马尔科夫链蒙特卡洛采样（Markov Chain Mento Carlo Sampling，简称 MCMC）等统计方法，变分推断（Variational Inference，简称 VI）将这一求出近似分布的问题视作一个优化问题：假设一个近似的后验分布 $q_\phi(z|x)$（为方便求解，该后验分布往往

是高斯分布）去逼近真实的后验 $p(z|x)$，其中后验分布由变分参数 ϕ 进行参数化。

为了学习到一个合理的变分参数 ϕ，模型优化如下的对数似然函数：

$$\begin{aligned}\log p_\theta(x) &= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x)\right] \\ &= \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{p_\theta(z\mid x)}\right]\right] \\ &= \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{q_\phi(z\mid x)}\frac{q_\phi(z\mid x)}{p_\theta(z\mid x)}\right]\right] \\ &= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{q_\phi(z\mid x)}\right]\right]}_{=\mathcal{L}_{\theta,\phi}(x)} + \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{q_\phi(z\mid x)}{p_\theta(z\mid x)}\right]\right]}_{=D_{KL}\left(q_\phi(z|x)\|p_\theta(z|x)\right)}\end{aligned} \tag{2-6}$$

由于第二项真实后验和近似后验的 KL 散度恒非负，便可以得到对数似然函数的下界，或者称之为 ELBO（Evidence Lower Bound，简称 ELBO），即

$$\begin{aligned}\text{ELBO} &= \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{q_\phi(z\mid x)}\right]\right] \\ &= \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x\mid z)p_\theta(z)}{q_\phi(z\mid x)}\right]\right] \\ &= \underbrace{\mathbb{E}_{q_\phi(z|x)}\log p_\theta(x\mid z)}_{=-\mathcal{L}_{\text{rec}}} - \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{q_\phi(z\mid x)}{p_\theta(z)}\right]\right]}_{=D_{KL}\left(q_\phi(z|x)\|p_\theta(z)\right)}\end{aligned} \tag{2-7}$$

从中可以看出 ELBO 可以分解为一项重建损失 $\mathcal{L}_{rec}$[①]和先验后验的逼近损失，它往往起到正则化的作用。而在实际计算重建损失时，往往采用蒙特卡洛采样的方式进行近似，即

$$\max_{\theta,\phi}\frac{1}{N}\sum_{i=1}^{N}\log p_\theta\left(x^i|z^i,c^i\right) \quad \text{s.t. } \forall i \quad z^i \sim q_\phi(z|x,c) \tag{2-8}$$

而在实际中，N 取值为 1 即可，这是由于训练数据往往会进行乱序，从整体“均摊”意义而言，仍然近似于做了平均。同时，由于对于隐变量采样函数是一个离散值，无法通过梯度回传的方式进行参数更新，往往采用一种“重参数化技巧”（reparameter trick）[102]。以高斯分布为例，$z = \mu + \sigma\epsilon$，其中 ϵ 来自一个单位高斯分布，即 $\epsilon \sim \mathcal{N}(0,1)$，图示参考 2-1(c)。

值得注意的是，通过最大化 ELBO，实则最小化真实后验和近似后验的距离，也就是尽可能使得近似后验逼近真实后验，如图 2-1(d) 所示。然而，模型毕竟优化的是一个上界，当上界不紧时，便有可能出现近似误差（approximation error）[104]，

① 这一术语来自编码理论（coding），在其中未被观察到的隐变量可以被解释为一个隐表示或者一个编码（code）。类似于自编码模型[103]，而该编码从后验分布中采样，又用于恢复原始数据，因而有“重建”的含义。

如何减小这一误差也是 VAE 系列方法需要考虑的问题。

由于神经网络强大的表征能力，往往利用它来得到近似的后验分布。具体而言，模型学习两套参数：模型参数 θ 和变分参数 ϕ，其实 ϕ 是从 x 推断 z，这部分又被称为编码器（encoder）或者识别模型（recognition model），最终网络的输出则是 z 的均值和方差；θ 是从 z 生成 x，这部分则被称为解码器（decoder）或者生成模型（generative model），如图 2-2所示。对于变分参数的学习而言，往往将从模型中推断的两个值作为均值和对数方差①，即：

$$\begin{aligned}(\mu, \log\sigma) &= \text{EncoderNeuralNet}_{\phi}(x)\\ q_{\phi}(z \mid x) &= \mathcal{N}(z; \mu, \sigma)\end{aligned} \tag{2-9}$$

尽管每个样本点都可以推断出各自隐变量的分布，但是它们共享同一套变分参数 ϕ。与传统的推断每个样本点各自的变分参数不同（例如高斯过程[105]），这种策略极大地提升了效率，也可以帮助模型拓展到大数据的场景下。这种方式也被称为分摊变分推断（amortized variational inference）[106]。

2.3　条件变分自编码器

上述 VAE 模型常常适合于无监督式的任务，对于大多数的监督式任务（如本文研究的视频描述），除了要重构的目标变量 x 以外，还有本身以 x 为标签的一个输入变量 c（对于视频描述任务而言，x 对应于文本，c 对应于视频）。图 2-2(c) 和 (d) 展示了 CVAE[107]的识别模型和生成模型。CVAE 的损失函数（ELBO 取负数）很容易由 VAE 推导出来，即，

$$\mathcal{L}_{\theta,\phi} = -\text{ELBO} = -\frac{1}{N}\sum_{i=1}^{N}\log p_{\theta}(x|z,\text{c}) + \text{D}_{\text{KL}}[q_{\phi}(z|x,\text{c}), p_{\theta}(z|\text{c})], \text{s.t. } \forall i\ z^{i} \sim q_{\phi}(z|x,\text{c}) \tag{2-10}$$

值得注意的是，CVAE 的提出[107]是用于结构化预测问题，即输出是个多峰（mode）的映射，这与本文的多样性描述输出是一致的。视觉的描述空间本身就是一个巨大的（连续）空间，存在很多输出的可能，因此将本文建立在 CVAE 上是合理的。

2.3.1　后验坍塌

后验坍塌（posteriori collapse）是 VAE 架构中很常见的问题，指后验分布几乎完全等同于先验分布，而隐变量被忽视掉了（相当于没有从数据中学习到有用的信息）。在 CVAE 的场景下，往往是由于过于强大的解码器[108-109]，使得模型无须

① 这里假设 z 变量的不同维度之间相互独立，即协方差矩阵为对角阵，为简化起见，下述公式表示每一维度上的采样过程。

z，直接从 c 就可以恢复出 x 了，使得隐变量无法发挥自身的作用（例如建模多样性）。一个经典的缓解后验坍塌的方式是对公式 (2-10)中起正则作用的第二项进行“退火”约束（KL annealing）①，即，

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \beta \mathrm{D}_{\mathrm{KL}}(q_\phi || p_\theta) \tag{2-11}$$

其中，β 即为一个“退火”函数，用于平衡忠实性（fidelity，即重构的准确性）和正则程度，进而平衡准确性和多样性。

2.4　不确定性

日常场景处理各种领域的不确定性，从投资机会和医疗诊断到体育比赛和天气预报，在所有情况下，目标都是根据收集到的观察结果和不确定领域知识做出决策。使用机器学习和深度学习开发的模型被广泛用于所有类型的推理和决策制定，这意味着在人工智能（AI）系统应用于实践之前评估其可靠性和有效性变得越来越重要[110]，因为此类模型所做的预测会受到噪声和模型推理错误的影响[111]。因此，在任何基于 AI 的系统中都非常希望以可信赖的方式表示不确定性。

不确定性主要可以分为两类，一类是数据不确定性，又称作 aleatoric uncertainty②；一类是模型不确定性，又被称作 epistemic uncertainty③。[111]在机器学习的背景下，数据不确定性是指标注的标签（人为认定的“真值”）与实际的“真值”之间的偏差。它刻画数据集中固有的随机因素，这往往与数据的生成过程④的随机性相关，也因此是通过增加数据量无法消去的。例如一般数据集的生成过程可以视作是由实验者收集的过程，比如在某些众包平台上进行标注类别，那么总会产生一部分的数据带有噪声；或者利用工具测量的长度，也无法避免精度上的误差等。模型不确定描述的是通过有限的训练集学习到的模型预测值之间的偏差，它往往是由于缺乏足够多的测试集相关的信息（知识）所导致的，且往往可以通过增大数据量来减小模型不确定性。

本文涉及到的多样性描述可以视作一种数据不确定性的表达。视频描述收集的过程中，由于视频场景的复杂交互以及语言结构的变化多样（参考引言部分），导致一段片段必须由多个受试者从不同的角度给出自己的描述 x^i。而假设真实模型对应的不可能出现的“真值”为 x^*，不同的受试者则可以视作在 x^* 附近增加了

① 尽管文献中往往把这种方式称作“退火”，但实际的操作并不是减少该项的值，而是增加，或许采用 warm-up 更加合适。
② 源自拉丁语，本意为“骰子”，含有随机性的意味。
③ 源自拉丁语，表示“与知识来源相关的”。
④ 这里假设数据是由一个真实模型产生。

不同程度的噪声 σ_i^2，即：

$$x^i = x^* - \sigma_i^2 \tag{2-12}$$

借用语言学中言语的意义起源于沟通交流这一说法[112]，这些噪声可以理解为各自的受试者对于目标视频自身的不同认知或意图（i）（比如着重留意视频哪一部分，表达整体还是细节，是否使用反语、幽默、讽刺、隐喻等），头脑中出现的对于某一片段的不同理解（主体发生了什么动作，要传达出什么意义），最终根据自己的语言教育、习得经历等给出一段语言表达（e）。值得注意的是，这些语言表达往往含有一些仅带有“表面含义”（s，standing meaning）的词语或者短语，它们往往是一些隐喻式的指代。但由于本文研究的描述场景是通用的场景，标注者往往被告知描述“表面”发生的行为即可，且句子仅只有一句话，不含复杂的上下文，因此往往不含有复杂的表面含义，这里将它忽略。可以将不确定性的来源，即噪声正式化为如下公式：

$$\mathrm{U} = \mathrm{var}(i) + \mathrm{var}(e) \tag{2-13}$$

例如在图 4-1中，来自 i 的差别可能在于不同的侧重点，有些人侧重发短信、有些则强调敲击、坐，有些另辟蹊径，认为这些画面是来自播放的电影等，这都来源于不同的主体的认知不同；而来自 e 的偏差则起源于形形色色的言语表达，引用部分已做了详细说明，这里不再赘述。

本文通过 VAE 去建模这种数据不确定性，不过与其它工作直接建模数据的方差[113]不同，VAE 将数据的不确定归因于隐变量 z 的不确定上，即通过推断 z 的后验分布（包含反映不确定性的方差），来间接反映数据中的不确定性。值得注意的是，由于每一个数据都对应一个生成它的隐变量（图 2-2 b），因此这种不确定性属于异向数据不确定性（heteroscedastic aleatoric uncertainty）[113]，即每个数据点都假设有不同的噪声①。

2.5　本章小结

本章是后面工作的背景技术介绍。首先回顾了包含自回归模型和隐变量模型的生成模型，它们构成了本文所用到的变分自编码器技术。之后就是针对隐变量模型中的 VAE 和 CVAE 进行了详细介绍。最后，由于一对多的分布是数据集收集过程中不可避免的“噪声”来源，它造成了之后不确定（多样性）的结果，因此最后一节对于不确定性的理论做了介绍。

① 与此相对的是同向数据不确定性，即 homoscedastic uncertainty，它假设所有数据都拥有一样程度的噪声。在神经网络损失函数的贝叶斯解释中，最小均方差损失即认为预测值服从高斯分布估计函数，由于认为该高斯分布的方差（噪声）对所有数据都一致，即同向不确定性假设，损失中与它相关的一项消去了。

第 3 章　基于动作-模板分离的隐变量模型建模

用自然语言描述视频对于人类来说本质上是一对多的翻译任务，因为受试者可能会从不同的角度给出几个侧重不同的描述。然而，大多数传统的视频描述模型都是以准确性为导向的，并且为视频生成单个“平均”描述，缺乏对人类、物体和环境之间复杂**交互**的全面理解。受此启发，本文基于条件变分自动编码器（CVAE），提出了动作和模型分离的隐空间模型，以最大程度地捕捉隐藏在交互中的多样性因素。具体而言，模板隐空间旨在编码潜在的语言方面的多样性，而动作隐变量编码蕴含在动作中的多样性。此外，一个新的对比正则化项通过增大为同一视频注释的两个随机句子之间的距离，来缓解潜在的模式崩溃（mode collapse）问题，从而在不牺牲描述的准确性的情况下提高多样性。三个数据集上的表现说明了本文提出的模型可以得到准确性相当，但更加多样化的描述。

3.1　引言

视频描述生成任务旨在生成多个有意义且流畅的句子来描述视频中的突出活动，这一领域在计算机视觉和自然语言处理领域越来越受到关注。随着几个大规模数据集[30-31,114]和先进的跨模态融合技术[51-53,115]的引入, 现有模型已经有可能产生令人印象深刻的性能。不过从这些模型中解码得到的序列通常遵循类似的模式，即选择具有最大概率的单词。因此，它们相对较短、简单[74,89,91]，未能捕捉到像人类标注一样的多样性。本文旨在生成更加多样的描述。

当前用于面向多样性描述的最先进模型[86,116]主要针对图像领域。虽然图像领域和视频领域在外观和目标上有很多共同之处，但视频自身就具有导致多样性的因素，即人、物体和环境之间的复杂交互，而这总是被人们所忽视。一方面，短片中的一个完整的突出活动由多个连续的动作组成。例如图 3-1中，在滑落降落伞的末尾吗，还有一个明显的抬升动作，这无法在静态图像中捕获到。这也可以反映在语言线索中，即时间连词，例如*之后*、*之前*以及*随着*等。另一方面，即使在同一帧内，不同的对象与环境的交互也很复杂，这给标注者提供了多种选择来表达场景的不同部分。例如，如图 3-1 所示，一些标注者强调人与船之间的交互，因此只呈现*推*和*拉*的动作，而如果标注者关注演员和降落伞的关系上，倾向于使用*骑着*和*降落伞*。然而，对于多样性图像描述的方法，由于复杂的相互作用而导致的这种多样性被忽略了。

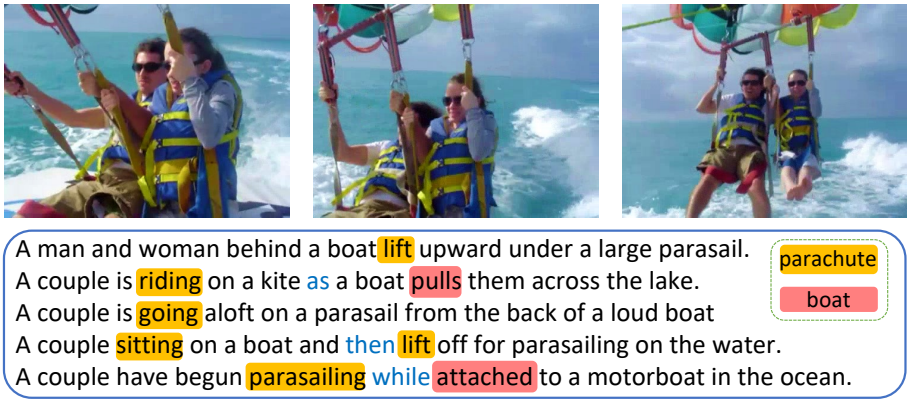


图 3-1　一个来自 VATEX 数据集中的视频描述的例子

为了对隐藏在交互中的不确定性进行建模，本文提出了一种具有结构化隐空间（structured latent space）的条件变分自动编码器的方法。具体来说，隐空间被分解为动词空间和模板空间，以更好地编码动作的多样性和语言上下文条件。同时，结构化的解耦的隐变量还可以增加隐空间的复杂性，潜在地提高真实后验分布近似的保真度（fidelity）[88]。本文将这种方法称为模板-动作分离的 VAE 模型，简称 ATVAE（**A**ction-**T**emplate split VAE）。此外，这种设计也与人类描述场景的方式一致：人类先想到突出的活动类型，之后填入到大脑习惯构建的句法模板[117]。

然而，这种带有一个强解码器的生成模型可能会忽略潜在变量，从而导致 VAE 框架中常被观察到的后验坍塌[108-109]问题，从而损害到多样性的表达。本文认为原因可能是视觉条件（即视频）的不同句子的后验分布重叠太多，因此在测试阶段从包含这些后验分布的先验分布中抽取的样本可能因为抽到了同一个后验分布而相似，最终导致缺乏多样性。为了缓解这个问题，本文设计了一种对比正则化项来“拉大”一个视频对应生成的句子，从而隐式地分散后验分布。这可以视作是对隐空间的一个正则化约束，与 ELBO 中的先验后验 KL 散度的一项起到的作用是类似的。

文章贡献如下：

- 文章提出了一个新的结构化隐空间来捕捉隐藏在复杂交互背后的多样性。
- 文章设计了一个对比正则化项来减轻潜在的后验坍塌问题并有效的增强了生成句子的多样性。
- 在三个基准数据集（即 MSVD、MSRVTT 和 VATEX）上，本文提出的模型在三个数据集上的大量实验都显示出在不牺牲准确性的前提下，结果有显著的多样性的提升。

3.2　方法

3.2.1　问题正式化描述

传统判别式的视频描述任务学习从一个视频[①]c 到一条描述 x 的映射，即条件似然概率：$p_\theta(x|c)$，其中 x 由 T 个词汇依序列构成，即，$x = (x_1, x_2, \cdots, x_T)$；$\theta$ 为似然概率的参数，如神经网络的参数。多样性描述问题可以视作一个一对多（多峰的）函数映射，或由输入 c 映射到一个由多个独立的描述句组成的集合，即：$f_\theta : c \to 2^{\mathcal{X}}/\emptyset$，后者表示所有可能输出句子 x 构成的空间 $\mathcal{X}$ 的有效子集（去除空集）。然而全集 $\mathcal{X}$ 在训练阶段无法完全获得，事实上只能获得一个稀疏的[②]训练子集，称之为参考集合 $\mathcal{R}$。整个学习过程就是试图估计这个多峰函数 f（或者对应的参数 θ），并且对于一个新的视频数据点 $\hat{c}$ 推断出一个合理的假设[③]集合 $\mathcal{H}$。

VAE 是一个基于低维隐空间 $\mathcal{Z}$ 假设的概率生成模型。这里同时假设一个序列 VAE，即组成一句话 x 的每个词汇 x_i 都对应一个生成它的低维 d 隐变量 z_i，其中 $z_i \in \mathbb{R}^d$。之后将序列生成模型定义到全数据空间 x, z 上，其基于隐变量的似然函数可以表示为：

$$\log p_\theta(x|z,c) = \log \sum_t p_\theta(x_t|x_{<t}, z_{\leqslant t}, c) \cdot p_\theta(z_t|z_{<t}, x_{<t}, c) \tag{3-1}$$

需要注意的是，由于训练集中的数据假设满足数据同分布，即 i.i.d.，且只显示出了一个数据点对应的似然函数。这一项表示的是目标函数（公式 (3-5)）中的第一项：重构损失。在训练阶段，由于数据是已知的，模型可以从不断学习中的近似后验分布 q_ϕ 中采样 z；测试阶段，模型则从学习好了的先验分布 p_θ 中采样 z。

3.2.2　分离空间的条件变分自编码器

增加隐空间的复杂性，如假设更多隐变量、假设它们之间存在更加复杂的依赖关系等可以提高对于真实后验分布的近似的拟合程度[88]。本文模型架构是建立在对于物体和上下文两类隐变量进行分离的 COS-CVAE[86] 上。然而，考虑到动作对于视频描述的重要性，因而，本文则对 COS-CVAE 进行改进，将针对于图像的物体空间 z^o 改为针对于视频的动作空间 z^v，仍保留去除动作后的模型空间 z^m。因此，整个隐变量空间可以被构建为 $z = [z^v, z^m]$，详细参考图 3-2。在生成的第 t

① c 代表 condition，在 CVAE 背景下，视频所代表的视觉信息相当于提供了一个条件信息，故未用传统的 v 来表示。

② 由于在收集描述时候，受于资源的限制，受雇的标注者有限，因此只得到有限集合，而理论上，对于一个场景可以有无限种表达语句。

③ 假设集合即 hypothesis set 的中译，表示预测集合。但传统的方法将生成的每一条语句并非认为是确定的、正确的，而用假设代替预测一词。

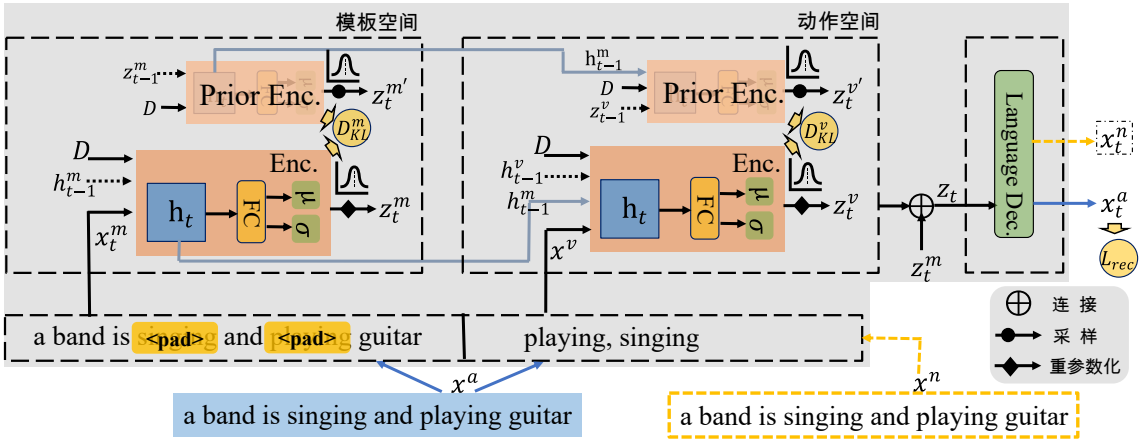


图 3-2　分离了模板空间和动作空间后的隐变量建模示意图

词汇前的隐变量可以表示为：

$$p(z_t) = p(z_t^v | z_{\leqslant t}^m, z_{<t}^v) \cdot p(z_t^m | z_{<t}^m) \tag{3-2}$$

先验分布 p 和后验分布 q 也可以被相应地分解为动作和模板这两部分，即，

$$p_\theta(z_t | z_{<t}, x_{<t}) = p_{\theta^v}(z_t | z_{<t}^o, z_{\leqslant t}^m, x_{<t}^v) \cdot p_{\theta^m}(z_t^m | z_{<t}^m, x_{<t}^m) \tag{3-3}$$

和，

$$q_\phi(z_t | z_{<t}, x) = q_{\phi^v}(z_t | z_{<t}^o, z_{\leqslant t}^m, x^v) \cdot q_{\phi^m}(z_t^m | z_{<t}^m, x^m) \tag{3-4}$$

最终，可以推导出分解的隐变量所对应的 KL 项为，

$$\mathrm{D_{KL}}(q_\phi || p_\theta) = \mathrm{D_{KL}}(q_{\phi^m} || p_{\theta^m}) + \mathrm{D_{KL}}(q_{\phi^v} || p_{\theta^v}) \tag{3-5}$$

3.2.3　带有对比学习正则化的解码器

为了进一步利用同一视频对应的句子之间的关系，拉开多个后验分布的距离，本文提出一个对比学习来帮助训练 VAE 模型。整体来说，一个视频下对应的句子之间应该尽可能不同，因此本文设立当前要解码的一个序列设为 anchor 句 x^a，将对应于同一视频下，不同的其它句子视作一个负样本 x^n。在时间 t 时，带有多样性奖励的目标函数定义如下：

$$\log p_\theta(x_t^a | x_{<t}, z_{\leqslant t}) + \lambda \Delta(x_t^a, x_t^n | x_{<t}, z_{\leqslant t}) \tag{3-6}$$

其中 Δ 是一个距离函数，用于衡量 x_t^a 和 x_t^n 之间的差异，λ 是一个平衡系数，用来控制多样性的程度。Δ 的选项有很多，比如汉明距离，本文采用一个简单的方式，即负的样本 x_t^n 对应的概率。注意到这里本文没有使用对数函数。因此，最终

的重建损失定义为：

$$\mathcal{L}_{\text{recon}} = -\log p_\theta(x_t^a | x_{<t}, z_{\leqslant t}) + \lambda p_\theta(x_t^n | x_{<t}, z_{\leqslant t}) \tag{3-7}$$

注意到这里使用的对比学习有些类似于 DBS[91]对于 BS 的改进，即在采样时候增加多样性奖励，但区别在于这里的对比学习的方式适用于训练阶段，而 DBS 主要是测试的采样阶段，因此不影响损失函数的设计等。

本文采用的解码器旨在恢复原始数据 x。由于模型专注于构建隐空间和提高多样性，解码器的设计都尽可能简单。具体来说，如图 2 所示，模型采用了两层的 LSTM：第一层带有自顶向下的注意力层，它以 z 和平均池化后的视觉特征作为输入，输出的隐变量则用于之后关注视觉的运动特征；第二层是语言解码层，用于接收第一层的输出特征并在每个时间点获得每个单词的生成概率。

3.2.4　模型训练和推断

整体损失则包含两部分：重构损失和 KL 损失，如下：

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \beta \mathrm{D}_{\mathrm{KL}}(q_\phi || p_\theta) \tag{3-8}$$

其中 β 是一个（反）退火函数（annealing function）[118]，用来平衡准确性和多样性。之后的所有实验 β 的在第一轮训练的值都设为 0.2，剩下的设置为 $\beta_{\max}$。

训练阶段，近似后验分布用于采样隐变量，这里由于整个训练阶段可以分摊采样，这里仅抽取一次，该隐变量作为解码器的特征，重建出原始的数据。在这个过程中，代表后验分布的隐变量编码器逐渐学习到一个条件于视频的、可以恢复原始句子的低维向量（流形）。而在测试阶段，模型只需从学习到的条件先验分布中采样，就解码出该隐变量对应的描述了。通过采样多次，可以产生多样的描述。

3.3　实验验证与分析

3.3.1　数据集

实验主要采用了三个规模较大的开放话题的常见数据集。MSVD 数据集[114]包含 1970 段从 Youtube 上截取的开放领域的视频。每个片段主要包含一个主要动作，大概持续 10 到 25 秒。数据集提供多语的视频描述，文章中只用到了英语，其中共有 85500 条描述，每个片段大约 40 个描述。文章采用常见的训练/验证/测试集的划分方式，分别为 1200 段，100 段和 670 段。MSRVTT 数据集[30]是另一个规模更大的数据集，它由 1 万段网络视频片段构成，每个片段包含 20 条描述。其中视频场景可被划分为 20 类。训练/验证/测试数据集的划分分别为 6513/497/2990。VATEX[31]是另外一个规模更大的双语视频描述数据集，它共包含 41250 段视频片

段以及每段视频附有 10 句中文描述和 10 句英文描述，共 825000 条。本文仅使用英文描述。与前两个数据集不同的是，VATEX 拥有更大规模的视频-描述对。同时 VATEX 数据集中的视频场景复杂且内容丰富，总共有 600 个人类活动①，也因此对应描述的词汇更加复杂多样，句型结构、语义表达更加丰富多变。该数据集保证所有的描述在数据集均只出现一次。

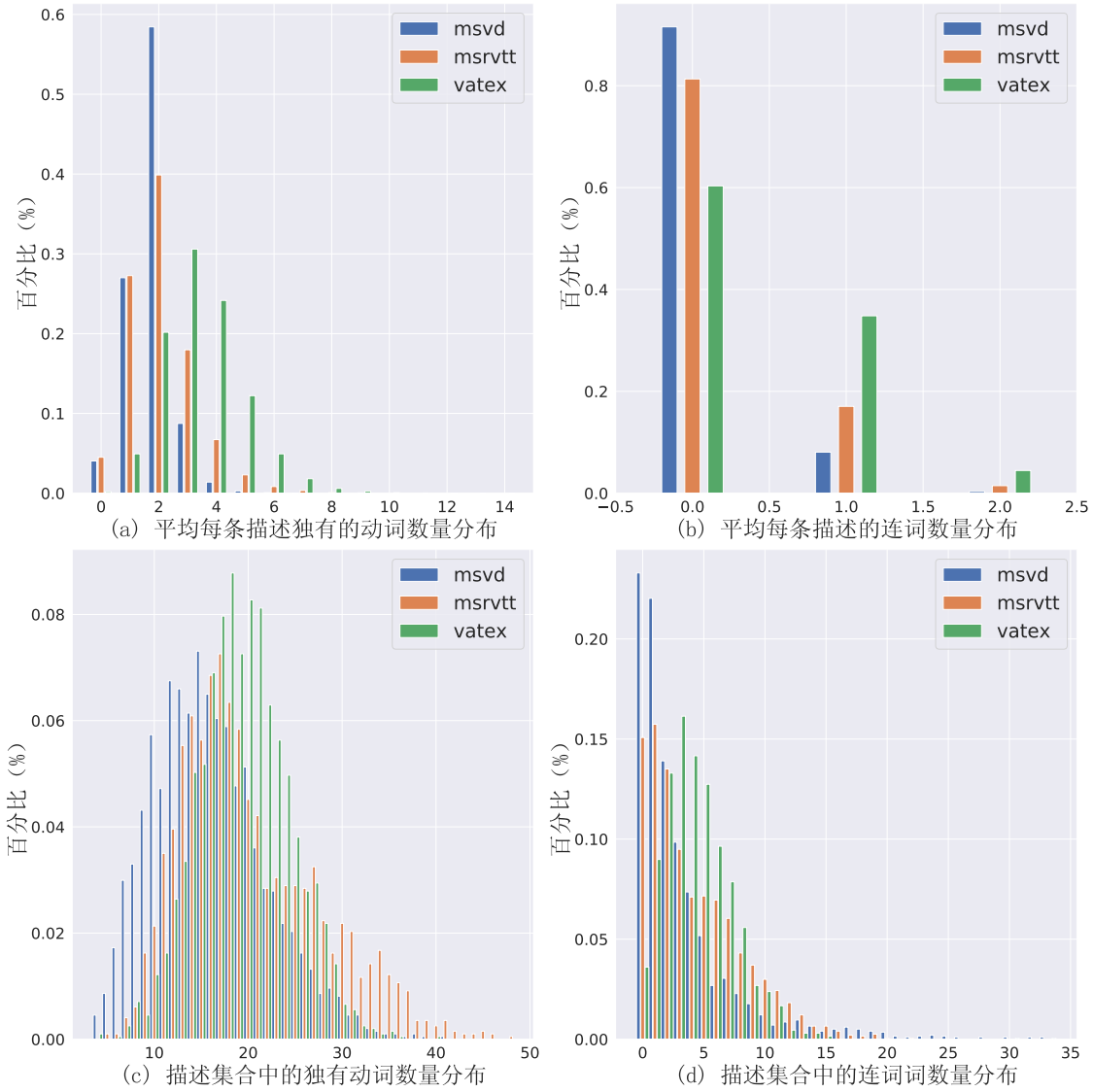


图 3-3　三个数据集中一些感兴趣的统计量的分布

图 3-3显示出了三个数据集的统计差异。由于动词可以刻画一句描述中的突出事件，图 (a) 和图 (c) 分别展示了平均每条描述所包含的独有动词数量分布（全局）和平均每个描述集合所包含的独有动词数量分布（局部）。可以看出 VATEX 多于 MSRVTT，MSRVTT 多于 MSVD，这与它们三个数据集的复杂性是成正相关的。值

① 这是由于 VATEX 数据集的视频是基于 Kinetics600[119] 的。

得注意的是，VATEX 平均在一个描述集合里面便有超过 20 个动词，说明其视频场景中发生了较为复杂的人物、环境互动，进一步表明单靠一句描述严重不足以刻画视频中复杂的场景，应该将多样性描述考虑其中。同时观察到 VATEX 数据集中往往有更加复杂的句型结构，一个突出表现是，很多描述中都会出现如：和（and）、或者（or）等连词，图 3-3中的 (b) 和 (d) 则输出了平均每条描述和描述集合中的连词数量分布。可以观察到，VATEX 平均一句描述可能具有 1 个及以上的连词，这也反映在图 (a) 中 VATEX 单句往往具有较多的动词上（连词往往连接多个语义动词）。这也间接说明 VATEX 单句便可以反映更多的语义，这也体现了 VATEX 数据集在收集过程中有着严格的质量控制。

3.3.1.1　评估指标

整体　文章使用新提出的 hau 衡量描述集合的整体性能。第五章详细介绍了 hau 指标。

准确性　文章报告 Oracle 条件下的（即最佳挑选出的描述的结果）Bleu-4（**B4**）、Meteor（**M**）、Cider（**C**）、Rouge（**R**）和 Spice（**S**）来表示一个上限值。其中使用视频的真值字幕搜索最佳字幕。由于在测试期间无法获得真值，本文按照 COSCVAE[86] 的方式使用共识-重排序（Consensus Re-ranking，简称 CR）的方式获取，即采用图文预训练模型 CLIP2Video[120] 检索出与预测视频最相似的训练视频，用该训练样本的标注去搜索最佳描述。

多样性　文章将多样性指标分为局部的和全局的：（1）全局多样性考虑为所有视频生成的所有字幕。**Gunis** 计算所有候选字幕中独一无二的句子百分比。文章考虑到动词对于视频反映主题的重要性，计算了 **guniv**，即每个句子的独特动词的平均数量；（2）局部多样性衡量一个视频对应的字幕集的多样性，然后返回所有视频的平均分数。**Div1** 和 **Div2** 是不同的 unigrams 和 bigrams 与 *n*-grams 总数的比率；**luniv/lunis** 表示集合中独一无二的动词/句子的数量。MBLEU-n 表示视频的每个字幕和其余字幕之间的平均 BLEU-n 分数。简约期间，文章只报告 **mix-mB** $= 1 - \frac{1}{4}\mathrm{mBLEU_n}$；Self-Cider（**Self_C**）[121] 计算成对相似度矩阵的最大特征值的比值，与人类评价的相关性比 mBLEU 强。对于所有指标，更高的值代表更好的多样性。

3.3.2　执行细节

预处理　文章参照论文 RMN[51]，词汇的收集是在去除标点符号和训练集中的稀有词，并将描述转换为小写，最终 MSVD 数据集含有 7,351 个词汇，MSR-VTT 则

含有 9,732 个词汇，VATEX 含有 10,525 个词汇。每个词对应一个 300 维的嵌入向量。描述的词类标注由 StanfordCoreNLP 提取。动词的词汇表是由在数据集中至少出现了 4 次的动词构成，最终 MSVD，MSR-VTT，VATEX 中的动词数量大小分别为 1,137，2,333 和 4,294。特殊符号用来进行掩膜动词。

特征提取　在数据集 ImageNet[122]上预训练的模型 InceptionResNetV2[123]用于提取视频每一帧的空间特征，而在 Kinetics 数据集[119]上预训练的 I3D 模型[124]则用于提取时空特征。这两种类型的特征连接起来，得到的平均池化后的输出作为视觉特征向量 c。

实现细节　模型使用动量为 0.9 且权重衰减为 0.001 的 SGD 优化器。第一阶段和第二阶段的初始学习率分别为 1.5×10^{-2}，并随着每轮迭代线性下降。隐变量的维数和隐藏层维度大小分别为 128 和 1024。公式(4-9)中的 $\beta_{\max}$ 超参数用于调节生成描述的多样性和准确性，是一个重要的超参数。最终，MSVD 数据集的 $\beta_{\max}$ 设置为 0.5；MSRVTT 数据集和 VATEX 则分别为 0.7 和 0.9。文章采用度量 *hau* 作为实验早期停止的标准，最大的容忍轮数为 20，最大轮数设置为 100。

3.3.3　性能比较

与多句描述方法相比较　表 3-1和表 3-2分别展示了几种常见的视频多样性描述方法的准确性和多样性的表现。为了获得更高的 Oracle 的得分，模型需要关注某些可能的描述上，并且把预测这些的描述的概率生成的足够高。[86] 从结果上看，本文提出的模型 ATVAE 可以取得最佳的 Oracle 评分。可以注意到，ATVAE 引入分离隐空间主要是为了增加多样性，相比另一个同样是隐空间模型的 Seq-CVAE，ATVAE 则在 Cider 和 Bleu-4 上都有提升。这可能是因为强调动作的分离空间可能更加可以挖掘运动特征，而运动特征在视频表示中起着重要作用。从多样性的指标看，ATVAE 则超出其它方法很多，尤其在两个规模比较大的数据集 MSRVTT 和 VATEX 上。注意到 MSVD 上本文的方法效果并不显著，这可能是由于 MSVD 的描述集合本身的多样性有限，动作描述单一（图 3-3）导致的。相比传统经常用到的 BS 采样方式，ATVAE 可以大幅提升多样性，这也表明学习到的先验隐变量空间很好的捕捉到了原始的多峰分布。同时高效的采样方式也是相比 BS 的一大优点。与 Seq-CVAE 相比，ATVAE 的优势有所缩小，这是由于它也是隐变量生成模型，与 ATVAE 仅仅在于没有分出动作隐空间，但提升仍然很明显，尤其在 mix-mB 的指标上，这说明 ATVAE 可以生成区别性更大的句子。。此外，平均而言，ATVAE 生成的句子或者句子集合中含有具有独特的动词，这反映在 luniv 和 guniv 中。

数据集	方法	整体性	上界准确性				
		hau	B4	C	R	M	S
MSVD	COS	22.5	45.9	105.3	78.9	52.8	-
	Seq-CVAE	24.1	50.7	113.4	81.0	**57.8**	8.4
	BS	21.5	47.8	108.6	78.9	52.7	-
	ATVAE（本文）	**24.4**	**53.1**	**113.9**	**81.6**	57.5	**8.4**
MSR-VTT	COS	19.1	41.8	63.6	68.5	41.8	-
	Seq-CVAE	19.6	**44.9**	64.5	**69.7**	43.2	**10.8**
	BS	16.1	39.1	59.4	67.2	40.7	8.3
	ATVAE（本文）	**19.6**	44.0	**65.1**	69.3	**43.2**	10.5
VATEX	COS	17.9	34.4	67.4	55.1	28.2	-
	Seq-CVAE	18.1	36.9	69.6	56.7	28.6	-
	BS	17.4	**38.5**	**71.0**	56.9	**28.8**	-
	ATVAE（本文）	**18.2**	37.5	70.3	**56.9**	28.7	-

表 3-1　两个数据集在不同方法上的整体性和准确性结果

与单句描述方法相比较　表 3-3使用 Cider、Rouge 和 Meteor 的 Oracle 得分来表示单句描述方法的准确性。多样性方法从全局角度衡量，即整个数据集中每个句子的独特句子和动词。数据表明 ATVAE 在不损失准确性的前提下，可以大大提高生成句子的全局多样性，即生成更多独特的句子和动词。

3.4　消融实验

本节针对实验中的一些主要功能模块进行消融实验，主要探究三个方面的影响，一是分离空间的有效性；二是对比学习的作用；三是 KL 正则项系数的影响。以下实验结果是在 MSRVTT 数据上的，注意到以下数据与表 4-1和表 4-2报道的不一致，原因是这里选择超参 β 选择为 0.8，而非 0.7。

3.4.1　分离隐空间

表 3-4展示了隐空间的不同假设下的结果。其中，w/o 表示完全没有隐空间假设，这和普通的确定性的描述模型是一样的，不过这里采用的特征和带有隐变量的模型是一致的，都是比较简单的视觉特征。之后比较了不同类型的隐空间，包含分离出名词空间（名词和去掉名词的上下文）z^n 以及随机去掉一些词和上下文的

数据集	方法	多样性							
		Div1	Div2	lunis	gunis	mix-mB	Self-C	luniv	guniv
MSVD	COS	26.4	34.6	49.1	40.6	11.7	28.9	1.46	**1.02**
	Seq-CVAE	28.0	37.8	55.7	44.0	13.8	35.1	1.52	**1.02**
	BS	**30.3**	**45.0**	**98.3**	**53.4**	**20.4**	**58.7**	**1.83**	0.95
	ATVAE（本文）	27.7	37.9	57.5	45.0	14.0	36.2	1.49	1.01
MSR-VTT	COS	38.5	58.6	88.1	75.7	35.1	66.6	3.01	**1.22**
	Seq-CVAE	39.9	61.0	89.3	75.2	36.4	68.2	2.78	1.17
	BS	28.6	41.1	**96.9**	39.2	14.2	-	1.87	1.09
	ATVAE（本文）	**43.6**	**66.7**	93.1	**77.4**	**43.6**	**73.4**	**2.88**	1.20
VATEX	Seq-CVAE	**32.2**	51.0	90.0	88.0	25.3	61.1	5.29	2.55
	BS	26.4	37.6	**99.7**	77.6	12.3	-	3.87	2.24
	ATVAE（本文）	27.7	**58.3**	96.6	**94.9**	**33.5**	**69.3**	**6.11**	**2.62**

表 3-2　两个数据集在不同方法上的多样性结果

z^r。准确性方面报道了采样 20 次后的 Oracle Cider 和 Oracle Meteor，其中 w/o 的情况则是从 BS 中采样 20 次。它们根据 CR 排序后的前五名用于测试多样性，这里报道了部分的多样性指标。可以看出分离了动词隐空间的情况可以得到最有的准确性以及超出大多数的多样性，表明该策略的有效性。

3.4.2　对比学习

针对文章中提出来的对比学习，表 3-5报道了 MSVD 和 MSRVTT 上使用（CL）和不使用对比学习（NO CL）下的 kl 损失（KL）和重建损失（即负的的对数似然值）（NLL）。这里使用先验后验的 kl 损失去表征后验坍塌的程度：kl 损失越大表明坍塌的程度越轻微，否则则越严重。重建损失则反映重建的效果，如果重建的效果越好，该项的损失越小。从结果中可以看出对比学习可以有效的增加还原的忠实度（fidelity）即准确度，却不造成过大的坍塌，从而保证生成句子的多样性。

3.4.3　正则化系数

公式 (3-8)中 β 即正则化系数对于控制准确性和多样性的平衡起到重要作用。图 3-4展示了不同的 β 值对于模型多样性和准确性的影响。其中准确性使用 Oracle Cider 指标，多样性使用 mix_mB。不难发现，随着 β 值的增大，多样性是单调递

数据集	方法	C	R	M	gunis	guniv
MSVD	COS	87.2	69.3	34.3	40.57	**1.02**
	SAAT	81.0	69.4	33.5	44.0	1.00
	GRU-EVE	78.1	**71.5**	35.0	-	-
	ATVAE（本文）	**90.9**	71.4	**35.8**	**45.0**	1.01
MSR-VTT	COS	46.6	54.0	25.9	75.7	**1.22**
	SAAT	**49.1**	**60.9**	**28.2**	31.0	1.04
	POS-CG	43.4	60.1	26.8	-	-
	ATVAE（本文）	46.7	53.7	25.5	**77.4**	1.20
VATEX	COS	45.3	42.8	20.6	99.6	1.20
	VATEX[31]	44.3	46.9	**21.6**	-	-
	BS	45.0	**47.6**	21.3	**99.9**	2.24
	ATVAE（本文）	**48.7**	45.0	21.4	94.9	**2.60**

表 3-3　与生成单句话的视频描述方法的准确性和多样性比较

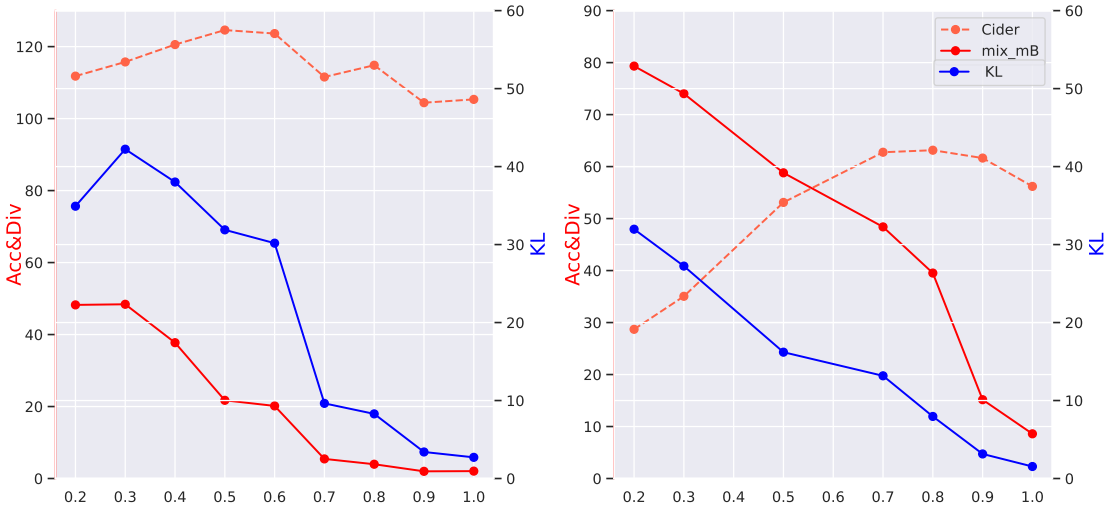
z	C	M	Div1	Div2	mix-mB	lv
w/o	63.6	43.6	25.6	32.6	6.0	2.61
z^n	65.1	43.5	33.1	48.5	26.1	3.54
z^r	66.2	43.8	37.9	56.4	**34.2**	**4.18**
z^v	**67.7**	**44.7**	**38.6**	**57.6**	33.4	3.90

表 3-4　不同策略下的分离隐空间方式对应的准确性和多样性评估

减的，而 Oracle 准确性则先上升后下降，这与实验前的预期相符合，可能的原因是：β 值越大，对于正则化的优化越优先于重构损失，因而后验分布越接近简单的先验分布，导致更加严重的后验坍塌（KL 损失值越来越小），从而导致从不同的后验分布采样就像从一个单一的先验分布中采样一样，生成越来越不多样的描述。而对于 Oracle 的准确度而言，即选择可能的最好的描述的准确度先是随着描述间的混乱程度变大而增加，而之后也会受到后验坍塌的影响，导致准确性受到

Dataset	Method	KL ↑	NLL ↓
MSVD	NO CL	6.38	21.16
	CL	**12.76**	**20.51**
MSR-VTT	NO CL	7.96	31.48
	CL	**8.30**	**31.45**

表 3-5　两个数据集上对比学习的效果比较



(a) MSVD上的性能随 β 的变化　(b) MSRVTT上的性能随 β 的变化

图 3-4　随着 β 值的变化，MSVD 和 MSRVTT 数据集在准确性（红色虚线）、多样性（红色实线）和 KL 损失（蓝色实线）的变化。其中 KL 损失反映了坍塌程度。

损失。因而会出现一个拐点。这个拐点，类似于 K-means 中选择 K 的拐点，也刚好可以成为准确性和多样性的一个权衡，因而可以作为选择超参（例如对于该例中 MSRVTT 数据集的 β 值选择为 0.8，MSVD 数据集的 β 值选为 0.5）或者程序提前终止的一个条件。

3.5　本章小节

本章提出了带有对比学习正则化的分离动词隐空间的方法（ATVAE）来生成多样化描述。它可以有效地捕捉物体与环境之间的复杂互动，从而尽最大程度恢复原始数据中的多峰分布。值得注意的是：下一章提出的训练方式每一阶段都是建立在本章提出的方法上的。

第 4 章　基于渐进式训练的多样视频描述

本章在第一部分首先对于 VAE 背景下的视频多样性描述生成任务做了数学上的正式化定义，第二部分介绍了方法部分，第三部分介绍实验部分。

4.1　引言

视频描述通常指使用单个命题式的句子描述一个短的视频片段。“命题式”要求描述的语句为可判断真假的陈述句，而非无法判定真假的开判断或者疑问句；“描述性”则侧重于通过语言去如实反映视频中“表面”上发生的情景，而非采用隐喻、夸张、反语等修辞手段。这些特点表明该任务主要适用于人类使用自然语言①的日常场景下，提到提示、总结或者教育意义。如，新闻的标题摘要，对于盲人的道路导航，短视频标题，指导儿童看图说话等。同时，不追求语言技巧以及对于场景的深刻理解也表明了单独完成该任务对人类而言并不费力，但大量的重复性工作（如，海量的短视频平台）却仍然有机器标注的市场需求。然而视频场景和自然语言的复杂性却对机器而言异常困难，这属于典型的一个人工智能完全问题（AI-complete）。

另一方面，给定一个视频片段，往往有多个合适的候选描述，可以称其为一个“一对多”的映射，或者统计意义上的“多峰性”（multimodality）的任务。现有的多数视频标注的模型[51-52]大都忽视了这一点：它们往往是确定性模型，只是在最终庞大的时序概率空间采用一种称为集束搜索（Beam Search，简称 BS）的贪心算法进行采样，得到的结果往往多样性很差，其目的也是搜索到最好的一个描述，而非考量多样的结果。多样性并不受到主流方法的重视，然而基于以下几点多样性有必要被考虑进去：1）**重构**的要求。在机器学习和不确定性理论中，测试集的目标就在于复现训练集中的样本分布，即若测试集中的样本点距离训练集中的样本点越近，其分布就应该与训练集中该样本点的标签分布越一致，这一处的不确定性越小，反之亦然[111]。通过观察主流的一些基准数据集，不难发现其训练集本身就以一对多的方式进行标注，例如 MSRVTT[30] 数据集平均一个视频有高达 20 条可能的表述。因而，对于一个给定的测试样本，理应反映出这种多峰性。2）**视频场景**本身就存在比较复杂和不确定性的因素，在给定一句简要的描述限制下，不同的标注者会根据自己的偏重找到不同的侧重点。3）**自然语言**具有歧义性、模糊

① 自然语言（Natural Language）是指人类通过重复使用习得的语言，以语音信号、文本、手势等形式展现，与程序语言相比，具有歧义性、非结构性、多样性等特点。

性，对于同一个场景本身就可以使用多种方式去表述。

为了捕捉这种多样性，句子生成过程可以看作是一个基于隐变量 z 假设的生成模型，并且使用变分自编码器的方式推断出 z 的后验分布 q（posterior distribution），从而将多样性建模成一个隐空间采样的过程。值得注意的是，基于 VAE 的方法已经在多样的图像字幕描述领域中有相关应用。它们大多数构建一个新的 z 空间，例如，AGVAE[89] 探讨了可加高斯隐空间和高斯混合空间；Seq-CVAE[85] 构建了一个时序隐空间；COS-CVAE[86] 则在时序隐空间基础上结构化变量，即分离了目标物体（object）和上下文（context）；VSSI-VAE[87] 从语言端着手，分离了句子中的词汇（lexicon）方面和句法（syntax）方面的隐变量。但这些模型都只针对单句描述，只发掘了句子内部词汇或短语短句之间的关系，忽视了原始的一对多的分布，没有充分挖掘句子与句子之间的联系，或者（句子所构成的）集合内的关系。由于模型的目标是重现训练集中的这种一对多映射，因此有必要发掘数据中本身就有的多峰分布。

多方面的原因导致了这种多峰性：1）**话题不同**：即使是包含一个主要动作的短视频，其视觉场景也往往比较复杂，标注者（annotator）仍可能因强调不同的侧面给出不同话题的描述，如不同的突出动作，不同时间段下的微动作等。2）**语言表达不同**：不同标注者之间的语言表达方式不同，反映在不同的词汇、句法、时态等语言特征上的差异。图 4-1展示了当前的一个基准数据集中由将近 20 位标注者对于“一个小女孩坐在床上休息发短信”的短场景进行的描述。根据它们语义进行聚类后，不难发现，可能的标注便可能侧重于“发短信”（组别 1），或者“坐着”（组别 3），因为这两个动作几乎同时发生；但期间发生的一个微动作，如“敲墙”（组别 2），也可能会被多个标注者捕捉到；值得注意的是，另一类的句子是以“画外音”的形式对这个短视频做一个整体的描述，如描述这是一个电影镜头（组别 4）。另一方面，即使是描述同一个话题，语言表达也不尽相同，以第 3 组的“坐着”为例，有的放置在表语处（句 2），有些放置在状语处（句 4），有些用“在床上”（in bed）（句 1）表示，有些则突出“坐”这个动作（句 2），或者更具体化（句 3，4）。其中可以把描述同一话题，但不同的语言表达的句子称之为“视觉同义句”（visual paraphrase）[125]。其中将一个最普通、简要的句子称为该组内的一个“中心句”（core sentence）（在图 4-1中用红色标注出该句）。因此其它的句子都可以看作是根据这个中心句进行同义改写。

为了捕捉到这种集合层面的多样性，隐变量的学习阶段被解耦，即采用一种渐进式的训练方法。具体而言，第一阶段，聚类方法可以构造出一个只包含中心句的训练子集 $\mathcal{D}'$。模型从这个稀疏的集合中期待学习到一个**话题**导向的隐变量。第

分组1: texting
1. girl sending sms to some one
2. someone texting on an old phone
3. a woman lying in bed texting
4. a woman taps on her bedroom wall before texting on her phone

分组2: knocking
1. a girl in bed knocking on the wall
2. a girl is knocking on the wall
3. a girl knocking on a wall
4. a woman is laying in bed knocking on the wall
5. girl knocking on the wall

分组3: sitting
1. a girl in bed
2. a woman is sitting
3. a woman is sleeping in a bed in a bedroom and looking at a phone
4. a woman sitting on a bed

分组4: movie playing
1. a clip from a movie is playing
2. a young girl in a horror movie is haunted
3. scene from a tv show

图 4-1　来自 MSRVTT 基准数据集的一段视频的例子

二阶段利用整个训练集合 $\mathcal{D}$ 去微调第一阶段得到的模型。由于第一阶段已经学习到一个稀疏的话题导向的模型，第二阶段期望在这个基础上，模型学习到**表达**导向的隐变量。总结来说，第一阶段的隐变量空间期待是稀疏的、话题的、普通的；第二阶段则是稠密的、表达的、丰富的。第二阶段是在第一阶段识别足够的话题后，用多种多样的语言表达去丰富它，从而更多捕捉到集合层面的多峰性。

文章的贡献如下：（1）本章提出了一个渐进性的训练策略，称之为 STR（**S**how **T**ell and **R**ephrase）[①]去分别构建话题导向的和表达导向的隐空间，从而捕捉更多集合层面的多峰性。（2）本文在多个数据集上做了详尽的实验，可以验证提出的方法可以得到更优的性能。

4.2　问题正式化描述

传统判别式的视频描述任务学习从一个视频[②]c 到一条描述 x 的映射，即条件似然概率：$p_\theta(x|c)$，其中 x 由 T 个词汇依序列构成，即，$x = (x_1, x_2, \cdots, x_T)$；$\theta$ 为似然概率的参数，如神经网络的参数。多样性描述问题可以视作一个一对多（多峰的）函数映射，或由输入 c 映射到一个由多个独立的描述句组成的集合，即：$f_\theta : c \to 2^X/\emptyset$，后者表示所有可能输出句子 x 构成的空间 $\mathcal{X}$ 的有效子集（去除空

① 本文将这一渐进式的训练理解为人类在理解一个场景的过程：面对（show）某场景，人类首先根据突出的事件做出描述（tell），之后用不同的方式去表达这一描述（rephrase）。

② c 代表 condition，在 CVAE 背景下，视频所代表的视觉特征提供了条件信息，故未用传统的 v 来表示。

集）。然而全集 $\mathcal{X}$ 在训练阶段无法完全获得，事实上只能获得一个稀疏的①训练子集，称之为参考集合 $\mathcal{R}$。整个学习过程就是试图估计这个多峰函数 f（或者对应的参数 θ），并且对于一个新的视频数据点 $\hat{c}$ 推断出一个合理的假设②集合 $\mathcal{H}$。

VAE 是一个基于低维隐空间 $\mathcal{Z}$ 假设的概率生成模型。这里同时假设一个序列 VAE，即组成一句话 x 的每个词汇 x_i 都对应一个生成它的低维 d 隐变量 z_i，其中 $z_i \in \mathbb{R}^d$。之后将序列生成模型定义到全数据空间 x, z 上，其基于隐变量的似然函数可以表示为：

$$\log p_\theta(x|z,c) = \log \sum_t p_\theta(x_t|x_{<t}, z_{\leqslant t}, c) \cdot p_\theta(z_t|z_{<t}, x_{<t}, c) \tag{4-1}$$

需要注意的是，由于训练集中的数据假设满足数据同分布，即 i.i.d.，且只显示出了一个数据点对应的似然函数。这一项表示的是目标函数（公式 (3-5)）中的第一项：重构损失。在训练阶段，由于数据是已知的，模型可以从不断学习中的近似后验分布 q_ϕ 中采样 z；测试阶段，模型则从学习好了的先验分布 p_θ 中采样 z。

4.3　方法

本工作主要可以分为两个步骤，第一个步骤，训练集中的参考集合 $\mathcal{R}$ 根据语义进行聚类，将语义空间分割为相互独立的子空间；第二步，本文提出一个渐进式训练机制，来充分挖掘句子之间的关系。

4.3.1　话题聚类

本方法在第一阶段对 $\mathcal{R}$ 中的句子进行语义聚类，以挖掘样本之间的关系。对于一个含有 M 个描述的参考集合 $\mathcal{R} = \{x^m\}$，本文通过一个预训练的特征提取模型 F 将每句话 x^i 映射到一个语义空间。之后采用两种类型的预训练模型，一种是只依赖于语言的 Bert 模型[126]，即：

$$e^i = \text{Bert}(\mathbb{E}(x^i)) \tag{4-2}$$

其中 $\mathbb{E}$ 是将文本中的词汇变成词嵌入的形式。最终的句子表示使用了两种方式：（1）选择了使用特殊符号 [CLS] 对应的输出变量；（2）考虑到动作对于视频的重要性，文章使用了句子中的动词③对应的输出变量的平均向量。

另外一种方式是基于 CLIP 的语言和视频特征 c 共同训练好的特征提取模型

① 由于在收集描述时候，受于资源的限制，受雇的标注者有限，因此只得到有限集合，而理论上，对于一个场景可以有无限种表达语句。

② 假设集合即 hypothesis set 的中译，表示预测集合。但传统的方法将生成的每一条语句并非认为是确定的、正确的，而用假设代替预测一词。

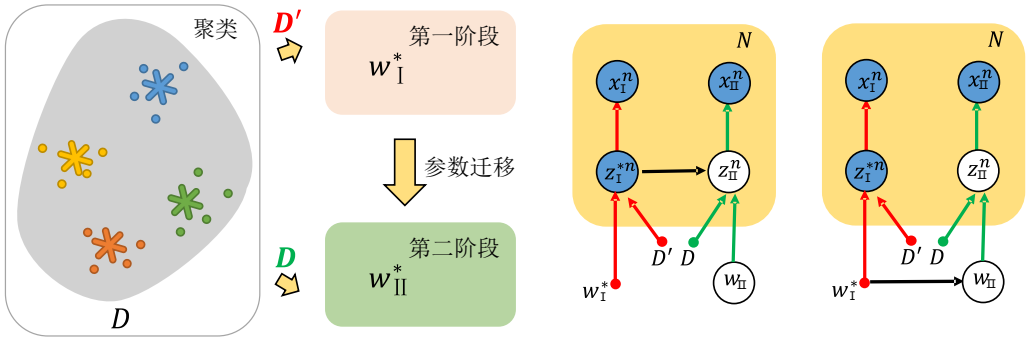
③ 使用 StanfordCoreNLP[127] 中的提取词类的工具检测句子中的动词。

Clip2video[120]，它基于大规模视频和描述配对语料库进行训练，因此可以学习到一些对齐的特征，表示如下：

$$e^i = \text{Clip2video}(\mathbb{E}(x^i), c) \tag{4-3}$$

在得到每句话所对应的表征后，再对其进行聚类。这里采用 K-means 方法将 R 聚为 K 个子集，其中一个示例如图 4-1。其中每个子集期望它们具有相似的语义组，它们都可以表达某类话题（这个使用动词聚类时，可以更加直观地理解，每个语义相似的动词都会被分在一个组内）。由于它们描述同一视频的同一话题，可以称之为视觉同义句（visual paraphrase）[125]。同时，聚类过程中可以找到一个类别中心，该类中心到其它句子的平均距离最短，可以称之为中心句（如图 4-1中的红色句子）。该中心句往往比较简单、具有概括性、没有过多修饰语等，可以用来作为该类的一个代表。之后将每个类别的中心句单独抽取出来，构成一个缩小版的数据集 $\mathcal{D}'$，其中 $\mathcal{D}'$ 可以作为整个数据集一个核心语义的一个代表；而对应的全集 $\mathcal{D}$ 则可以认为在 $\mathcal{D}'$ 的基础上做了些同义转换和丰富润色。

4.3.2　渐进式训练



(a) 模型架构　　(b) 局部依赖　　(c) 全局依赖

图 4-2　渐进式训练模型架构以及两种依赖关系

本文采用渐进式方法训练模型，共分为两个阶段。每一阶段采取的基础模型都是分离动词和上下文空间的 CVAE 模型。第一阶段，模型在由中心句组成的数据集 $\mathcal{D}'$ 中进行训练，这一阶段训练所得的最优模型记作 w^*_{I}，由于它是针对于所有样本的，本文称之为全局（global）或者同质（homogeneous）变量；由模型推断出的隐变量（高斯分布对应的均值（μ）和方差（$\log\sigma^2$））记作 z^*_{I}。第二阶段模型的训练依赖于第一阶段，类似地，将最优的模型参数和隐变量分别记作：w^*_{II} 和 z^*_{II}。因此，整个隐空间可以被构建成 $z = [z_{\mathrm{I}}, z_{\mathrm{II}}]$。其中第二阶段隐变量 z_{II} 的边缘

分布[①]可以通过将联合分布中的 z_{I} 积掉所得，即：

$$p(z_{\mathrm{II}}) = \int p(z)dz_{\mathrm{I}} = \int p(z_{\mathrm{I}}, z_{\mathrm{II}})dz_{\mathrm{I}} \tag{4-4}$$

其中，

$$p(z) = p(z_{\mathrm{II}}|z_{\mathrm{I}}^*, w_{\mathrm{I}}^*, w_{\mathrm{II}}, \mathcal{D}) \cdot p(z_{\mathrm{I}}^*|w_{\mathrm{I}}^*, \mathcal{D}') \tag{4-5}$$

针对公式 (4-5)第一项条件依赖，本文基于 w_{I}^* 和 z_{I} 是一对强相关的变量，而认为它们对于和 z_{II} 发挥的作用是一致的假设，设计了两种依赖关系：

• 局部依赖。如图 4-2(b) 所示。即第二阶段的隐变量依赖于第一阶段的隐变量。同时基于上述假定可以得出如下条件独立：在给定 z_{I}^* 和 w_{II} 的条件下，z_{II} 独立于 w_{I}^*，即：

$$p(z) = p(z_{\mathrm{II}}|z_{\mathrm{I}}^*, w_{\mathrm{II}}, \mathcal{D}) \cdot p(z_{\mathrm{I}}^*|w_{\mathrm{I}}^*, \mathcal{D}') \tag{4-6}$$

• 全局依赖。如图 4-2(c) 所示。即第二阶段的模型权重依赖于第一阶段的模型权重。同样有如下条件独立假设：在给定 w_{I}^* 和 w_{II} 的条件下，z_{II} 独立于 z_{I}^*，即：

$$p(z) = p(z_{\mathrm{II}}|w, \mathcal{D}) \cdot p(z_{\mathrm{I}}^*|w_{\mathrm{I}}^*, \mathcal{D}') \tag{4-7}$$

本文最终采取了全局依赖的方式，也就是模型权重建模为 $p(w) = p(w_{\mathrm{II}}|w_{\mathrm{I}}^*)$。值得注意的是，这种权重上的依赖方式相当于一种参数迁移的方式（见模型架构图 2-2，即常见的“训练-微调”模型[②]模型期待在第一阶段学习到了一个准确的、稀疏的参数空间；第二阶段在第一阶段的前提下可以学习到更加丰富、表达更加多样的参数空间。

4.3.3　模型训练

模型训练采用的两阶段训练的渐进方式，并没有联合训练，两个阶段的损失函数是一样的，如下：

$$\mathcal{L}_{\mathrm{recon}} = -\sum_{v \in \mathcal{V}} \sum_{t} \log p_\theta(x_t|x_{t-1}, z_t, v) \tag{4-8}$$

整体损失则包含两部分：重构损失和 KL 损失，如下：

$$\mathcal{L} = \mathcal{L}_{\mathrm{recon}} + \beta \mathrm{D}_{\mathrm{KL}}(q_\phi || p_\theta) \tag{4-9}$$

① 求 z_{II} 的边缘分布可以采用蒙特卡洛采样方式使用数值进行逼近，即先从分解后的联合分布采样 z_{I}] 和 z_{II}，然后再隐式地扔掉 z_{I}。

② 模型与常见的“训练-微调”模型并非完全一致，由于它一般用于迁移学习中，两个数据集的分布往往差异较大，例如图像分类数据集 ImageNet[] 和语义分割数据集 CityScape[]，但本文两个阶段对应的数据集则存在子集和全集的关系，分布的差异也比较小。同时这种方式有些类似课程学习，不过并没有显式定义简单和复杂样本。

其中 β 是一个（反）退火函数（annealing function）[118]，用来平衡准确性和多样性。之后的所有实验 β 的在第一轮训练的值都设为 0.2，剩下的设置为 $\beta_{\max}$。

4.4　实验

本章节介绍实验部分。第一部分介绍实验细节，包括数据集、评估指标、基准方法以及执行细节；第二部分是实验结果的定量比较结果，包含和最优结果比较和自身的消融实验；第三部分是评测的相关性，包含评测设计和评测结果；最后一部分通过几个实例对结果进行定性分析。

4.4.1　实验细节

4.4.1.1　数据集

实验主要采用了三个规模较大的开放话题的常见数据集。MSVD 数据集[114]包含 1970 段从 Youtube 上截取的开放领域的视频，共有 85500 条描述，每个片段大约 40 个描述。MSRVTT 数据集[30]是另一个规模更大的数据集，它由 1 万段网络视频片段构成，每个片段包含 20 条描述。VATEX[31]是另外一个规模更大的双语视频描述数据集，它共包含 41250 段视频片段以及每段视频附有 10 句中文描述和 10 句英文描述，共 825000 条。从数据规模、视频场景复杂性、语言多样性的角度出发，VATEX 数据集要强于 MSRVTT，而 MSRVTT 则要强于 MSVD。三个数据集的更加详细的介绍可以参考上一章的相关介绍。

4.4.1.2　评估指标

根据不同图像字幕的评估指标，文章从整体、准确性和多样性方面评估模型。

- 整体。文章使用新提出的 hau 衡量描述集合的整体性能。
- 准确性。文章报告 Oracle 条件下的（即最佳挑选出的描述的结果）Bleu-4（**B4**）、Meteor（**M**）、Cider（**C**）、Rouge（ **R**）和 Spice（**S**）指标。
- 多样性。文章将多样性指标分为局部的和全局的：（1）全局多样性考虑为所有视频生成的所有字幕，其中 **Gunis** 计算所有候选字幕中独一无二的句子百分比；**guniv** 则计算每个句子的独特动词的平均数量；（2）局部多样性衡量一个视频对应的字幕集的多样性。**Div1** 和 **Div2** 是不同的 unigrams 和 bigrams 与 *n*-grams 总数的比率；**luniv** 或者 **lunis** 表示集合中独一无二的动词或者句子的数量。MBLEU-n 表示视频的每个字幕和其余字幕之间的平均 BLEU-n 分数。简约起见，文章只报告 **mix-mB** $= 1 - \frac{1}{4}\mathrm{mBLEU_n}$；Self-Cider（**Self_C**）[121]计算成对相似度矩阵的最大特征值的比值。对于所有指标，更高的值代表更好的多样性。

4.4.1.3　基准方法

文章重新实现了多样性图像表述任务下的一些 VAE 模型，来作为文章中用于比较的基准模型。其中，**COS**[86]采用了目标（名词）和上下文分离的输入空间，并以时序的方式进行建模。**Seq-CVAE**[85]采用时序 CVAE，但没有分割隐空间。**ATVAE** 是本文上一章提出的方法，它是对于 COS[86]方法进行的改进，即将图像领域内占据重要位置的“名词”换做“动词”，因此这种方法将隐空间分割为动作隐空间和模板隐空间（详情可参考上一章）。同时文章还包含了传统的集束搜索（**BS**）的文本解码方法，它并没有隐空间假设，取而代之的是在生成句子的词汇构成的联合大空间中贪婪采样。如前文所述，文章首先通过与真值匹配，搜索出（理论上）最佳的描述，求得其对应的准确性。之后，再根据基于共识-重排序的方式报告它们最佳描述对应的准确性。按照 COS[86]论文中的惯例，文章的多样性衡量是在重排序后选择前 5 个最佳的描述中计算。之后还比较了单句视频字幕方法，如 **SAAT**[52]、**GRU-EVE**[128]、**POS-CG**[55]。

4.4.1.4　执行细节

预处理　文章参照论文 RMN[51]，词汇的收集是在去除标点符号和训练集中的稀有词，并将描述转换为小写，最终 MSVD 数据集含有 7,351 个词汇，MSRVTT 则含有 9,732 个词汇，VATEX 含有 10,525 个词汇。每个词对应一个 300 维的嵌入向量。描述的词类标注由 StanfordCoreNLP 提取。动词的词汇表是由在数据集中至少出现了 4 次的动词构成，最终 MSVD，MSRVTT，VATEX 中的动词数量大小分别为 1,137，2,333 和 4,294。特殊符号用来进行掩膜动词。

特征提取　在数据集 ImageNet[122]上预训练的模型 InceptionResNetV2[123]用于提取视频每一帧的空间特征，而在 Kinetics 数据集[119]上预训练的 I3D 模型[124]则用于提取时空特征。这两种类型的特征连接起来，得到的平均池化后的输出作为视觉特征向量 c。对于语言部分，词汇对应的权重从头训练；而对于主题聚类，则采用预训练后的 BERT 模型[126]进行提取。

实现细节　模型使用动量为 0.9 且权重衰减为 0.001 的 SGD 优化器。第一阶段和第二阶段的初始学习率分别为 1.5×10^{-2}，并随着每轮迭代线性下降。隐变量的维数和隐藏层维度大小分别为 128 和 1024。两个数据集的聚类组数 M 均设置为 4。公式(4-9)中的 $\beta_{\max}$ 超参数用于调节生成描述的多样性和准确性，是一个重要的超参数。最终，MSVD 数据集在两个阶段的 $\beta_{\max}$ 分别为 0.2 和 0.5；MSRVTT 数据集则分别为 0.5 和 0.7；VATEX 数据集则分别为 0.9 和 0.9。文章采用度量 hau 作为实

验早期停止的标准，最大的容忍轮数为 20，最大轮数设置为 100。

4.4.2　定量比较

4.4.2.1　性能比较

数据集	方法	整体性	上界准确性				
		hau	B4	C	R	M	S
MSVD	COS	22.5	45.9	105.3	78.9	52.8	-
	Seq-CVAE	24.1	50.7	113.4	81.0	57.8	8.4
	BS	21.5	47.8	108.6	78.9	52.7	-
	ATVAE（本文）	24.4	53.1	113.9	81.6	57.5	8.4
	STR（本文）	**25.3**	**54.8**	**117.8**	**82.4**	**58.9**	**8.9**
MSR-VTT	COS	19.1	41.8	63.6	68.5	41.8	-
	Seq-CVAE	19.6	44.9	64.5	69.7	43.2	**10.8**
	BS	16.1	39.1	59.4	67.2	40.7	8.3
	ATVAE（本文）	19.6	44.0	65.1	69.3	43.2	10.5
	STR（本文）	**20.0**	**45.3**	**65.3**	**70.2**	**43.7**	10.5
VATEX	COS	17.9	34.4	67.4	55.1	28.2	-
	Seq-CVAE	18.1	36.9	69.6	56.7	28.6	-
	BS	17.4	**38.5**	**71.0**	56.9	**28.8**	-
	ATVAE（本文）	**18.2**	37.5	70.3	**56.9**	28.7	-
	STR（本文）	17.8	35.4	66.6	56.2	28.2	-

表 4-1　两个数据集在不同方法上的整体性和准确性结果

与多句描述方法相比较　表 4-1报告了几种多样视频描述方法在 MSVD，MSR-VTT 以及 VATEX 三个数据集上的准确性和整体性能，其中采样数都为 20。与其他基于 VAE 的模型（即 COS，Seq-CVAE 以及 ATVAE）以及传统的 Beam Search 相比，文章提出的方法几乎在所有指标上都取得了最好的成绩，这表明预测的字幕集更加准确。尤其参考整体性的指标，文章提出的模型比最相似的模型 ATVAE 模型高 0.9 分，这表明了模型提出的两阶段学习策略的有效性。同时再参考表 4-2，它展示了从前 20 个采样样本中选择前 5 个最准确的句子的集合中的多样性结果。本文提出的模型 STR 在很多多样性上的指标都比其它方法好。值得注意的是，传统的

数据集	方法	多样性							
		Div1	Div2	lunis	gunis	mix-mB	Self-C	luniv	guniv
MSVD	COS	26.4	34.6	49.1	40.6	11.7	28.9	1.46	**1.02**
	Seq-CVAE	28.0	37.8	55.7	44.0	13.8	35.1	1.52	**1.02**
	BS	30.3	45.0	**98.3**	53.4	20.4	**58.7**	**1.83**	0.95
	ATVAE（本文）	27.7	37.9	57.5	45.0	14.0	36.2	1.49	1.01
	STR（本文）	**35.3**	**50.6**	76.4	**62.9**	**27.8**	53.2	1.61	1.00
MSR-VTT	COS	38.5	58.6	88.1	75.7	35.1	66.6	3.01	1.22
	Seq-CVAE	39.9	61.0	89.3	75.2	36.4	68.2	2.78	1.17
	BS	28.6	41.1	**96.9**	39.2	14.2	-	1.87	1.09
	ATVAE（本文）	**43.6**	**66.7**	93.1	77.4	**43.6**	**73.4**	2.88	1.20
	STR（本文）	43.3	66.6	93.5	**78.2**	43.2	73.3	**3.32**	**1.30**
VATEX	Seq-CVAE	**32.2**	51.0	90.0	88.0	25.3	61.1	5.29	2.55
	BS	26.4	37.6	**99.7**	77.6	12.3	-	3.87	2.24
	ATVAE（本文）	27.7	**58.3**	96.6	**94.9**	**33.5**	**69.3**	**6.11**	**2.62**
	STR（本文）	31.7	50.1	91.4	88.9	24.9	61.1	5.12	2.52

表 4-2　两个数据集在不同方法上的多样性结果

集束搜索的方法由于使用贪心算法和树形的搜索策略，其多样性远远小于基于采样的方法。同时随着搜索宽度（beam width）的增大，其搜索效率也会变得非常低下。

与单句描述方法相比较　表 4-3表明，STR 可以轻松迁移到传统的单句的描述任务。相比以前模型，本文提出的模型大大提高了全局多样性，这表明 STR 可以生成更多独特的句子和更多的独特的动词。值得注意的是本文使用的基础模型没有像其它模型会针对任务过多的设计，例如 SAAT 的自注意力模块等。STR 以较大的优势在多样性上超越了其他模型，同时仍保持相当的准确性（多样性和准确性往往是一个需要平衡的两个指标），这表明了所提出方法的有效性。

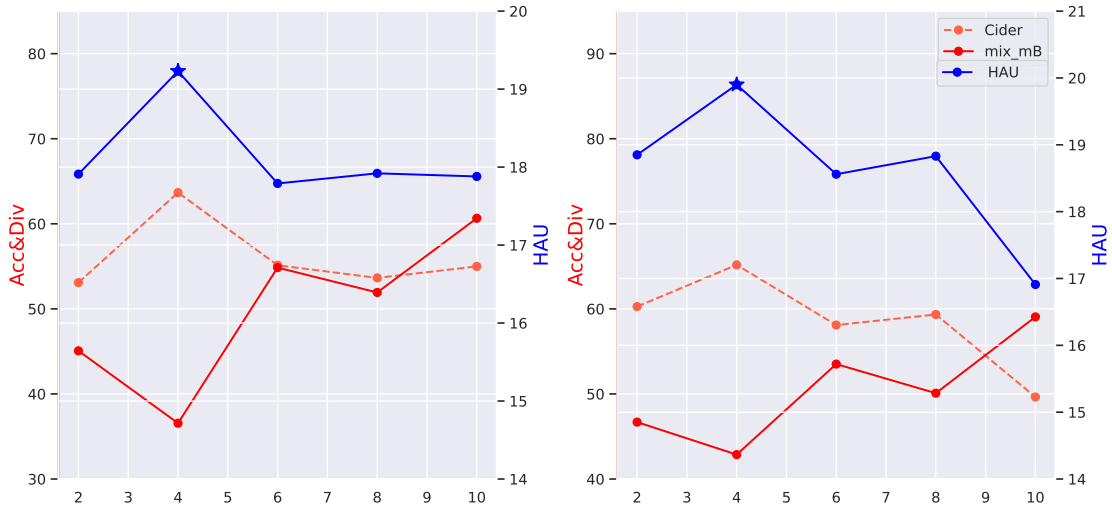
数据集	方法	C	R	M	gunis	guniv
MSVD	COS	87.2	69.3	34.3	40.57	**1.02**
	SAAT	81.0	69.4	33.5	44.0	1.00
	GRU-EVE	78.1	**71.5**	35.0	-	-
	ATVAE（本文）	**90.9**	71.4	**35.8**	45.0	1.01
	STR（本文）	87.9	69.3	34.6	**62.9**	1.00
MSR-VTT	COS	46.6	54.0	25.9	75.7	1.22
	SAAT	**49.1**	**60.9**	**28.2**	31.0	1.04
	POS-CG	43.4	60.1	26.8	-	-
	ATVAE（本文）	46.7	53.7	25.5	77.4	1.20
	STR（本文）	46.4	54.8	25.8	**93.0**	**1.30**
VATEX	COS	45.3	42.8	20.6	99.6	1.2
	VATEX[31]	44.3	46.9	**21.6**	-	-
	BS	45.0	**47.6**	21.3	**99.9**	2.24
	ATVAE（本文）	**48.7**	45.0	21.4	94.9	**2.60**
	STR（本文）	47.7	45.5	21.5	91.4	2.52

表 4-3　与生成单句话的视频描述方法的准确性和多样性比较

4.4.2.2　消融实验

本文设计了一系列的消融实验，还验证一些模块的重要性，主要集中在训练策略和聚类策略上。所有结果汇总在表 4-4 中。

聚类数量选择　图 4-3展示了 MSRVTT 数据集在两个阶段的准确性、多样性和整体性的表现，其中准确性采用 Oracle_Cider 指标表示，多样性使用 mix_mB 代表，整体性则使用 HAU 表示。其中 Cider 和 mix_mB 的坐标轴为红色的左半轴坐标；HAU 为蓝色的右半轴坐标。从 (a) 图可以看出准确性和多样性整体有一个相反的趋势，而整体性而言在 K 为 4 时候可以达到一个较高的值，可能是由于过度小的 K 可能会导致训练集的规模减小，而在 K 过度大又可能夸大了视频中突出的视觉



(a) 第一阶段性能随K的变化　　(b) 第二阶段性能随K的变化

图 4-3　MSRVTT 中随着聚类数量在准确性、多样性以及整体性能的变化

信息，导致结果不准确。值得注意的是 K 值越大，第一阶段训练的规模就越大，该阶段消耗的资源、所用的时间就越大，而在 K 较小时就可以达到一个比较好的值，也对在资源匮乏情况下应用改模型提供了一些启发。从 (b) 图看，第二阶段与第一阶段有类似的变化趋势，而该阶段 hau 指标将作为最终模型选择 K 值的依据，即如图所示选择 K 为 4。

学习策略　表格 4-4的第一部分，本文比较了 MSRVTT 数据集下，分别采用单阶段（也就是 ATVAE）训练和双阶段的第一阶段的结果。文章中提出的方法在准确性和多样性两个方面都比被比较的方法好，证明了文章提出的新模型 STR 两阶段学习机制的有效性。其中反映多个话题的局部指标 luniv 比其它方法高比较多，说明了 STR 可以产生更加多样性的话题。有趣的是，第一阶段在仅给定字幕全集的一个稀疏子集的条件下，仍然可以得到与 STR 相比的性能，这表明，一个稀疏的语义空间足以捕捉比较准确的句子。这种情形可以指导实际资源不足的情形，如小语种的描述标注问题。

聚类策略　为了验证聚类中心（中心句）发挥的作用，本文选择了每个小组中的随机一个句子构成的子集（random）和句子数量最大的一个组的 K 个句子构成的子集（mode）做了对比实验，结果参考表 4-4。结果显式出中心句子集仍可以得到更好的正确性和多样性，暗示一个语义更加稀疏的空间对于结果的重要性。同时文章尝试了不同的方式进行聚类：（1）用于聚类的句子的表示获取方式：SBERT[129] 采用对比学习对句子表示进行了微调，最终使用代表句子整体的特殊符号 CLS 作为

模型	C	M	S	hau	luniv	guniv
STR（本文）	**65.2**	43.4	10.6	**19.9**	3.30	1.30
单阶段	65.1	43.2	10.5	19.6	2.88	1.20
第一阶段	63.7	43.0	10.0	19.2	2.69	1.20
mode	65.2	**44.1**	10.4	19.8	2.90	1.22
random	64.7	43.3	10.3	19.8	3.13	1.34
SBERT	61.3	41.3	**10.9**	19.4	**3.55**	**1.38**
视觉语言特征	61.4	41.4	10.8	19.4	3.22	1.36
分层聚类	62.3	41.9	-	19.0	3.06	1.23
凝聚聚类	61.3	41.1	-	18.8	2.97	1.22

表 4-4　在 MSRVTT 上的与训练策略（上）和聚类策略（下）相关的消融实验

该句子的表示，视觉语言特征采用 CLIP2Video[120]。（2）聚类的方式，包含分层聚类和凝聚聚类两种方式。最终根据 hau 确定本文采用的模型。

4.4.3　定性分析

本节展示了模型预测出的句子结果，如图 4-4所示。除了真值标注，传统模型 BS、与本文提出最相似的单阶段模型 COS 以及本文提出的模型 STR 都对预测结果进行可视化。其中，视频依照序列截出几个关键帧，每个模型预测的输出展示了五句输出作为结果。红色部分高亮出了唯一的动作词汇来表示不同的话题。从图中可以看出，人类标注的真值集合很显著的特征就是本文一再强调的“多样性”，对上图示例而言，尽管视频中的主要行为是“摔跤”，其主谓宾的搭配，句式结构，描述重点①。下图踢足球的例子仍是类似的。而对比模型的输出而言，BS 的结果几乎没有发生大的结构上的变化，而 STR 相比其它两种方法可以捕捉到更加多样化的表达，值得注意的是，它也可以表现出一定的语义偏差，如第二幅图中既有容易检测出的“踢足球”，也有“跑步”这一话题。因此 STR 更适合对于复杂场景下的物体之间的互动关系进行建模。

① 如图中第二句并没有很明显地提到“摔跤”，而是更换了一种表述方式——“当宣布者介绍摔跤手时候，二者开始比赛”，发生了语义偏差。

GT: {'an amatuer wrestling event', 'two wrestlers begin their match as the announcer introduces them', 'guys wrestling each other in a competition', 'two people are fighting', 'wrestling champions are fighting on the ground'}

BS: {'two men are wrestling in a competition', 'two people are wrestling in a competition', 'two people are wrestling', 'two people are wrestling', 'two men are wrestling'}

COS: {'a man is wrestling', 'two people are wrestling', 'two men are wrestling each other in a competition', 'two men are wrestling', 'people are wrestling'}

Ours: {'two men are wrestling', 'two people are wrestling on a competition', 'a man is talking about a wrestling match', 'a wrestling match on the mat', 'two wrestlers are wrestling on a competition'}

GT: {'a soccer player hops over a wall', 'soccer players are running around on a field', 'there are people playing soccer and scores a goal', 'a soccer highlight video', 'compilation of popular soccer clips'}

BS: {'men are playing soccer', 'people are playing soccer', 'people are playing soccer', 'people are playing soccer', 'some people are playing soccer'}

COS: {'men are playing soccer', 'people are playing soccer', 'people are playing soccer games', 'people are playing during a soccer game', 'people are playing soccer games'}

Ours: {'a man is playing a soccer game', 'a man is running in a soccer game', 'a soccer player kicks the ball', 'a soccer match being played', 'the players were playing and playing soccer'}

图 4-4　MSRVTT 中一些带有真值标注和预测模型标注的示例

4.5　讨论：与单阶段模型的联系和区别

上一章提出的单阶段模型与本章所提出来的双阶段方式有很多联系和区别，故在模型架构、实验设置、实验结果等都有相似之处。本节探讨它们之间的主要区别和联系，并且探讨它们背后设计的哲学原理。

二者联系紧密，在设计上是一脉相承的，主要体现在如下方面：

（1）二者分离思想类似。单阶段的模型分离的是**单个**句子的“联合”隐空间，它将句子中的单词视作是基本单位，并且在时序依赖的前提下，进一步区分出动作和上下文；双阶段模型将单个视频对应的整个句子集合作为一个更大意义的“隐空间”，不过，它将整个句子集合中的单独的句子视作集合中的一个元素。整个句子空间分离出的“主题”句，之后再在整体的“表达”集合中去进一步发掘语言的多样性。也就是说它们都是将一个大的输入空间（或者隐空间）分离成多个部分。

（2）二者都属于一种结构化隐变量分离模式。如（1）所述，二者都进行了空间上的分离操作。同时，它们都分离的两部分空间都是后一个以前一个为条件：单阶段模型的动作空间是以上下文空间为条件的；而双边界模型的表达空间是以主题空间为条件。这种条件性的设计除了出于可解释性①的考量，主要可以增加隐空间结构的复杂性，从而增加模型的重构能力[88]。

① 这里的可解释性主要是指符合人类生成句子的过程：比方先留意一个突出的动作，再把它套入到某个“习得”的语言模板中。

（3）从设计结构上看，双阶段在每一个阶段模型的结构与单阶段的设计完全一致。这是由于双阶段的设计事实上是一种“预训练-微调”的模式，有些类似于课程学习[130]，即两个阶段仅仅在输入空间有所不同，模型结构不需要有所区别，所以可以自然地使用单阶段的模型。因而双阶段方法具有很强的可拓展性。值得注意的是，尽管双阶段的训练方式强调挖掘输入句子之间的关系，但只是从整个输入空间来看的，在实际的训练中，仍然是单个句子去计算损失，进而更新模型的。

（4）从实验设计上看，二者由于模型相似，实验中的很多设置都可以相互参考，例如正则化系数 β 对于多样性和准确性的权衡作用在两种方法中都非常敏感。在章节的叙述中也可以看到两组实验的相似性。

不过，二者也有一些主要的区别：

（1）归根到底，单阶段模型的分离空间还是只用到了单个描述的隐空间，而双阶段的方式利用了整个描述集合中的所有句子间的关系，文章论证了后者要更利于恢复原始的“一对多”的分布。

（2）两阶段的方式在第一阶段的输入数据是整个数据集合的子集，因而相比整个单阶段，有些超参数的数值需要调整，例如第一阶段的 β 的取值相比单阶段的选取的更小。实验过程也提示，超参数对于数据量的变化仍有很敏感的影响，在更改了数据后，也仍需要关注超参数的变化。未来如何保证超参数更小的波动仍然是个有挑战性的课题。

总而言之，双阶段的模型是建立在单阶段的基础上的，在有效分离出动作空间基础上又可以充分利用句子之间的关系，从而进一步挖掘多样性的关系。同时，二者在实验设计、超参数选择等等上面都有共通一致的地方，可以相互借鉴和参考。

4.6　本章小结

本章介绍了多样视频描述任务中做出的另一个改进：渐进式训练机制。它是一种两阶段训练方式，第一阶段利用句子集合中的中心句子集；第二阶段在前一阶段基础上进一步训练。它可以有效利用多句标注之间的关联，从而直接针对一对多的目标。本章可以视作是上一章方法的延续，其中的每一阶段训练方式正是上一章提出的方法。本章详细介绍了实验设置以及结果分析，它们可以验证方法的有效性。最后一节介绍了和上一个方法的区别和联系，体现了工作的一脉相承。

第 5 章　集合水平的评估指标设计

本章详细探讨关于视频描述指标的问题，同时介绍本文设计的指标以及与对指标的评估。具体而言，第一节调研了传统的针对描述的准确性指标；第二节介绍了本文设计的指标 hau 和 o2o；第三节则分析了本文提出的指标与人类评估的相关性，以证明其有效性。

5.1　已有指标

现有的对于多样性视频描述生成任务主要从两个方面对于描述进行自动化评估：准确性和多样性。准确性衡量生成的句子的质量，反映预测的句子与“黄金准则”，即标注的句子集合之间的相似性，进而表现出句子的流畅度、可否反映突出视觉对象等；多样性衡量生成的句子与句子之间以及句子内部词汇间的差异性，进而衡量句子表达的丰富性、多样性、表达视觉对象的全面性等等。由于自然语言和视觉场景的复杂性（如异性同义的同义表达不能简单判错），并不能像图像分类等任务简单通过标签匹配等“硬”匹配，往往需要数个较多的自动指标从不同的角度进行衡量。以下分别对各个指标进行说明。值得注意的是，以下指标都用于单句准确性评估，多样性指标在上文实验部分已做过介绍，这里不再赘述。

现有评测准确性的指标大多针对单条预测描述（到一个真值描述集合），对于一个预测集合来说，仍然是每个单条描述计算准确性后再取平均。在这些指标中，有三个指标是完全借鉴自然语言任务中的：BLEU（Bilingual Evaluation Understudy）[131]，ROUGE（Recall Oriented Understudy of Gisting Evaluation）[132]和 METEOR（Metric for Evaluation of Translation with Explicit Ordering）[133]；另外两个指标 CIDER（Consensus-based Image Description Evaluation）[134]和 SPICE（Semantic Propositional Image Caption Evaluation）[135]则是专门针对图像描述而设计的，不过它们也都适用于视频描述。下表 5-1中列出了这几种指标的对比。

5.1.1　语言任务相关

BLEU　BLEU[131]是用于量化机器生成文本的质量的常见指标。BLEU 分数考虑了预测的 unigrams（单个单词）或更高阶的 n-gram（n 个相邻单词的序列）与一组一个或多个候选参考句子之间的重叠程度。高 BLEU 值的描述应该在长度上基本与真值句子匹配，即单词以及它们的顺序要精确匹配。如果完全匹配，BLEU 评估

指标名称	设计缘由	方法归纳
BLEU[131]	机器翻译	基于 n-gram 精度（precision）
ROUGE[132]	文档摘要	基于 n-gram 召回率（recall）
METEOR[133]	机器翻译	基于同义词的 n-gram 匹配
CIDER[134]	图像描述	由 TF-IDF 加权的 n-gram 匹配
SPICE[135]	图像描述	基于场景图的同义词匹配

表 5-1　常见视频描述评估指标对比。

将获得 1 分。不过，需要注意的是，每个视频的真值句子集合中的参考句子数量越多，获得更高 BLEU 分数的机会就越大（命中的可能性越大）。它主要用于在语料库级别评估文本，因此，将其用作单个句子的评估指标可能不公平。BLEU 的计算公式为：

$$\log \text{BLEU} = \min\left(1 - \frac{l_r}{l_c}, 0\right) + \sum_{n=1}^{N} w_n \log p_n \tag{5-1}$$

上式中，$\frac{l_r}{l_c}$ 为真值描述长度与候选描述的比值，w_n 为正权重，p_n 为修正后的 n-gram 精度的几何平均值。注意的是，第二项才是实际的匹配分数，而第一项是对比参考描述更短的描述进行的惩罚。

ROUGE　ROUGE[132]指标于 2004 年提出，用来评估文本摘要。它计算 n-gram 对应于参考句子的生成句子的召回（recall）得分。与基于精确度的 BLEU 不同，Rouge 基于召回值。图像和视频描述评估中使用的版本是 Rouge（L），即它计算的是预测描述和每个参考描述之间最长的公共子序列（longest common sequence）的召回值和精度，背后的直觉是预测句子和参考句子之间较长的公共部分可以更好的反映两个摘要之间的相似性。要匹配的单词无需连续，但应该出现在序列中，其计算公式为：

$$\text{ROUGE} = \frac{\sum_{S \in R_{Sum}} \sum_{g_n \in S} C_m(g_n)}{\sum_{S \in R_{Sum}} \sum_{g_n \in S} C(g_n)} \tag{5-2}$$

其中，n 代表 n-gram 长度，R_{Sum} 代表真值描述集合，$C_m(g_n)$ 代表同时出现在参考集和预测集的 n-gram 的最大数量。

METEOR　METEOR[133]可以缓解 BLEU[131]等指标的一些缺点。注意到 METEOR 引入了**语义**匹配，而不是其它指标所要求的**词形**匹配。METEOR 使用了 WordNet[136]，这是一个英语词汇数据库，可用于各种匹配的级别，包括精确的词

形匹配、词干匹配、同义词匹配和释义匹配等。

METEOR 分数计算预测句子和参考句子的对齐程度。每个句子都被视为一组 unigram 的集合，并通过匹配预测句子和参考句子的 unigram 来完成对齐。在匹配过程中，预测句（或参考句）中的 unigram 要不映射到参考句（或预测句），要不未命中。如果有多个选项可用于两个句子之间的对齐，则首选具有较少交叉数的对齐配置。在完成对齐过程后，计算 METEOR 分数。

5.1.2　视觉任务相关

CIDER　CIDEr[134]是最近引入的用于图像描述任务的评估指标。它评估预测句子和相应图像的参考句子之间的一致性。它先提取词干并将候选句子和参考句子中的所有单词转换为它们的词根形式。CIDEr 指标将每个句子视为一组包含 1 到 4 个单词的 n-gram 集合。为了编码预测句子和参考句子之间的一致性，它测量了两个句子中 n-gram 的共同出现的频率。最后，在所有图像的参考句子中非常常见的 n-gram，例如“the”等，被赋予较低的权重，因为它们可能对图像内容的信息较少，常常只表明句子的结构信息。每个 n-gram 的权重是使用词频-逆文档频率（Term Frequency Inverse Document Frequency，简称 TF-IDF）[137]计算的：TF 对图像参考句子中频繁出现的 n-gram 赋予更高的权重，而 IDF 对整个数据集中常见的 n-gram 赋予较低的权重。最后，CIDEr_n 的分数计算为：

$$\mathrm{CIDEr}_n(c_i, S_i) = \frac{1}{m}\sum_j \frac{g^n(c_i)\cdot g^n(s_{ij})}{\|g^n(c_i)\|\cdot\|g^n(s_{ij})\|} \tag{5-3}$$

其中 $g(c_i)$ 代表预测句 c_i 的词频统计向量。最终的 CIDEr 得分为 n 取 1 到 4 的得分的平均。

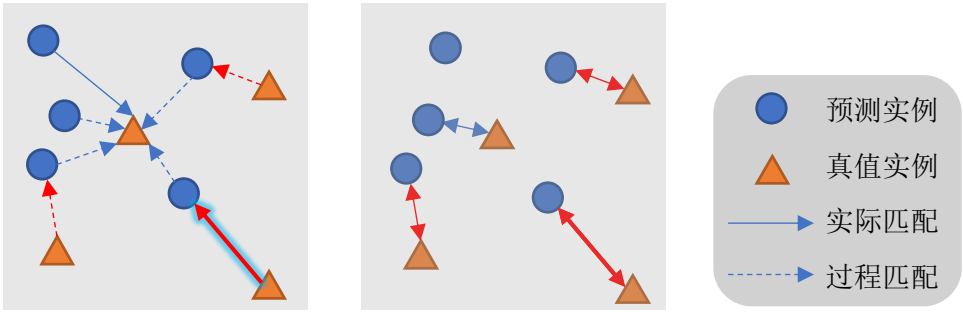
SPICE　SPICE[135]是最新提出的用于图像和视频描述评估指标。SPICE 衡量预测描述中解析的场景图元组与真值之间的相似性。语义场景图通过依赖分析树对对象、属性和关系进行编码。句子 c 的场景图元组 $G(c)$ 由语义标记组成，例如对象类 $O(c)$、关系类型 $R(c)$ 和属性类型 $A(c)$，即：

$$G(c) = \langle O(c), R(c), A(c)\rangle \tag{5-4}$$

SPICE 是基于预测描述和真值元组与之间的 F1 分数计算的。与 METEOR 一样，SPICE 也使用 WordNet 来查找同义词来进行匹配。尽管如此，受限于低效率以及关系提取可能存在的错误，SPICE 指标并没有被广泛应用。同样值得注意的是，尽管 METEOR 和 SPICE 指标都是基于语义匹配，而非词形匹配，它们仍然无法解决同义短语或者同义句的匹配问题，仍然是一种形式优先的匹配。

5.2　指标设计

本文将该任务视作一个假设集合 $\mathcal{H}$ 和参考集合 $\mathcal{R}$ 的匹配任务，根据在匹配过程是否严格“一对一”的配对，提出了两个指标：hau 和 o2o。图 5-1展示了二者的示例图[①]。图 (a) 中，hau 示例所展示的虚线箭头表示每个实例都需匹配的最近的另一个集合中的实例，最终整个集合的距离是由最远距离确定，这里用实线箭头表示，带阴影的箭头表明最终的距离由集合层面最远的距离确定。图 (b) 展示了“一对一”的匹配模式，即一个实例仅可以匹配另一个集合的一个实例，这样可能会造成孤立点的存在。



(a) hau示例　　(b) o2o示例

图 5-1　两个衡量整体性的新指标 hau 和 o2o 的示意图

5.2.1　融合召回率的集合距离

本文采用集合间的豪斯多夫距离（Hausdorff distance）ℓ_H 去衡量 $\mathcal{H}$ 和 $\mathcal{R}$ 之间的距离。假设 x_i 和 y_i 分别出自 $\mathcal{H}$ 和 $\mathcal{R}$，相似性算子 $\mathcal{S}(\cdot,\cdot)$ 衡量两句话之间的匹配程度（例如 Meteor），定义如下：

$$\ell_H = \min(\underbrace{\min_i \max_j \mathcal{S}(x_i, y_j)}_{\text{precision}}, \underbrace{\min_j \max_i \mathcal{S}(x_i, y_j)}_{\text{recall}}) \tag{5-5}$$

正如公式中所指出的，该公式考虑了精度和召回率两个指标，最终最小值算子取了二者的下界。

在实际计算时候，公式 (5-5)外括号内部的 min 算子表示将最差匹配的句子作为集合与集合之间的精度或者召回结果，这样容易导致“灾难性崩溃“的得分，事实上由于采样中噪声的影响，难免会采到一个性能不好的句子，如此严苛的筛选条件最终造成整个集合之间的结果表现很差。因此，本文将严格的 min 算子更改

① 注意这里图中展示的其实是未经改进后的 hau，即公式 (5-5)中的 ℓ_H，这是出于绘图的方便，并不妨碍示意图对于这两者整体计算和各自特征的展示。

为更为温和的 mean 算子，新的指标命名为 hau，定义如下:

$$\ell_{hau} = \min(\underset{i}{\text{mean}} \max_{j} \mathcal{S}(i,j), \underset{j}{\text{mean}} \max_{i} \mathcal{S}(i,j)) \tag{5-6}$$

需要指出的是，正如图 5-1（a）所示，hau 并不限制多个实例点可以匹配另一个集合中的同一个实例点，是一种“多对一”的匹配模型，这种模式的弊端在于，无法惩罚聚集点（它们并不多样但都很准确的情形），不过由于还有另一端，即从真值集合到预测集合的计算，而真值集合的实例大多是分散的（如前文所述人类的标注倾向于较高的多样性），因而这一端提供一种补偿机制。之后所采用的 o2o 的“一对一”式的匹配则没有这个问题。

5.2.2　一对一匹配的集合距离

为了避免聚集点获得较低惩罚这一不合理现象，o2o 采用“一对一”匹配的模式，也就是一个集合的一个实例只可以唯一地匹配另一个集合的一个实例，如图 5-1（b）所示。该指标的提出将两个集合的匹配视作是线性指派问题（linear sum assignment problem）。一般而言该问题需要定义一个距离矩阵 $\mathcal{C}$，该矩阵横向索引 i 表示工人编号，纵向索引 j 表示任务编号，最终需要指派哪名工人完成哪项任务。解决该类问题常常看作求解一个基于下述目标函数的优化问题：

$$\max \sum_{i} \sum_{j} \mathcal{S}_{i,j} \mathcal{X}_{i,j} \tag{5-7}$$

其中，$\mathcal{S}$ 仍为相似性函数，而 $\mathcal{X}$ 则一个布尔型的指示矩阵，用来表明工人编号 i 到任务编号 j 的指派情况：

$$\mathcal{X}_{i,j} = \begin{cases} 0 & i \text{ 未指派给 } j \\ 1 & i \text{ 指派给 } j \end{cases} \tag{5-8}$$

本文采用改进的 Jonker-Volgenant 算法[138]去求解这一优化问题。具体可以参考 python 的 API 接口①，该算法的介绍超出了本文叙述的范围，在此略过。

需要注意的是，考虑到计算复杂性，本文在之前的评估中主要使用 hau 作为综合性评分标准，o2o 仅作为本章的一个参考指标。不过，从后文的评测结果来看，这两个指标仍然具有较强的相关性。

① 参考：https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.linear_sum_assignment.html

5.3　评测相关性

为了验证本文提出的用来刻画整体性能（融合了准确性和多样性）的指标 hau 和 o2o 的有效性，文章收集了人工对于模型的预测集整体性能进行打分，之后再与 hau，o2o 以及其它常用的指标计算相关性。本节第一部分介绍了评测收集的过程，第二部分是对于结果的分析。

5.3.1　问卷评测设计

评测以问卷的形式向受试者发放。问卷以表格的形式，对于每个视频样本，首先提供它对应的真值描述集合，之后表格中依次列出 5 个模型（表 5-2列出了不同的模型的预测结果在准确性和多样性两个维度上的大致特点）对于同一视频样本的预测描述集合，为了保证客观性，这里没有列出本文所采用的模型，而是选择了代表不同准确性和多样性的其它模型。其中的 CR 方法，并不是一个学习出来的模型，而是通过共识-重排序的方式选出与该视频最邻近的视频对应的描述的集合。之后，为了保证可比性，预测的描述与真值描述保持一致。同时模型的顺序对于每个视频都是随机打乱的，这样可以防止受试者受到干扰。

模型	准确性	多样性
集束搜索	✓	✗
多样性约束的集束搜索	✓	✓
COS (verb)	✓	✓
SCVAE	✓	✓
CR	✗	✓

表 5-2　问卷采用的模型的预测描述集合的特点

每个视频样本对应的 5 个模型都需要受试者打分，共 5 个等级，即 1 到 5 分，5 分表示整体性最好，1 分则表示最差。在问卷的注意事项中指明：该整体性主要是参考“集合”方面的相似性来做出判断，可以参考两个方面：准确性在于将真值中常见的、突出的一些视觉概念都刻画出来，并且语句流畅，无明显语法错误；多样性则是指生成语言的语法、词汇要尽可能多样（仍然需要参考真值集合），可以看出，以上也只是一些参考准则，实际主要是一个主观上的评测，其终极目的还是类似图灵测试，去评测机器是否可以生成足够逼真的语句从而人机无法分辨。图 5-2给出了问卷的示例展示，其中，阴影部分是每位受试者都会看到的示例，白

色部分为受试需要预测的描述示意。实际问卷由 EXCEL 表格形式发放。受试需要参考示例部分的打分情况，对之后的预测集合进行打分。示例对应了不同类型的预测集合以及对应的分值，它们包含不同准确性和多样性高低的组合，以便更好地指引受试做合理的打分。在受试的实际打分过程中，不要求受试给出具体的理由。

真值集合	预测集合_0	预测集合_1
two men are fishing one in green top and one in red top	a man putting a gummy worm as bate on his fishing hook	there is a man talking about something
a man in a red and black shirt sunglasses and a backward ba	a man is going towards the river with a material for fishing	a man in a black shirt is talking about a
two persons are catching fishes on a river	a man is walking and talking about fishing how delicious the s	a man in a white shirt is talking about a
a person with a red shirt on is fishing and catches a fish	there is a man on a golf course talking about the fishing lure	a man in a blue shirt is talking about a
a man in a red and black shirt and a black hat catches a fish	a person is carrying a pole with a fishing line while walking thr	a man in a white shirt is standing next to the camera
wearing a black cap backwards and dark sunglasses stands	a man goes fishing in the watertrap at a golf course	a man in a white shirt is standing next to a camera
a smiling man wears his cap backwards and holds a glistenir	guy holding a fish net and walking	a man in a white shirt is standing in front of water
a man catching the fish on the water and posing to the came	a man preparing to fish for bass in a golf course pond	a man in a white shirt is standing in front of a camer
a fisherman shows a fish that he caught as his friend takes a	a guy walking on a golf course with a fishing pole while talkin	a man in a white shirt is standing in front of a mount
guy in red tishirt taking photos of the fish he caught in a lake	man is putting some bait on his fishing strike on the grass	a man in a white shirt is standing in front of a buildin
two men are fishing when one catches a small fish they take	person is putting some baits on the fishing strike and preparir	a man in a white shirt is standing in front of a water
two men transform a one pound fish into a world record catc	a man is walking with a gummy worm on a fishing pole at a c	a man in a white shirt is standing in front of a body
famous youtuber named dude perfect catching a fish and be	a water scene grass beside and holding in hand long stick sp	a man in a white shirt is standing in front of a
man in red shirt is pulling and catching a fish	a man holding a fishing pole with a lure on the end talks abo	a man in a white shirt is standing in front of a white b
man is catching some fish with his friend and taking photo	a man prepares to fish on a golf course lake with a gummy w	a man in a white shirt is standing in front of a large v
two men standing by a a river catching fish	man walking around a golf course with a fishing rod	a man in a white shirt is standing in front of a large b
a couple of men out fishing in a lake and telling jokes	first person footage of a man carrying a fishing pole across a	a man in a white shirt is standing in front of a white b
a person is having a very big size fish in his hand	a person showing the fishing tricks in the camera	a man in a white shirt is standing in front of a large b
a man in red t-shirt gathering fish with needle	the man uses a gummy worm as bait on the fishing line	a man in a white shirt is standing next to a man is st
a guy with a red shirt and hat is fishing outside	a person is carrying a gummy worm on a golf course	a man in a white shirt is standing in front of a large
整体评分：预测集合相对于真值集合的相似程度(包含描述的准	4（准确性较好，多样性好）	1（多样性差，准确性差）

	预测集合_2	预测集合_3
people wrestling at a match and one of a referee	man in black shirt and black shirt is doing a tutorial on a tr	man in red shirt is fishing and catching a fish
a couple of people wrestling on a mat with a referee	a man in a jacket and a black shirt and a cap is demonstr	man in red shirt is fishing and gathering a fish
men are wrestling with a referee and then a referee is laying c	one persons on a show with a long shot on a track	A man in red shirt is fishing and catching a fish
people are wrestling to a mat and one of a referee	two men are in a gym in a large pool	A man in red shirt is fishing and catching a fish
two people competing in a gym and red motion	a man with a cap on a table and a man in a blue shirt with	A man in red shirt is fishing and catching a fish
men are wrestling at the ground in a gym	the men are in a large tank and the camera on the side of	A man in red shirt is fishing and catching a fish
two people are wrestling against each other in a gym	a man is standing on a table and a man is holding a piece	man in red shirt is fishing and catching a fish
men are wrestling at an event	there is a man is talking about a bike	man in red shirt is fishing and catching a fish
men wrestle on a mat and red motion music	a man in a jacket and shirt and a cap is demonstrating ho	man in red shirt is fishing and catching a fish
wrestling match and a guy in yellow shorts and red and red sh	man in black shirt taking some tricks on a table	man in red shirt is fishing then catches a fish
wrestlers fighting match and down a ground and a referee is	a bald man is standing on a table and a man is holding a	man in red shirt is fishing and catching a fish
people are wrestling match	two men are on a show with a large camera on a track	man in red shirt is fishing and catching a fish
two people wrestle on the ground and a referee goes on the	a bald man in a black and black shirt is doing a video gam	man in red shirt is fishing and catching a fish
wrestlers are fighting and music with a referee	man is taking a selfie for a race	man in red shirt is fishing
wrestlers fighting match with each other	a man in a jacket is demonstrating different exercises with	man in red shirt is fishing and catching a fish
people wrestling while fighting at a wrestling match	the man is on the camera at the table and the camera	man in red shirt is catching a fish and fishing
a person is sitting on a mat and another man is getting ready	a man is in a large tank with a large body of water	man in red shirt is fishing and catching a fish
men wrestling and competing on a mat	a man in a jacket is demonstrating different exercises of a	man in red shirt is fishing and catching a fish
wrestlers are fighting with each other	one persons are doing a tutorial on a race	man in red shirt is fishing and catching a fish
two men are wrestling over a match in a gym	a guy wearing a cap and shirt and jeans on a table	man in red shirt is fishing and catching a fish
	1（多样性好，准确性差）	2（准确性好，多样性差）

图 5-2　受试评测问卷展示

问卷随机选择了 100 段 MSRVTT 和 100 段 VATEX 中的视频片段，其预测集合和真值集合的描述数量都是 20。其中，每个片段对应的模型至少保证被 3 个人独立地打分。为了保证答卷质量，每个受试需打出 10 个片段的 5 个模型，共 50 个分值。为了保证打分过程尽可能独立进行，多个模型之间的顺序是随机打乱的，视频之间随机采样并无衔接关系①，采样者之间不被允许互相交流。同时，在打分时受试者被鼓励横向参考其它模型的预测结果，相对着打分，如评分 5 实则表示相对比其它模型而言，该预测集合整体效果最好，以避免潜在的集中打分情况。

① 这是由于在 MSRVTT 等数据集上，都存在多个视频片段对应原始的一段长视频。

5.3.2　问卷评测结果分析

最终，问卷招募到了 60 位受试者，他们都是受过良好英文教育的大学生。本文对他们对不同模型的打分结果做了如下处理：首先对于同一个视频点的两个受试相应模型下的得分取平均，最终将每个视频的每个模型的结果拉成一列，即文章是每个模型的结果都作为一个独立的评分点，再将不同受试的不同视频对应的结果连接在一起，最终便可以得到一个 500（100 乘以 5）维的打分向量。同时这 500 组预测集合都可以得到一组机器自动指标（如上述表示准确性的 C，M 等；多样性的 luniv 等；整体性的 hau 等）。值得注意的是，本文同样计算公式 (5-5)中的得分，记作：hau_min，来验证本文对于它的 min 算子的改进是有效的。得到相同大小的人工打分和指标打分后，本文采用了三类统计上常用到的相关性进行分析，即：Pearson's r，Kendall's τ 和 Spearman's ρ。它们的值越大，说明相关性越大。

首先，图 5-4展示了不同的受试针对不同数据集和不同模型之间的打分相关性的分布。如上所述，每 3 位受试都会对于同 10 段视频进行打分，因此每个数据集会有 10 组不同的打分，每组的每个模型（共 5 个）都有三个分数，这三个分数两两组合，并可以得到三个相关性。为了显示出这三个相关性的平均情况和波动情况，文章采用了盒图进行可视化。可以看出对于特定数据集，特定模型而言，其相关性变化都比较大，有些相关性降到了零下，这显示不同人之间的主观性差异仍然比较大。对于 MSRVTT 而言，SCVAE 模型上的波动会更小些，而 VATEX 上的 BS 波动会相对小些。值得注意的是，上述情况正是反映了人类标注的不确定性，更加说明了评估这个问题的复杂性。为了尽可能减少偏差，本文因而采用三个受试打分的平均值作为最终结果。

相关性类型	MSRVTT				VATEX			
	hau	hau_min	o2o	oracle	hau	hau_min	o2o	oracle
Overall Pearson's r	0.406	0.363	**0.417**	0.270	**0.505**	0.419	0.487	0.225
Overall Kendall's τ	0.306	0.236	**0.312**	0.224	**0.351**	0.303	0.34	0.156
Overall Spearman's ρ	0.424	0.328	**0.431**	0.313	**0.477**	0.419	0.465	0.219

表 5-3　人类评估和 hau 指标以及其两种变种的三类相关性

表 5-3展示了在 MSRVTT 和 VATEX 两个数据集上，中间相似性计算都是基于 Meteor 的三类指标。本文提出的指标 hau 和 o2o 取得了最强的相关性，尤其比现在很多方法都用到的 oracle 的标准而言。不过，两个数据集上的表现没有很一致：

在 MSRVTT 上 o2o 的相关性要略高于 hau，而在 VATEX 上该趋势恰好相反。由于 o2o 涉及到的算法更加复杂，本文主要使用 hau 作为主要的衡量整体性的指标。同时，相比起 hau 的变种——hau_min 而言被本文提出的“软”指标以较大幅度超过（尤其在 MSRVTT 数据集上），这证明了改进的有效性。

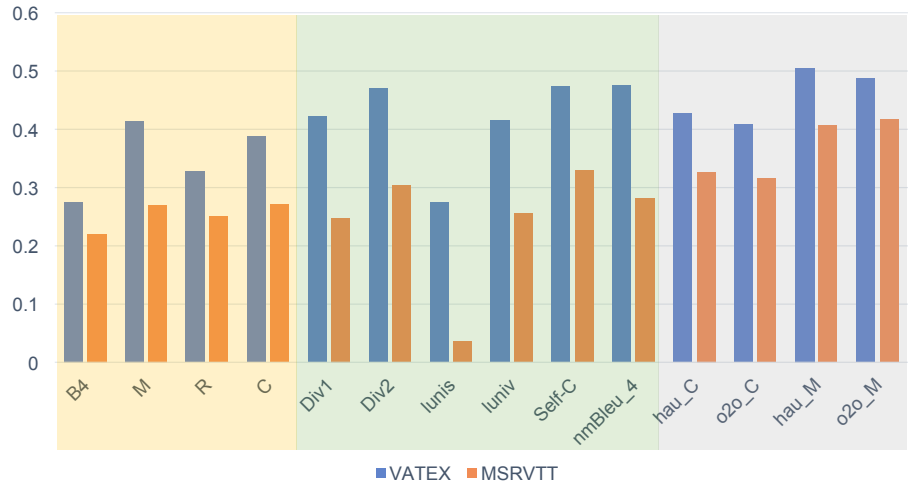


图 5-3　人类评估和其它各个指标的 Pearson’s 相关性

图 5-3则详细地输出了两个数据集的各个自动指标与人类打分相关性大小，其中，不同底色的阴影表示不同类型的指标：黄色表示准确性指标，绿色表示多样性指标，灰色表示本文提出的一系列指标。由于三类相关性相互之间的趋势变化几乎是一致的，这里仅输出 Pearson’s 相关性。可以观察到，涉及到准确性的指标往往和人类评估具有较高的相关性，尤其以 Meteor 和 Cider 为主。多样性的指标除了 lunis 这个指标，其余的多个指标也都有较高的相关性。对比而言，本文提出的指标 hau 和 o2o 则具有最强的相关性。对比两个数据集而言，在不同维度上的评估，它们的趋势都类似，然而，在衡量多样性的 lunis 上，MSRVTT 具有更小的相关性，这可能由于各个模型在 VATEX 数据集生成的句子普遍都是互异的，因此该指标对于区分不同的模型帮助较小，而 MSRVTT 数据集由于体量和多样性仍弱于 VATEX，不同模型对于生成独特句子的程度较为不同，因而该指标仍具有较好的可区分性，因而和人类的相关性也相应更高。

尽管如此，仍然可以看出这两个指标和人类的“黄金准则”标注相比，仍并非具有非常大的相关性，很可能的原因是 hau 和 o2o 出于设计的简单，其中间所利用的距离度量（核函数）仍基于句法表面相似性，例如 Cider，Meteor 等；并没有深层次地利用复杂的与词汇、语义相关的信息，比如：同义转换等。这样在匹配的过程中仍旧是一个“硬”匹配——只有表面**形式**上完全一致才会获得最好的得分，而形式不一致，语义一致的句子却不能得到应有的奖励。

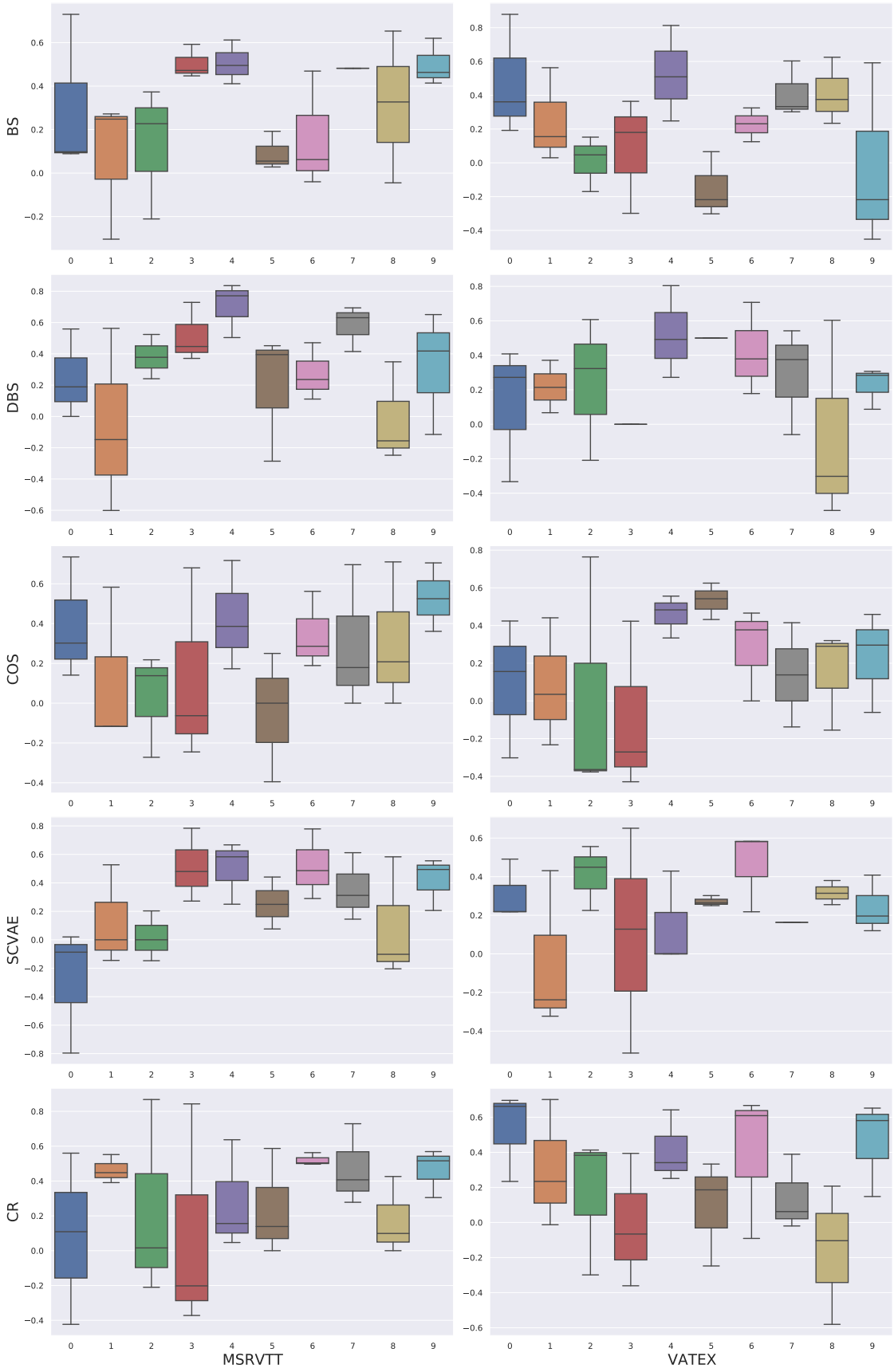


图 5-4　MSRVTT 和 VATEX 上，针对不同模型三位受试者之间两两评估相关性分布图

结 论

本论文研究如何对于一段短视频，得到一组描述准确且多样的句子，即多样视频描述任务，正如标注者所标注的是一个描述集合一样。本文探讨了人类（生成）标注的多样性的可能来源：视频场景中复杂的互动、可能出现的多个视觉主题、自然语言的歧义多变等，并将这些因素编码到隐变量生成模型——VAE 中，从而从生成式的角度得到像人类一样的描述。具体而言，本次研究的创新点和贡献如下：

（1）本文在 VAE 基础上提出了一个动作和上下文（模板）的分离隐空间，即 ATVAE。通过分离出动作空间，模型更加专注于场景之间的互动关系，同时模板部分可以模拟多样的语言表达。同时额外增加了一个对比学习机制，可以进一步拉开句子间的可区分性。实验结果表明这种方法相比传统的基于句子空间贪婪采样的 Beam Search 方法、以及缺乏分离空间的 Seq-CVAE 等方法在准确性可比的情形下，有效地提升了多样性。

（2）本文的第二部分在 ATVAE 基础上，进一步挖掘了同一组描述之间的关系，通过将输入的全体句子构成的全集分割为不同子集，解耦出主题和表达两个输入空间。之后文章提出了一种类似“训练-微调”的双阶段训练机制 STR，有效地捕捉不同的视觉信息（主题）和丰富多样的语言表达。实验结果表明 STR 可以在保持准确性的前提下，进一步提升生成句子的多样性。

（3）针对现有指标单独衡量准确性和多样性，同时准确性仅仅考虑了精度一侧，因而无法从集合层面衡量与真值的匹配问题，本文在第三部分提出了两个可以融合这两方面的指标：hau 和 o2o。它们将整体的准确性对应于集合之间的距离问题，不仅仅考虑到从预测集合到真值集合的精度一侧，也充分考虑到另一方向的召回率一侧，因而可以全面反映生成的描述集合的性能。与人类对于“整体性能”的得分相比，本文提出的指标相比其它指标具有更强的相关性。

未来本研究方向的展望和设想：

（1）从上述定性分析中仍然可以看出模型生成出的描述集合仍然无法和人类匹敌，不同于人类可以生成更加的丰富、自然和多样（多峰分布）的描述，模型还无法全面捕捉到（重构出）这种分布。之后可以进一步采用更先进的模型（如 transformer），更大的数据集（如 WebText），更新颖的学习方式（如 prompt learning）等，进一步提高模型的重构能力。

（2）尽管文章中赋予隐变量一定语义（如动作隐空间），但此种语义并不能真

正在实验中体现出来，只是增加了隐空间结构的复杂性。之后的工作可以进一步研究隐变量的可解释性，以及反映在低维流形的不同维度下更细粒度的可解耦性。

（3）虽然文章提出的双阶段建模的方式在输入空间中利用了多个句子之间的关系，但在实际训练时仍将它视作单个句子，只是试图让模型“隐式”地学习到它们之间的关系。之后的研究可以直接利用它们之间的关系，显式学习到这种一对多的分布。

（4）双阶段训练中先在第一阶段采用一个子集训练，之后再在更大的全集上训练，这符合当前流行的课程学习（curriculum learning）的范式。之后的工作可以结合课程学习，着重从不同的角度如训练的难易程度去划分新的训练子集。

（5）针对短视频的描述大多只是客观命题式的描述，但受到开放视频场景和自然语言的丰富性仍有一部分描述采取了更加人格化的修饰，如幽默、反语、隐喻、夸张、指代等。一个真正反映人标注特征的模型仍然需要考虑这些（潜在的）生成因素。未来研究可以针对风格相关的细粒度方面进行建模。

参考文献

[1] BUSCHMAN T J, MILLER E K. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices[J]. science, 2007, 315(5820): 1860-1862.

[2] KOLLER D, HEINZE N, NAGEL H H. Algorithmic characterization of vehicle trajectories from image sequences by motion verbs[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 1991: 90-91.

[3] BRAND M. The "inverse hollywood proble": From video to scripts and storyboards via causal analysis[C]//Fourteenth AAAI Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence. AAAI Press, 1997: 132-137.

[4] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]// Computer Vision and Pattern Recognition. IEEE, 2001: I-I.

[5] TORRALBA A, MURPHY K P, FREEMAN W T, et al. Context-based vision system for place and object recognition[C]//Proceedings Ninth IEEE International Conference on Computer Vision: volume 2. IEEE, 2003: 273-273.

[6] LOWE D G. Object recognition from local scale-invariant features[C]//IEEE International Conference on Computer Vision: volume 2. IEEE, 1999: 1150-1157.

[7] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2009, 32(9): 1627-1645.

[8] FELZENSZWALB P, MCALLESTER D, RAMANAN D. A discriminatively trained, multiscale, deformable part model[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-8.

[9] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D. Cascade object detection with deformable part models[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 2241-2248.

[10] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2005: 886-893.

[11] CHAUDHRY R, RAVICHANDRAN A, HAGER G, et al. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions [C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 1932-1939.

[12] HONGENG S, BRÉMOND F, NEVATIA R. Bayesian framework for video surveillance application[C]//International Conference on Pattern Recognition. IEEE, 2000: 164-170.

[13] GONG S, XIANG T. Recognition of group activities using dynamic probabilistic networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2003: 742-749.

[14] BOBICK A F, WILSON A D. A state-based approach to the representation and recognition of gesture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(12): 1325-1337.

[15] ZHU S C, MUMFORD D. A stochastic grammar of images[M]. Now Publishers Inc, 2007.

[16] MOORE D, ESSA I. Recognizing multitasked activities from video using stochastic context-free grammar[C]//AAAI Conference on Artificial Intelligence. 2002: 770-776.

[17] POLLARD C, SAG I A. Head-driven phrase structure grammar[M]. University of Chicago Press, 1994.

[18] NISHIDA F, TAKAMATSU S. Japanese-English translation through internal expressions[C]// Ninth International Conference on Computational Linguistics. 1982.

[19] NISHIDA F, TAKAMATSU S, TANI T, et al. Feedback of correcting information in postediting to a machine translation system[C]//International Conference on Computational Linguistics. 1988.

[20] HAKEEM A, SHEIKH Y, SHAH M. CASE^ E: a hierarchical event representation for the analysis of videos[C]//AAAI Conference on Artificial Intelligence. 2004: 263-268.

[21] KHAN M U G, ZHANG L, GOTOH Y. Human focused video description[C]//IEEE International Conference on Computer Vision Workshops. IEEE, 2011: 1480-1487.

[22] LEE M W, HAKEEM A, HAERING N, et al. Save: A framework for semantic annotation of visual events[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2008: 1-8.

[23] NEVATIA R, HOBBS J, BOLLES B. An ontology for video event representation[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshop. IEEE, 2004: 119-119.

[24] GUADARRAMA S, KRISHNAMOORTHY N, MALKARNENKAR G, et al. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition[C]//IEEE International Conference on Computer Vision. 2013: 2712-2719.

[25] THOMASON J, VENUGOPALAN S, GUADARRAMA S, et al. Integrating language and vision to generate natural language descriptions of videos in the wild[R]. University of Texas at Austin Austin United States, 2014.

[26] CHEN D, DOLAN W B. Collecting highly parallel data for paraphrase evaluation[C]//The 49th Annual Meeting of the Association for Computational Linguistics. 2011: 190-200.

[27] ROHRBACH A, ROHRBACH M, QIU W, et al. Coherent multi-sentence video description with variable level of detail[C]//German Conference on Pattern Recognition. Springer, 2014: 184-195.

[28] ROHRBACH A, ROHRBACH M, TANDON N, et al. A dataset for movie description[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3202-3212.

[29] TORABI A, PAL C, LAROCHELLE H, et al. Using descriptive video services to create a large data source for video annotation research[A]. 2015.

[30] XU J, MEI T, YAO T, et al. Msr-vtt: A large video description dataset for bridging video and language[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5288-5296.

[31] WANG X, WU J, CHEN J, et al. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research[C]//IEEE International Conference on Computer Vision. 2019: 4581-4591.

[32] ROHRBACH M, QIU W, TITOV I, et al. Translating video content to natural language descriptions[C]//IEEE International Conference on Computer Vision. 2013: 433-440.

[33] KOEHN P, HOANG H, BIRCH A, et al. Moses: Open source toolkit for statistical machine translation[C]//The 45th Annual Meeting of the Association for Computational Linguistics of the demo and poster sessions. 2007: 177-180.

[34] KOJIMA A, TAMURA T, FUKUNAGA K. Natural language description of human activities from video images based on concept hierarchy of actions[J]. IJCV, 2002, 50(2): 171-184.

[35] DAS P, XU C, DOELL R F, et al. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2634-2641.

[36] KRISHNAMOORTHY N, MALKARNENKAR G, MOONEY R, et al. Generating natural-language video descriptions using text-mined knowledge[C]//AAAI Conference on Artificial Intelligence. 2013.

[37] XU R, XIONG C, CHEN W, et al. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework[C]//AAAI Conference on Artificial Intelligence: volume 29. 2015.

[38] YU H, SISKIND J M. Learning to describe video with weak supervision by exploiting negative sentential information[C]//Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[39] CORSO J. Gbs: Guidance by semantics-using high-level visual inference to improve vision-based mobile robot localization[R]. STATE UNIV OF NEW YORK AT BUFFALO AMHERST, 2015.

[40] SUN C, NEVATIA R. Semantic aware video transcription using random forest classifiers[C]// European Conference on Computer Vision. Springer, 2014: 772-786.

[41] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Conference on Neural Information Processing Systems, 2012, 25.

[42] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[A]. 2014.

[43] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[44] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[45] CHO K, VAN MERRIËNBOER B, BAHDANAU D, et al. On the properties of neural machine translation: Encoder-decoder approaches[A]. 2014.

[46] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [J]. Conference on Neural Information Processing Systems, 2014, 27.

[47] GRAVES A, JAITLY N. Towards end-to-end speech recognition with recurrent neural networks [C]//International Conference on Machine Learning. PMLR, 2014: 1764-1772.

[48] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2015: 2625-2634.

[49] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3156-3164.

[50] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Conference on Neural Information Processing Systems, 2017, 30.

[51] TAN G, LIU D, WANG M, et al. Learning to discretely compose reasoning module networks for video captioning[A]. 2020.

[52] ZHENG Q, WANG C, TAO D. Syntax-aware action targeting for video captioning[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2020: 13096-13105.

[53] PAN B, CAI H, HUANG D A, et al. Spatio-temporal graph for video captioning with knowledge distillation[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2020: 10870-10879.

[54] PEREZ-MARTIN J, BUSTOS B, PÉREZ J. Improving video captioning with temporal composition of a visual-syntactic embedding[C]//Workshop on Applications of Computer Vision. 2021: 3039-3049.

[55] DESHPANDE A, ANEJA J, WANG L, et al. Fast, Diverse and Accurate Image Captioning Guided by Part-Of-Speech[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 10695-10704.

[56] KRISHNA R, HATA K, REN F, et al. Dense-captioning events in videos[C]//IEEE International Conference on Computer Vision. 2017: 706-715.

[57] JOHNSON J, KARPATHY A, FEI-FEI L. Densecap: Fully convolutional localization networks for dense captioning[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4565-4574.

[58] WANG J, JIANG W, MA L, et al. Bidirectional attentive fusion with context gating for dense video captioning[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7190-7198.

[59] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]//European Conference on Computer Vision. Springer, 2016: 21-37.

[60] WANG J, JIANG W, MA L, et al. Bidirectional attentive fusion with context gating for dense video captioning[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7190-7198.

[61] YANG D, YUAN C. Hierarchical context encoding for events captioning in videos[C]//IEEE International Conference on Image Processing. IEEE, 2018: 1288-1292.

[62] WANG T, ZHENG H, YU M, et al. Event-centric hierarchical representation for dense video captioning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(5): 1890-1900.

[63] IASHIN V, RAHTU E. Multi-modal dense video captioning[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2020: 958-959.

[64] IASHIN V, RAHTU E. A better use of audio-visual cues: Dense video captioning with bi-modal transformer[A]. 2020.

[65] LI Y, YAO T, PAN Y, et al. Jointly localizing and describing events for dense video captioning [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7492-7500.

[66] ZHOU L, ZHOU Y, CORSO J J, et al. End-to-end dense video captioning with masked transformer[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8739-8748.

[67] WANG T, ZHANG R, LU Z, et al. End-to-End Dense Video Captioning with Parallel Decoding [C]//IEEE International Conference on Computer Vision. 2021: 6847-6857.

[68] KRAUSE J, JOHNSON J, KRISHNA R, et al. A hierarchical approach for generating descriptive image paragraphs[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2017: 317-325.

[69] LUO Y, HUANG Z, ZHANG Z, et al. Curiosity-driven reinforcement learning for diverse visual paragraph generation[C]//ACM International Conference on Multimedia. 2019: 2341-2350.

[70] MELAS-KYRIAZI L, RUSH A M, HAN G. Training for diversity in image paragraph captioning[C]//Conference on Empirical Methods in Natural Language Processing. 2018: 757-761.

[71] XIONG Y, DAI B, LIN D. Move forward and tell: A progressive generator of video descriptions [C]//European Conference on Computer Vision. 2018: 468-483.

[72] SONG Y, CHEN S, JIN Q. Towards Diverse Paragraph Captioning for Untrimmed Videos[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2021: 11245-11254.

[73] QIAN J, DONG L, SHEN Y, et al. Controllable Natural Language Generation with Contrastive Prefixes[J]. CoRR, 2022, abs/2202.13257.

[74] CHEN L, JIANG Z, XIAO J, et al. Human-like Controllable Image Captioning with Verb-specific Semantic Roles[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2021: 16846-16856.

[75] KIM D J, CHOI J, OH T H, et al. Dense relational captioning: Triple-stream networks for relationship-based captioning[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6271-6280.

[76] CORNIA M, BARALDI L, CUCCHIARA R. Show, control and tell: A framework for generating controllable and grounded captions[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2019: 8307-8316.

[77] LINDH A, ROSS R J, KELLEHER J D. Language-driven region pointer advancement for controllable image captioning[A]. 2020.

[78] CHEN S, JIN Q, WANG P, et al. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2020: 9962-9971.

[79] ZHONG Y, WANG L, CHEN J, et al. Comprehensive Image Captioning via Scene Graph Decomposition[C]//Lecture Notes in Computer Science: volume 12359 European Conference on Computer Vision. Springer, 2020: 211-229.

[80] PONT-TUSET J, UIJLINGS J R R, CHANGPINYO S, et al. Connecting Vision and Language with Localized Narratives[C]//Lecture Notes in Computer Science: volume 12350 European Conference on Computer Vision. Springer, 2020: 647-664.

[81] DENG C, DING N, TAN M, et al. Length-Controllable Image Captioning[C]//Lecture Notes in Computer Science: volume 12358 European Conference on Computer Vision. Springer, 2020: 712-729.

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

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

[84] LI D, HUANG Q, HE X, et al. Generating diverse and accurate visual captions by comparative adversarial learning[A]. 2018.

[85] ANEJA J, AGRAWAL H, BATRA D, et al. Sequential Latent Spaces for Modeling the Intention During Diverse Image Captioning[C]//IEEE International Conference on Computer Vision. IEEE, 2019: 4260-4269.

[86] MAHAJAN S, ROTH S. Diverse Image Captioning with Context-Object Split Latent Spaces [C]//Conference on Neural Information Processing Systems. 2020.

[87] CHEN F, JI R, JI J, et al. Variational structured semantic inference for diverse image captioning [Z]. 2019.

[88] BLEI D M, KUCUKELBIR A, MCAULIFFE J D. Variational Inference: A Review for Statisticians[J]. CoRR, 2016, abs/1601.00670.

[89] WANG L, SCHWING A G, LAZEBNIK S. Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space[C]//Conference on Neural Information Processing Systems. 2017: 5756-5766.

[90] MAHAJAN S, GUREVYCH I, ROTH S. Latent Normalizing Flows for Many-to-Many Cross-Domain Mappings[C]//International Conference on Learning Representations. 2020.

[91] VIJAYAKUMAR A K, COGSWELL M, SELVARAJU R R, et al. Diverse Beam Search for Improved Description of Complex Scenes[C]//Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, 2018: 7371-7379.

[92] WANG Z, WU F, LU W, et al. Diverse Image Captioning via GroupTalk.[C]//International Joint Conference on Artificial Intelligence. 2016: 2957-2964.

[93] CHEN F, JI R, SUN X, et al. Groupcap: Group-based image captioning with structured relevance and diversity constraints[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1345-1353.

[94] BISHOP C M, NASRABADI N M. Pattern Recognition and Machine Learning: volume 4[M]. Springer, 2006.

[95] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[96] DOERSCH C. Tutorial on variational autoencoders[A]. 2016.

[97] BEPLER T, ZHONG E, KELLEY K, et al. Explicitly disentangling image content from translation and rotation with spatial-VAE[J]. Advances in Neural Information Processing Systems, 2019, 32.

[98] ZHENG Z, SUN L. Disentangling latent space for vae by label relevant/irrelevant dimensions [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2019: 12192-12201.

[99] LOCATELLO F, BAUER S, LUCIC M, et al. Challenging common assumptions in the unsupervised learning of disentangled representations[C]//International Conference on Machine Learning. PMLR, 2019: 4114-4124.

[100] HIGGINS I, MATTHEY L, PAL A, et al. beta-vae: Learning basic visual concepts with a constrained variational framework[Z]. 2016.

[101] BLEI D M, KUCUKELBIR A, MCAULIFFE J D. Variational inference: A review for statisticians[J]. Journal of the American statistical Association, 2017, 112(518): 859-877.

[102] KINGMA D P, WELLING M. Auto-Encoding Variational Bayes[C]//International Conference on Learning Representations. 2014.

[103] KRAMER M A. Nonlinear principal component analysis using autoassociative neural networks [J]. AIChE journal, 1991, 37(2): 233-243.

[104] SHEKHOVTSOV A, SCHLESINGER D, FLACH B. VAE Approximation Error: ELBO and Exponential Families[C]//International Conference on Learning Representations. 2021.

[105] RASMUSSEN C E, WILLIAMS C K I. Adaptive computation and machine learning: Gaussian processes for machine learning[M]. MIT Press, 2006.

[106] KINGMA D P, WELLING M. An introduction to variational autoencoders[A]. 2019.

[107] SOHN K, LEE H, YAN X. Learning Structured Output Representation using Deep Conditional Generative Models[C]//Conference on Neural Information Processing Systems. 2015: 3483-3491.

[108] LUCAS J, TUCKER G, GROSSE R B, et al. Understanding Posterior Collapse in Generative Latent Variable Models[C]//International Conference on Learning Representations. 2019.

[109] LUCAS J, TUCKER G, GROSSE R B, et al. Don't blame the Elbo! a linear Vae perspective on posterior collapse[J]. Advances in Neural Information Processing Systems, 2019, 32: 9408-9418.

[110] ABDAR M, POURPANAH F, HUSSAIN S, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges[J]. Information Fusion, 2021, 76: 243-297.

[111] HÜLLERMEIER E, WAEGEMAN W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods[J]. Machine Learning, 2021, 110(3): 457-506.

[112] BENDER E M, KOLLER A. Climbing towards NLU: On meaning, form, and understanding in the age of data[C]//Fifty-eighth Annual Meeting of the Association for Computational Linguistics. 2020: 5185-5198.

[113] KENDALL A, GAL Y. What uncertainties do we need in bayesian deep learning for computer vision?[J]. Conference on Neural Information Processing Systems, 2017, 30.

[114] CHEN D, DOLAN W B. Collecting highly parallel data for paraphrase evaluation[C]//The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011: 190-200.

[115] ZHANG Z, SHI Y, YUAN C, et al. Object relational graph with teacher-recommended learning for video captioning[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2020: 13278-13288.

[116] XU G, NIU S, TAN M, et al. Towards Accurate Text-Based Image Captioning With Content Diversity Exploration[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 12637-12646.

[117] YANG X, ZHANG H, CAI J. Learning to collocate neural modules for image captioning[C]// IEEE International Conference on Computer Vision. 2019: 4250-4260.

[118] BOWMAN S R, VILNIS L, VINYALS O, et al. Generating Sentences from a Continuous Space [C]//Conference on Computational Natural Language Learning. ACL, 2016: 10-21.

[119] KAY W, CARREIRA J, SIMONYAN K, et al. The Kinetics Human Action Video Dataset[J]. CoRR, 2017, abs/1705.06950.

[120] FANG H, XIONG P, XU L, et al. CLIP2Video: Mastering Video-Text Retrieval via Image CLIP [J]. CoRR, 2021, abs/2106.11097.

[121] WANG Q, CHAN A B. Describing like humans: on diversity in image captioning[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2019: 4195-4203.

[122] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet Large Scale Visual Recognition Challenge [J]. Int. J. Comput. Vis., 2015, 115(3): 211-252.

[123] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Thirty-first AAAI Conference on Artificial Intelligence. 2017.

[124] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

[125] LIU L, TANG J, WAN X, et al. Generating diverse and descriptive image captions using visual paraphrases[C]//IEEE International Conference on Computer Vision. 2019: 4240-4249.

[126] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2019: 4171-4186.

[127] QI P, ZHANG Y, ZHANG Y, et al. Stanza: A python natural language processing toolkit for many human languages[A]. 2020.

[128] AAFAQ N, AKHTAR N, LIU W, et al. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2019: 12487-12496.

[129] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks[C]//Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2019: 3980-3990.

[130] WU X, DYER E, NEYSHABUR B. When Do Curricula Work?[C]//International Conference on Learning Representations. 2021.

[131] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Annual Meeting of the Association for Computational Linguistics. 2002: 311-318.

[132] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74-81.

[133] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Annual Meeting of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005: 65-72.

[134] VEDANTAM R, LAWRENCE ZITNICK C, PARIKH D. Cider: Consensus-based image description evaluation[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4566-4575.

[135] ANDERSON P, FERNANDO B, JOHNSON M, et al. Spice: Semantic propositional image caption evaluation[C]//European Conference on Computer Vision. Springer, 2016: 382-398.

[136] FELLBAUM C. WordNet[M]//Theory and applications of ontology: computer applications. Springer, 2010: 231-243.

[137] ROBERTSON S. Understanding inverse document frequency: on theoretical arguments for IDF [J]. Journal of Documentation, 2004.

[138] CROUSE D F. On implementing 2D rectangular assignment algorithms[J]. IEEE Transactions on Aerospace and Electronic Systems, 2016, 52(4): 1679-1696.

致　谢

本研究持续了研究生的半数时光，如今已到了尾声。许多人对此工作都有非常大的帮助：感谢导师郑锋老师帮我找到这个研究方向，并同意自己对其中较偏僻的小课题进行深入研究，对这方向的兴趣和日益增多的收获也是支撑自己坚持下来的一股动力。同时感谢日常科研的催促、对论文写作和科研展示的真诚批评，这些都促使我更好的提升和思考。感谢经常一起讨论、合作的王腾师兄，他严谨踏实的做事风格为我树立了很好的榜样，诚恳热心的帮助则经常让我有所收获，很难想象没有他，论文中的很多细节、想法如何真正付诸实践。感谢讲授《机器学习》的郝祁老师、《贝叶斯推断》的蔡敬衡老师以及经常一起讨论交流更宽泛的相关背景知识的骆京同学、岳凤鹏同学、刘之豪同学等，他们让我对概率生成模型有更深入的理解——而这是完成本次工作不可或缺的高屋建瓴的知识体系。感谢实验室的所有成员，他们不仅仅毫无怨言地帮助我完成了问卷评测的一部分，平常积极讨论的学术氛围、友善融洽的学习环境也常常让我倍感温馨。感谢一路无条件支持我的家人们，他们为我创造了舒适安心的环境，让我可以对科研心无旁骛。

感谢南科大提供的充实的学习平台和美丽的校园环境，它是我缓解心情的“大观园”，一草一木、一山一水都曾留下我的痕迹；它又像是增长知识、探寻真理的“雅典学院”，沉思激辩、静默灵动的场面都仍历历在目、难以忘却。能在心智极度渴望发展的青年时期遇到这样的校园和周围的朋友，不知内心几次感慨何其幸运！感谢仅仅出现在一页又一页故纸堆、却常不停地徘徊在脑海里的伟大先哲们，他们孜孜不倦地探寻真理、思考本质的精神常常像永不枯竭的泉水滋润着我的心灵，成为我科研前行路上的深层动力，“高山仰止，景行行止，虽不能至，心向往之”！

最后衷心感谢在百忙之中评阅论文和参加答辩的各位专家、教授。本课题承蒙国家自然科学基金资助，特此致谢。

个人简历、在学期间完成的相关学术成果

个人简历

1996 年出生于山西省灵丘县。

2015 年 9 月考入大连理工大学软件学院数字媒体技术专业，2019 年 6 月本科毕业并获得工学学士学位。

2019 年 9 月至今，在南方科技大学计算机科学与工程系电子科学与技术专业攻读工学硕士学位。

获奖情况：参加“太阳风暴识别和预警人工智能挑战赛”并获得全国第一名，之后获得学校 2021 年校竞赛奖学金；参加计算机视觉顶会 CVPR 中关于“大规模视频行为密集视频标注”的比赛，并获二等奖。校英语阅读比赛二等奖。