

电子科技大学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

MASTER THESIS



论文题目 结合视觉内容理解与语言信息分析的
视频字幕生成技术研究

学科专业 计算机科学与技术

学 号

作者姓名

指导老师

学 院 计算机科学与工程学院（网络空间安全
学院）

分类号 _____ 密级 _____

UDC 注 1 _____

学 位 论 文

结合视觉内容理解与语言信息分析的视频字幕生成技术研究

(题名和副题名)

(作者姓名)

指导老师

电子科技大学 成都

(姓名、职称、单位名称)

申请学位级别 硕士 学科专业 计算机科学与技术

提交论文日期 _____ 论文答辩日期 _____

学位授予单位和日期 电子科技大学 年 月

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号。

Research on Video Captioning with Visual content Understanding and Linguistic Information Analysis

A Master Thesis Submitted to
University of Electronic Science and Technology of China

Discipline: **Computer Science and Technology**

Author: _____

Student ID: _____

Supervisor: _____

School: **School of Computer Science and
Engineering**

摘要

当今社会是一个多媒体的时代。随着网络传输速度的不断发展，人类社会的生活发生了巨大的变化：短视频占据人们更多的空余时间，网络娱乐直播和在线购物行业愈发繁荣。在这样的时代背景之下，跨媒体内容分析和理解成为了人工智能及深度学习领域一个极具需求和挑战性的研究热点。

视频字幕生成作为一个结合视觉分析和语言生成的跨模态任务，需要首先从视频内容出发，将人类可感知的视觉转换成机器可感知的特征化符号，然后借助训练的语言生成器得到准确、生动、详细描述视频内容的语句。如何更好地进行视觉内容的理解、如何更好地进行语言信息的分析，成为了视频字幕生成技术研究的基本问题。为了更好地进行视觉内容理解和文本的理解生成，许多研究进行了针对性地提升，但仍然存在一些重要问题需要完善：例如在视频的理解中，如何不依靠成本更高的标注数据补充，充分利用自身已有数据集，挖掘数据内部存在的联系来支持字幕生成；并且在当今天大规模预训练模型高速发展并引领深度学习前进的情况下，如何跟上时代的步伐，将预训练模型中的“知识”充分利用，服务于自身的任务进步。

针对上述问题，在本文提出相应的解决方案：

1. 对于内部知识挖掘，本文提出了基于支持集的视觉表达增强，在跨模态的语义空间更好地对视觉内容和语言信息进行理解，对学习过程中的视觉表达进行增强。通过利用构造支持集并建立灵活映射，优化模型的学习过程，以此得到更加灵活丰富的字幕。

2. 对于来自外部的预训练模型视觉-语言知识的挖掘，本文提出了基于跨模态预训练模型 CLIP 的关键词辅助视频字幕生成，在引入文本模态理解和指导的同时，推动对视觉内容理解的进步。通过利用 CLIP 视觉-语言统一的语义空间以及预训练大模型中存储的丰富知识，得到对于视觉信息的关键词，以此指导字幕生成。

最后，在视频字幕生成对应的两个数据集进行了详细的实验，证明了以上方法的有效性和先进性。

关键词：跨模态，视频理解，字幕生成，支持集，预训练

ABSTRACT

Today's society is in a multimedia era. With the continuous development of network transmission speed, human life has undergone significant changes: short videos now occupy more of people's spare time, and online entertainment live broadcasts and online shopping industries are becoming more and more prosperous. In this context, cross-media content analysis and understanding has become a research hotspot with great demand and challenges in the field of artificial intelligence and deep learning.

As a cross-modal task that combines visual analysis and language generation, video caption generation needs to start by analyzing the video content and converting human-perceived vision into machine-perceived feature symbols. Then, with the help of a trained language generator, accurate, vivid, and detailed sentences describing the video content can be obtained. How to better understand the visual content and how to better analyze the language information have become the basic problems of video captioning technology research. Many studies have made targeted improvements to better understand visual content and text, but there are still some important problems that need to be addressed. For example, in video understanding, how can existing datasets be fully utilized, and how can the connections within the data be mined to support the generation of captions without relying on more expensive annotated data supplements? In addition, with the rapid development of large-scale pre-training models and the progress of deep learning, how can researchers keep up with the times, make full use of the "knowledge" in pre-training models, and serve their own tasks?

To address these problems, this thesis proposes corresponding solutions:

For internal knowledge mining, this thesis proposes visual representation enhancement based on a support set to better understand visual content and language information in the cross-modal semantic space and enhance visual expression in the learning process. By constructing a support set and establishing a flexible mapping, the learning process of the model is optimized, enabling more flexible and rich captions.

For the mining of vision-language knowledge from external pre-trained models, this thesis proposes a keyword-assisted video caption generation based on the cross-modal pre-trained model CLIP, which promotes the progress of visual content understanding while introducing text modal understanding and guidance. By utilizing the unified vision-

ABSTRACT

language semantic space of CLIP and the rich knowledge stored in the pre-trained large model, keywords for visual information are obtained to guide the generation of captions.

Finally, detailed experiments are carried out on two datasets corresponding to video caption generation, which prove the effectiveness and advancement of the above methods.

Keywords: Cross Modality, Video Understanding, Captioning, Support Set, Pre-training

目 录

第一章 绪 论	1
1.1 研究工作的背景与意义	1
1.2 视频字幕生成的国内外研究历史与现状	3
1.3 本文的主要贡献与创新	5
1.4 本论文的结构安排	7
第二章 视频字幕生成算法基础	9
2.1 相关技术	9
2.1.1 视频特征提取	9
2.1.2 注意力机制	17
2.1.3 循环神经网络	20
2.2 本章小结	21
第三章 基于支持集视觉表达增强的视频字幕生成	22
3.1 动机	22
3.2 基于支持集的视觉表达增强	24
3.2.1 模型框架设计	25
3.2.2 支持集构建	25
3.2.3 语义空间转换	26
3.3 实验结果	29
3.3.1 数据集	29
3.3.2 评价指标	30
3.3.3 执行细节	31
3.3.4 定量结果对比	32
3.3.5 消融实验	33
3.3.6 可视化展示	36
3.3.7 存在的不足以及未来工作	36
3.4 本章小结	38
第四章 基于预训练模型视觉-语言知识挖掘的视频字幕生成	39
4.1 动机	39
4.2 基于预训练模型的视觉-语言知识挖掘	40
4.2.1 辅助关键词的提取	41

4.2.2 多注意力的双层解码器	42
4.2.3 结合指针网络的生成模块	43
4.3 实验结果	44
4.3.1 数据集及评价指标	44
4.3.2 执行细节	44
4.3.3 定量结果对比	45
4.3.4 消融实验	46
4.3.5 可视化展示	48
4.3.6 存在的不足以及未来工作	48
4.4 本章小结	49
第五章 全文总结与展望	51
5.1 全文总结	51
5.2 后续工作展望	51
致 谢	53
参考文献	54
攻读硕士学位期间取得的成果	60

第一章 绪论

1.1 研究工作的背景与意义

随着数字媒体技术的不断发展和进步，视频已经成为人们获得知识、了解他人生活和休闲娱乐的重要手段。与文本和图像等传统媒体形式不同，视频利用各种感官模式，如视觉和声音，为我们周围的世界提供了更加身临其境和生动的表现。除此之外，新的数字媒体形式，如网上购物和直播，也成为了人们生活的常态。这些数字媒体形式的存在产生了大量的视频数据，并提高了对复杂视频内容理解能力的需求：跨媒体成为当今的一个研究热点^[1]。

跨媒体需要深入了解各种形式的媒体之间的联系和差异，需要具备先进的智能处理能力，以便更好地提取、整合多种数据形式的信息。只有通过对跨媒体的研究和应用，才能有效满足人们对数字媒体的需求，推动数字时代的发展和进步。

而视频字幕则是跨媒体领域一个经典并且仍然火热的研究课题^[2]。如图1-1所示，视频字幕是生成视频内容的文本描述的过程，通过网络设计实现从视频内容到文本内容的变换。它是计算机视觉和自然语言处理领域的一项重要任务，对各种应用：包括多模态检索^[3]、视觉问答^[4]等有重大影响。对视频字幕的需求来自于越来越多的在线视频数据，以及使这些内容更容易理解、获取和搜索的期望。视频内容在社交媒体平台、线上学习平台和优酷油管等流媒网站上越来越受欢迎，但与此同时，这些内容缺乏文字描述和对视频内容的有效总结。这种描述的缺乏使得人们很难找到他们要找的视频。又或者，抛开线上的业余生活，在某些现实场景下也会有对于视频描述的需求：比如视力残障人群，他们也需要通过这种辅助手段获取接触和理解这个世界的途径，获得更好的生活质量。

视频字幕作为一个跨模态领域的经典课题，历经多年的发展并取得了一定的进步。发展起初，人类通过人工智能手段进行图片内容的理解，该课题也称作图片描述或图片字幕生成^[5]。在此课题中，机器会理解图像的内容，并且对图片展示的主体和事件进行全面的描述。随着数字媒体的不断发展，视频取代图片，短视频平台取代过往的信息论坛和网站，成为人们生活的主流，视频字幕也逐渐获得了更多的关注。视频相较于图片，描述的难度更大：首先是视频具有一定的时长，在时间序列上视频产生了主体和事件的变化，机器需要识别变化并学会关注需要关注的点，给一个视频生成一个详细且准确的描述；其次，视频作为多种跨媒体形式的载体，它具有视听艺术的综合，并且在视频中文字也会作为一种信息载体需要被关注，这使得视频描述的困难程度相较图片而言有剧烈的上升。怎样

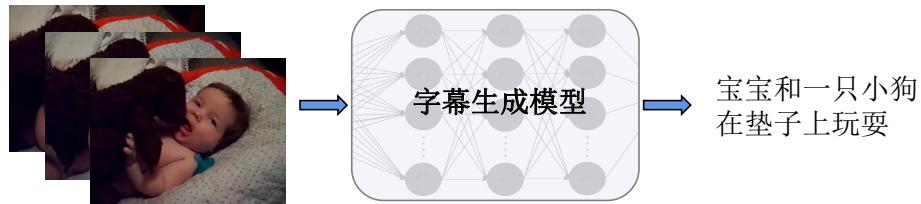


图 1-1 视频字幕生成任务示意图

更好的获取每个模态的信息，如何更好的综合多个模态的内容，成为了视频字幕生成的重要话题。

在探索更佳表现的过程中，视频字幕生成领域随着人工智能的发展走过了一段漫长的过程^[6,7]。随着卷积神经网络（CNN）的出现，视觉信号可以被转换为机器识别的特征向量，为视频字幕生成提供了更广阔的发展空间。之后，循环神经网络（RNN）的出现，在序列生成任务中大放异彩，这为视频字幕生成的发展提供了新的方向。但由于 RNN 对长期依赖的建模能力不足，生成的文字序列有时会缺乏连贯性和上下文信息的考虑，为了克服该问题，注意力机制在视频字幕生成中得到了广泛的应用。注意力机制可以在生成每个单词时动态地分配注意力，以便于更好地利用视频中的信息，从而提高了生成文字描述的准确性和流畅度。同时，注意力机制可以帮助模型更好地处理长期依赖的信息，生成更准确、连贯的文字描述。如今双层的注意力 LSTM^[8] 成为了当今视频字幕生成器的主流范式。就生成思想上，视频字幕生成也经历了许多模型和框架上的变更：从最初的模板化的描述，再到如今的深度学习编码器-生成器模式，不断的技术进步和创新推动着视频字幕生成的质量和多样性不断提升，使得对视频内容的概括能力得到了显著的改善。

尽管获得了非常多的进展，时至今日，视频字幕生成任务仍面临着重大挑战，比如说如何更加准确的进行语义理解：识别场景中不同物体和事件之间的关系，以及识别抽象概念，如情感或隐喻；又比如词汇限制：现有的视频字幕模型可能无法识别特定视频中存在的所有词汇和语言结构，这可能导致不准确或不完整的字幕，不能完全表达视频的含义；又或许该如何利用最新技术服务于视频字幕生成这种经典课题，使其跟上时代潮流的发展。所以，视频字幕仍然面临一些需要解决的挑战和困难，应对这些挑战需要开发更复杂的算法，以处理不同的视频类型，整合跨模式数据，并及时有效地生成准确和信息丰富的字幕。某种意义上，问题的关键也可以总结为，如何将视频表达出的“知识”，以人类的视角进行呈现。

总之，视频字幕生成是结合计算机视觉领域和自然语言处理领域的一项重要课题，它涉及生成视频内容的文本描述的过程，需要挖掘多种丰富、复杂的信息。

生成更合理生动的视频字幕对各种辅助应用和视觉理解任务有重大影响和意义。

1.2 视频字幕生成的国内外研究历史与现状

视频字幕生成是一个通过机器手段自动为视频内容生成文本描述的任务，其目的是创建一个连贯而有意义的句子或系列句子，描述视频中描述的主要事件、对象和行动。视频字幕生成是多模态任务的先驱，这一领域的研究经历了较长时间演化：从基于模板的生成形式到基于深度学习的编码器-解码器形式，国内外研究者不断推动着视频字幕生成的进步和发展^[9]。视频字幕生成领域是字幕生成的一个子课题，与图像的字幕生成互相补充，又略有不同。视频作为一个多模态载体，承载着比图片信息更为丰富的语义和内容，与此同时也包含着图片所具有的视觉特征，因此视频字幕的发展基于图片字幕生成，又在此之上针对视频更为复杂特性进行了深度的拓展，使其更加适配视频的特点。

字幕生成的研究发展的最初提出了第一批自动生成字幕的方法。这些早期的方法大多是基于人工创建的规则，并依靠手工制作的特征来描述视频的内容：这些方法通常包括将视频分解成不同的场景，并使用人为规定的规则来描述场景和正在发生动作^[10-15]。例如使用物体识别和运动分析来检测视频中的物体和运动，然后使用一组预定义的规则来生成描述物体动作的字幕^[12,13]；又例如结合语言规则，基于隐马尔科夫^[16]或知识本体论来生成描述视频视觉内容的字幕^[10,11]。然而，这些早期的方法在生成准确和连贯的字幕方面能力有限，因为它们依赖于手动创建的规则，可能无法捕捉和匹配不同视频的复杂性和可变性，与此同时，他们还缺乏学习和适应不同类型视频和内容的能力，这使得他们在为涉及领域更为广泛的视频生成字幕时效率较低。尽管存在这些局限性，但这些早期的方法在为开发更先进和更复杂的视频字幕方法奠定基础方面作出了非常重要的贡献。

近年来机器学习和深度学习技术逐渐涌现并不断发展成为当今人工智能的主流技术。在过去的十年里，视频字幕生成的研究从深度学习和神经网络的发展中获益良多。如今，最先进的视频字幕方法是基于深度神经网络的框架，学习将输入的视频帧映射到自然语言句子。该框架通常由两部分组成：一个从视频帧中提取特征的视觉编码器，以及一个逐字生成字幕的语言解码器，因此也被称为：编码器-解码器架构^[17]。在实践中，这一架构通常由一个视觉提取编码器（一般为卷积神经网络）和一个语言生成解码器（一般为循环神经网络）组成，卷积神经网络编码器处理输入的视频帧以提取其在特征空间的视觉表达，而长短期记忆循环神经网络解码器则利用输入的视觉信息在时序上逐步解码，生成相应的句子。近年来的研究工作还探索了其他深度神经网络架构的使用，如基于转化器（Transformer）

[18] 的模型和基于图形的模型，用于视频字幕。这些模型通常更复杂，需要更多的训练数据，与此同时，模型更好的学习能力和信息捕捉能力使它们可以产生更准确和富有语境的字幕。

在编码器-解码器的基础框架上，视频字幕生成经历了许多发展和进步，研究者从多种角度推动着模型性能的进步以及描述多样性和丰富性的优化。第一类为那些继续原有的框架，专注于多模态的融合或表征嵌入方法的工作^[19-25]。例如：“针对视频字幕生成的多模态注意力建模网络”^[19] 创建了一个共享的视觉和文本记忆，以模拟长期的视觉-文本依赖。它结合了多种模式，包括视觉特征、音频特征和语言特征，并使用注意力机制引导全局注意力在描述的具体目标上。具体操作上，所提出的方法使用了一个具有新颖注意力机制的递归神经网络，使网络能够在不同的时间步骤中关注输入特征的不同部分；模型还包括一个记忆模块，使网络能够存储以前生成的词语的信息，并利用这些信息为未来的词语预测提供信息，其中记忆模块由一个使用一个外部记忆矩阵来存储和检索信息的动态记忆网络组成。又例如工作“针对视频字幕的时空动态和语义属性增强的视觉编码”^[21] 提出了一种新的视频字幕编码方法，旨在捕获视频数据中的空间和时间信息以及与视频内容相关的语义属性，该方法采用两阶段方法，首先使用时空卷积神经网络对视觉信息进行编码，然后使用语义属性增强的视觉编码模块对视觉内容根据语义属性进行补充。在此基础上，引入的存储模块将对象类别和场景类型等语义属性合并到视觉编码中，由此可用于生成更准确和信息丰富的标题。

第二类为引入网格特征来捕获更多网格级空间信息，从而更好地进行空间关系表征和建模的^[26,27]。例如：工作“视频字幕的运动引导空间注意”^[26] 提出了一种新的视频字幕方法，该方法利用视频中不断变化的光流信息来引导神经网络中的空间注意机制，在此基础上，为了更好地捕捉光流的变化和联系，引入门控注意力递归单元以适应性地纳入之前时间步的注意力图，以生成字幕。此外，论文“视频字幕的运动引导区域消息传递”^[27] 基于以上工作又提出了一种新的视频字幕方法，该方法使用运动引导视觉区域表征之间的消息传递来捕获视频数据的空间和时间动态：本工作提出了基于网格特征的循环区域注意模块，以更好地提取多样的空间特征，并通过采用运动引导的跨帧信息传递使模型更好适应时间结构，并能够在跨帧的不同区域之间建立高阶关系。以上设计能够鼓励信息交流，使模型产生紧凑且更佳的视频表示。

第三类为引入目标检测器来增加物体之间的时空交互的方法^[28-30]。例如，工作“视频字幕生成领域符合语法的动作标记”^[29] 提出了一个语法感知动作定位模块，该模块通过同时参考物体目标主体和视频动态，明确地学习动作，具体的，首

先通过在主体和视频动态之间建立全局依赖关系来识别目标，然后从一个共同的空间解码目标的动作描述，由此更好的利用了目标检测器的指导作用。又例如工作“为视频字幕学习离散推理模块网络”^[30]提出了一种新的视频字幕方法，该方法为现有的编码器-解码器框架配备了推理能力，具体来讲，该工作采用了三个的时空推理模块：“目标定位”、“关系捕捉”和“功能文本”分别进行“名词”、“动词”和其他组成成分的预测，并设计了一个动态的、离散的模块选择器，该选择器由具有 Gumbel 近似^[31]的语言损失训练，根据其预测概率决定该时间步产生文本的具体模块，让生成的字幕更好地符合人类社会的语言逻辑。

第四类的视频字幕研究如何纳入外部知识源，如语义知识库或语言模型，并且通过利用这些外部信息，使视频字幕模型可以生成信息量更大、更连贯的字幕^[25,32]。在这种方法中，论文“使用检索-复制-生成网络的开放式书籍视频字幕”^[25]提出了一种新的视频字幕方法，该方法使用检索-复制-生成网络，通过利用开放式书籍的外部知识来生成字幕，该方法使用检索-复制-生成网络从外部知识源中检索相关信息，将检索到的信息复制到字幕生成过程中，这样生成的字幕将会更加详细丰富，取得更好的性能。论文“具有辅助任务的视频字幕和视频问答分层表示网络”^[32]提出了一种用于视频字幕和视频问题回答的新方法，该方法使用带有辅助任务的分层表示网络来学习视频数据的更全面表示，使用分层表示网络来模拟视频数据的时间动态。分层网络由多个抽象层次组成，每个层次都经过辅助任务的训练，以学习数据的更全面的表示。然后，该方法使用这种分层表示来生成字幕或回答关于视频的问题，使得生成的文本更加符合丰富生动的需求。

视频字幕研究是一个动态的、不断发展的领域，未来还有许多令人兴奋的研究方向和挑战有待探索。总的来说，视频字幕研究在过去十年中取得了重大进展，最先进的模型可以为广泛的视频内容生成信息量大、背景丰富的字幕。然而，这一领域仍有需要改进的部分，特别是在当前的研究现状下，应该如何在有限的资源下继续研究，任重而道远。

1.3 本文的主要贡献与创新

本文的研究任务名为视频字幕生成，旨在能够准确、生动、全面地概括视频中的人物及事件，该领域结合了视觉内容的理解表征以及语言的理解与生成，是一个跨学科交叉任务。视频字幕生成领域的研究，不仅可以帮助盲人等特殊人群更好地了解视频内容，还可以为工业界场景下的视频标注、内容总结等应用提供有力支持，具有重要的理论和实践意义。

如图1-2所示，由视频字幕生成出发，本文将解决两个核心问题：首先针对信



图 1-2 本文的研究思路

息挖掘不充分的问题，工作一研究内部知识的挖掘，提出了基于支持集的视觉表达增强来进行视频字幕生成；针对视频描述匮乏不生动的问题，工作二研究外部知识的引入，提出了基于预训练模型视觉-语言知识挖掘的视频字幕生成，以下将针对主要贡献和创新进行详细介绍。

工作一针对对现有视觉数据信息挖掘不充分的问题，提出利用支持集来进行表达增强，以此更好的挖掘内部知识。本工作提出了一个基于支持集的视觉表示增强 (SMRE) 模型来利用样本之间共享的语义子空间中的信息，优化模型视觉表达的能力。本质上，本工作的模型是基于一个典型的编码器-解码器框架，在此基础上，通过语言信息指导的支持集构建 (SC) 模块获取一个能够学习到与原视觉输入具有相似视觉元素的支持集，并将其通过一个新的输入分支（该分支与原始路径共享编码器-解码器），建立灵活的映射。此外，为了在视觉-语言共享的子空间中建模复杂的语义关系，本工作引入了具有额外损失设计的语义空间转换 (SST) 模块来约束多模态语义空间中的相对距离，并分别从模态间和模态内两个角度进行自监督约束。利用以上设计，编码器-解码器结构可以学习更好的多模态空间语义表示并生成语义丰富的字幕。总结起来，本工作的贡献有三个方面。1) 与传统的一对一映射相比，通过挖掘语义相关的视觉元素，在视频字幕生成课题提出支持集概念，构建了一种新颖灵活的映射框架，以更好地捕捉样本之间的内部细节联系 2) 为了进一步约束语义关系，本工作从模态间和模态内两个角度提出了基于自监督设计的语义空间转换模块，使得模型表达能力进一步增强 3) 本工作的 SMRE 在 MSVD 和 MSR-VTT 基准数据集上实现了最先进的性能，在 MSVD 上比现有的 SOTA 工作在 BLEU-4 和 CIDEr 上分别高出 5.3% 和 1.1%，在 MSR-VTT 上分别高出了出 1.5% 和 0.4%。

工作二针对视频字幕的表述匮乏和不生动的问题，设计利用大规模预训练模

型 CLIP 中的视觉-语言知识，进行外部知识的引入。本工作提出了一个基于 CLIP 统一语义空间知识进行辅助关键词提取，进而融合视觉-语言模态的视频字幕生成方式。具体来讲，本工作基于经典编码器-解码器框架，但与过往工作不同的是，本工作首先希望获取大规模预训练模型中的丰富知识来指导字幕的生成，于是引入了 CLIP 跨模态预训练模型，一个拥有共享的视觉-语言语义空间的多模态表征工具，通过直接对视频抽取的图像帧进行零样本预测，得到预训练模型知识指导的关键词，从而引入外部知识，并将其作为文本特征与视觉特征一起通过统一建模，进而进行多模态信息的理解和输入。并且，在解码器部分采用结合指针网络的生成模块，通过动态地从关键词词库和原有词库中选择词汇，扩展了原有的单一来源生成方式。与此同时，在原有双层注意力 LSTM 上，构建了文本注意力-视觉注意力的多注意力机制，通过增加文本模态注意力机制的指导，更好地融合多模态特征。总结起来，本工作的贡献有三方面。1) 通过利用大规模预训练模型中的跨模态外部知识，得到能够指导字幕生成过程的关键词文本，构建了多模态的视频字幕生成范式，使得字幕表达匮乏的问题得到缓解 2) 针对多模态的信息输入，构造了文本注意力-视觉注意力的多注意网络以及结合指针网络的生成模块，使得文本特征能更好的指导视频字幕生成 3) 本工作的模型在 MSVD 和 MSRVTT 上的性能表现证明了该方法的有效性和先进性。

1.4 本论文的结构安排

本文的整体章节结构安排如下：

在第一章，首先介绍了视频字幕生成的背景和意义。从现今视频作为一种跨媒体形式在人类日常生活中的愈发重要的现状出发，引出视频字幕生成任务，并且进一步介绍了视频字幕生成的定义及其应用价值，概述了视频字幕生成领域的发展过程和重要的成果产出，虽然获得了非常大的进步，但该领域仍然面临着重大挑战，在本章第一小节对于挑战进行分析总结。另外，在本章介绍了视频字幕生成的国内外研究历史与现状，分类汇总了先进工作及其提出的方法，对问题的思考具有重要意义。在本章最后，总结了本文的主要贡献与创新。

在第二章，主要围绕视频字幕生成算法需要的基础进行描述。本章主要介绍与视频字幕生成相关的多个经典和先进研究成果。首先介绍了视频特征提取方面的两类特殊网络：外观表征提取网络和运动表征提取网络。其中，对于 ResNet、InceptionNet、I3D、C3D 和 3D ResNeXt 等比较突出的工作进行了详细介绍。接着，从模态处理出发，介绍了注意力机制，并介绍了基于此的 Transformer 模型及其最新进展（ViT、CLIP）。这些模型在多模态领域和计算机视觉领域表现出卓越的建

模能力和理解水平。最后，对视频字幕生成的解码器部分进行了详细介绍，包括循环神经网络的基础和数据传输过程，以及从最早的 RNN 到 LSTM 门控机制的变化。这些网络都为视频字幕生成的基础打下了坚实的基础，并为其未来的发展提供了有力支持。

在第三章，主要介绍了“基于支持集视觉表示增强的视频字幕生成”。在本章中，针对信息挖掘不充分的问题，提出了一个基于支持集的视觉表示增强 (SMRE) 视频字幕模型，以建立灵活的映射关系，并在样本之间共享的视觉-语言语义子空间中挖掘信息。具体来说，通过设计支持集构建 (SC) 模块和语义空间转换 (SST) 模块来捕捉多模态语义空间中的内在联系，从而获得更好的语义表示。在 MSVD 和 MSR-VTT 上的实验结果证明了该方法的优越性。

在第四章，主要介绍了“基于预训练模型视觉-语言知识挖掘的视频字幕生成”。本章针对传统方法由于固定的训练数据以及缺乏适当指导导致的描述匮乏不充分的问题，引入了 CLIP 预训练模型存储的外部知识。通过基于 CLIP 视觉-语言跨模态共享的语义空间及其零样本预测能力，获得了外部知识指导下的辅助关键词，并利用多注意力的双层解码器和结合指针网络的生成模块将辅助关键词和生成模型有机结合，得到了新的字幕生成方法。在 MSVD 和 MSV-VTT 上的大量实验证明了该模型的有效性和先进性。

在第五章，对整体文章进行分析和总结，并且在此基础上进行反思和展望，最终陈述对未来工作的思考。

第二章 视频字幕生成算法基础

2.1 相关技术

2.1.1 视频特征提取

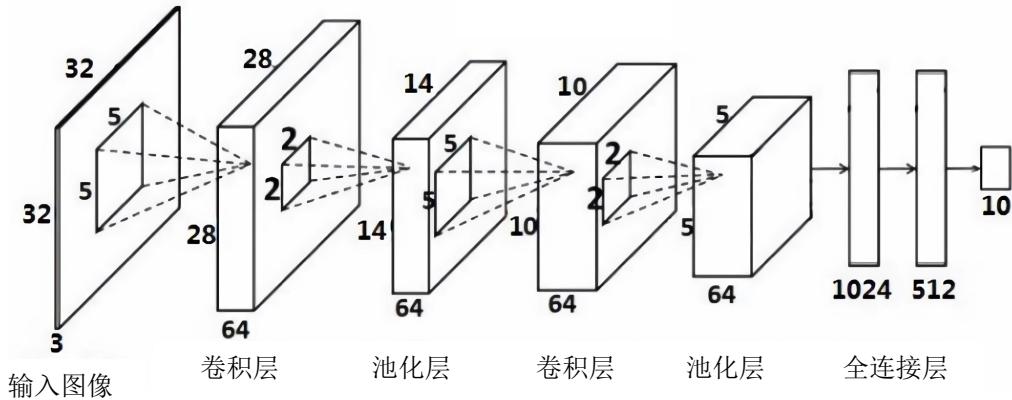
视频特征是研究和处理视觉信息的关键, 可以用于开发机器学习和深度学习模式, 并基于此进行进一步的内容理解。针对视频特征提取, 涉及从视频帧中提取两种主要类型的特征: 外观表征和运动表征。外观表征是一个物体的视觉外观的特征, 如颜色、纹理和形状。它们对于识别具有明显视觉特征的物体非常有用, 如人脸、动物或汽车。运动表征来自视频中物体的运动, 如速度、方向和加速度。它们对于跟踪物体在场景中的移动非常有用, 并且可以帮助区分不同类型的运动, 如行走、跑步或跳跃。为了从视频帧中提取这些特征, 研究人员使用了各种技术, 包括深度学习、光流和运动估计算法等。以下将分别详细介绍常用的外观表征提取网络和运动表征提取网络。

2.1.1.1 外观表征提取网络

外观表征提取网络主要基于 2D 的卷积网络, 与图像的特征提取类似, 在长宽两个维度对视觉进行建模, 通过对图像 RGB 三通道表征, 将图像数字化为 [3, H, W] 的机器表示, 并进一步通过二维的特征提取模型得到视觉的向量表征。

卷积神经网络经过几十年的发展和完善, 一路走来取得了无数的突破和里程碑。在 1980 年代, 第一个卷积神经网络^[33]由 Yann LeCun 团队开发释出, 他也被认为是卷积神经网络的开拓者。这些早期的模型是为识别手写数字而设计的, 在模型结构上简单采用卷积网络和全连接层的组合, 为人工智能的发展奠定了基础。在此基础上, 科学研究人员继续探索完善基于卷积神经网络的图像识别任务, 这些探索在提高模型的准确性方面取得了显著的进展, 最终不幸受制于当时有限的计算能力, 限制了模型的大小和复杂性。2000 年代, 强大的 GPU 的出现和大型数据集(如斯坦福大学李飞飞团队提出的 ImageNet^[34] 数据集)的出现使研究人员能够开发出更大、更复杂的卷积神经模型。在此之后, LeNet(1998)^[35]、AlexNet(2012)^[36]、VGG(2014)^[37]、Inception^[38]、ResNet(2015)^[39]、Xception(2016)^[40] 和 Inception-ResNet^[41] 系列不断的在图像分类方面取得突破, 模型更大更复杂, 表征能力更为强大, 在 ImageNet 等数据集上的表现不断进步, 不断地推陈出现, 使得人工智能进入了蓬勃发展的时段。

在这些模型中, 近期较为突出的工作是由微软团队提出的 Resnet 和谷歌团队

图 2-1 经典的卷积网络 LeNet-5^[33]

提出的 Inception-ResNet-v2^[41] 网络。接下来将进行详细介绍。

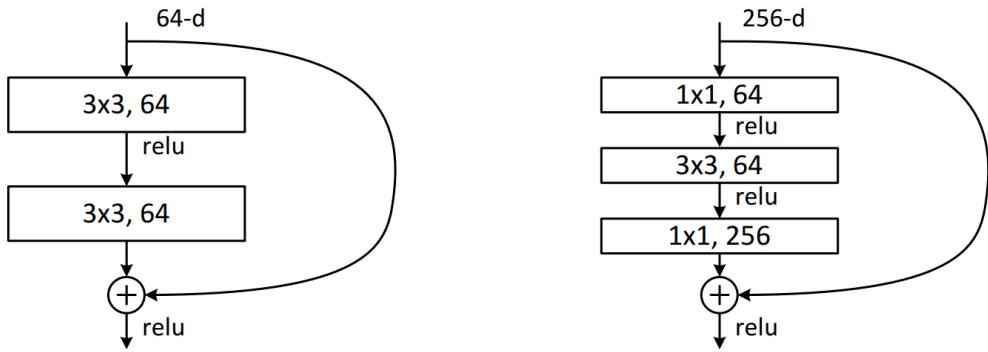
经典的卷积神经网络是一种在计算机视觉领域广泛应用的深度学习模型。其设计基于输入图像的像素表示，即图像的维度为 [3, H, W]，其中 3 代表图像的 RGB 三通道，H 和 W 分别代表图像的高度和宽度。如图2-1所示，通过堆叠的卷积层和采样层（也称为池化层），CNN 可以对输入图像进行特征提取和降维处理。卷积层采用卷积运算对图像进行滤波处理，提取出图像的局部特征。采样层则用于对卷积层的输出进行降采样处理，减少特征图的维度，同时保留重要的特征信息。经过堆叠的卷积层和采样层之后，卷积神经网络会得到一个特征图，其中每个像素代表一种特定的特征。为了进行分类或其他下游任务，这个多维特征需要被拉成一个一维向量表示。卷积神经网络通常会添加一些全连接层，将特征图中的多维特征转化为一个一维的向量，然后通过全连接层进行其他下游任务，例如图中所表示的 10 分类。在全连接层中，每个神经元都与上一层中的所有神经元相连，使得模型可以学习到特征之间的复杂关系和相互作用。

卷积层操作可以表示如下：给定一个张量 \mathbf{x} ，它的形状为 $C_{in} \times H \times W$ 和一组 C_{out} 滤波器 \mathbf{F} 形状为 $C_{in} \times k \times k$ ，每个卷积层基于填充 p 和步长 s 进行 2 维卷积操作，并且应用一个偏差值 b ，最终输出一个 $C_{out} \times H' \times W'$ 形状的输出 y ，见式 (2-1)：

$$\mathbf{y}_{i,j,k} = \sum_{c=1}^{C_{in}} \sum_{r=1}^k \sum_{s=1}^k \mathbf{F}_{i,c,r,s} \mathbf{x}_{c,(i-1)s+r+p, (j-1)s+s+p} + \mathbf{b}_i, \quad (2-1)$$

其中 i 从 1 到 C_{out} ； j 从 1 到 H' ； k 从 1 到 W' 进行计算。

池化层操作可以表示如下：给定一个张量 \mathbf{x} 它的形状为 $C \times H \times W$ 和一组大小为 $k \times k$ 的池化内核，步长设置为 s ，每个池化层在每个 $k \times k$ 的非重叠区域以步长

图 2-2 残差神经网络^[39]

s 进行最大池化或平均池化，公式表达为式（2-2）：

$$\mathbf{y}_{c,i,j} = \max_{r=1}^k \max_{s=1}^k \mathbf{x}_{c,(i-1)s+r,(j-1)s+s}, \quad (2-2)$$

其中 i 从 1 到 $\lfloor \frac{H}{s} \rfloor$; j 从 1 到 $\lfloor \frac{W}{s} \rfloor$; c 从 1 到 C 。

这样的范式统治了视觉表征很多年，卷积神经网络为了追求更好的性能不断进行网络层数的堆叠，在短期内得到了性能的飞跃。但与此同时，神经网络随着网络结构不断复杂化、网络层数的不断加深，反向传播过程中的梯度会变得非常小，这使得网络很难有效地学习。这个问题在深度网络中尤为突出，被称为深度神经网络的梯度消失问题，限制着网络规模的扩大和增长。而深度网络在图像分类和其他任务中越来越受欢迎，这个问题变得重要而亟待解决。

在这样的背景下，ResNet（残差网络）被提出了。它是一个深度神经网络架构，由微软的研究人员在 2015 年开发，旨在克服非常深的神经网络中梯度消失的问题。ResNet 引入了“短路”的概念，它允许信息绕过网络的某些层，直接传递到后面的层，这使得远距离的梯度依赖得到保留，有助于解决梯度消失的问题，并允许 ResNet 比以前的神经网络架构更深入，模型规模扩展到了 ResNet-152。ResNet 在许多计算机视觉任务上取得了最先进的成果，包括分类、检测任务和语义分割任务。它被广泛用于研究和工业领域，并被调整为计算机视觉以外的各种应用，如语音识别和自然语言处理。

ResNet 的基本构件被称为残差块，如图2-2所示它由两个或三个卷积层和一个跳过一个或多个层的捷径（也被称为“短路”）连接组成，ResNet 文章中主要给出了两种实现方式，左侧残差结构称为 BasicBlock，右侧残差结构称为 Bottleneck。这些残差块可以堆叠起来，形成一个具有数百甚至数千层的非常深的网络。残差

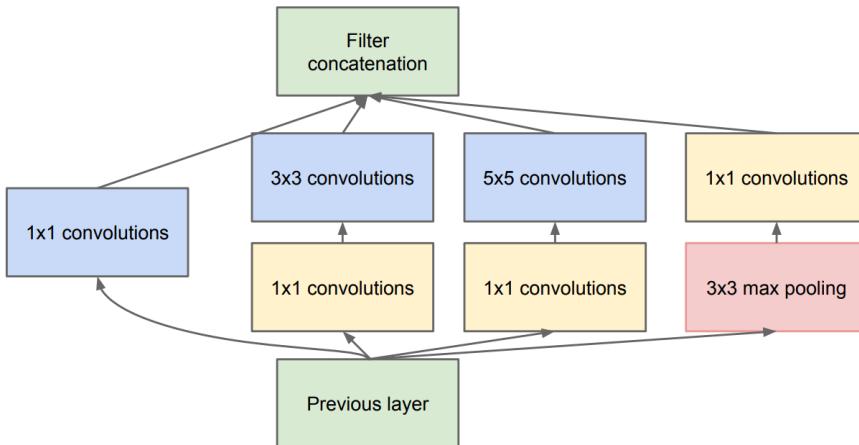
块的基本结构可以用以下公式（2-3）表示：

$$\mathbf{y} = \mathbf{F}(\mathbf{x}, \mathbf{W}_i) + \mathbf{x}, \quad (2-3)$$

其中 \mathbf{x} 是块的输入； \mathbf{y} 是输出； \mathbf{F} 是代表块所要学习的映射的残差函数； \mathbf{W}_i 是块中卷积层的权重。残差函数 \mathbf{F} 被定义为一系列卷积层的输出，然后是一个非线性激活函数，如 ReLU。这个残差映射被加到输入 \mathbf{x} 上，得到输出 \mathbf{y} ，然后被传递到下一层。捷径连接只是将输入 \mathbf{x} 加入到残差函数 $\mathbf{F}(\mathbf{x}, \mathbf{W}_i)$ 的输出中。这允许网络学习残差映射 $\mathbf{F}(\mathbf{x}, \mathbf{W}_i)$ ，而不是直接映射 $\mathbf{H}(\mathbf{x}, \mathbf{W}_i)$ ，其中 $\mathbf{H}(\mathbf{x}, \mathbf{W}_i)$ 是没有捷径连接的块的输出。通过学习残差映射，网络可以避免梯度消失问题并提高其性能。

关于这样的残差块能够融合缓解梯度消失的问题，作者在文中给出了解释。在深度卷积神经网络中，损失函数相对于网络参数的梯度在通过多层反向传播时可能变得非常小，这被称为梯度消失问题，它可能使训练深度网络变得困难，因为权重在训练过程中可能不会被正确更新。通过引入捷径连接，快捷连接绕过卷积层，将输入添加到残差函数的输出，有效地创建了一个“跳过连接”，使梯度直接从输出流向输入，而不需要经过太多的层。这可以帮助防止梯度变得太小，在网络中向后传播时消失。此外，通过使用剩余块，网络可以比传统网络“更深”，同时仍然保持其性能。这是因为残差块中的捷径连接允许网络在不牺牲性能的情况下增加新的层，因为残差映射可以简单地学习一个通过短路连接的特征映射。关于 ResNet 中的剩余块如何解决梯度消失问题的另一个角度是为网络提供一个更好的优化环境来学习：在深度神经网络中，每一层都会引入一个非线性，会导致梯度变得非常小，甚至是零，这可能使网络难以从数据中学习并在训练中优化损失函数。通过学习剩余映射而不是直接映射，网络可以被认为是试图学习输入和输出之间的剩余误差，而不是试图学习输出本身。这可以被看作是一种正则化的形式，因为它鼓励网络把注意力放在需要纠正的错误上，而不是试图完美地适应数据。此外，通过使用剩余块，网络可以以更有效的方式学习数据的分层表示。捷径连接允许信息绕过某些层，直接传输到更深的层，使网络能够学习更复杂的特征和模式。这可以帮助网络更好地捕捉数据的底层结构并提高其性能。

外观表征提取网络的另一篇突出文章为 InceptionNet，一个深度神经网络架构系列，由谷歌的研究人员在 2014 年开发。InceptionNet 的目标是创建一个网络，通过降低网络的计算成本，可以有效地从大量的数据中学习。Inception 系列与以往工作的不同点可以总结如下：Inception 模块由一系列具有不同滤波大小的卷积层组成，然后是池化层，再将这些层的输出特征图串联起来。这使得网络能够学习输入数据的局部和全局特征，并捕捉数据中更复杂的模式，如图2-3所示。另一个

图 2-3 Inception 网络^[41]

不同之处是使用了 1×1 卷积，在 InceptionNet 中，在计算成本更高的 3×3 和 5×5 卷积之前，它被用作降维的手段。这有助于减少网络中的参数数量，提高其效率。此外，InceptionNet 在训练过程中使用了一个辅助分类器，为网络增加了额外的监督和规范化，提高了网络的性能，减少了过拟合的风险。InceptionNet 有几个版本，每个版本都有不同的层数和不同的设计选择，也有自己独特的优势和探索。

Inception v1: 于 2014 年提出是 InceptionNet 的第一个版本，并引入了 Inception 模块的概念。这些模块由多个具有不同过滤器大小的卷积层组成，然后是池化层。然后将这些层的输出特征映射连接起来，形成模块的输出。Inception V1 有 22 层，在 ImageNet 数据集上进行训练，实现了当时最先进的性能。这个版本的缺点之一是它有大量的参数，使得它的计算成本很高，很难训练。

Inception v2: 这些是 InceptionNet 的更新版本，在 2015 年推出。对 Inception 模块进行了一些改进，比如用多个较小的卷积替换一些较大的卷积，并使用批处理规范化。这些变化有助于减少网络中的参数数量，提高网络效率。此外，Inception V2 引入了一个名为“Inception-A 模块”的新模块，该模块由一组具有不同过滤器大小的并行卷积组成，随后是池化层和级联。这个模块的设计目的是在不同的尺度上捕捉更多不同的特征。

Inception v3: 于 2016 年提出，对 Inception 模块进行了一些改进，比如分解卷积，它将大的卷积分解成小的卷积，并引入了一种名为“Inception -residual block”的新型模块。该块将 Inception 模块与 ResNet 的剩余连接结合在一起，允许信息在网络中更自由地流动，并提高其捕获复杂特征的能力，还引入了“全局平均池化”，它将网络末端的全连接层替换为平均输出的池化层。

Inception v4: 这是 InceptionNet 的最新版本，于 2016 年推出。它建立在 Inception

v3 的设计基础上，引入了对 Inception 模块的几个改进，例如使用“split-transform-merge”(STM) 模块，通过将输入特征映射分割成更小的组，对每个组独立地应用卷积，然后合并结果，减少了网络所需的计算量；还引入了“跨阶段部分连接”，将网络的不同部分连接在一起，允许信息在它们之间更自由地流动，这有助于提高网络跨多个尺度捕获复杂特征的能力。

Inception-ResNet 的结合：Inception-ResNet 是 InceptionNet 的一个变体，它将 Inception 模块与 ResNet 的剩余连接结合起来。它在 InceptionV3 中被引入，并在 InceptionV4 中进一步改进。这两种架构的结合使网络既高效又强大，能够在多种尺度上捕捉复杂的特征。

InceptionNet 已被广泛用于各种计算机视觉任务，它的高效设计使它在资源有限的设备上很受欢迎，如手机和嵌入式系统。在本文的课题视频字幕生成网络中一般会使用 Inception-ResNet-v2 版本模型。

2.1.1.2 运动表征提取网络

视频相关任务与单纯的图像任务最为不同的一点就是它是三维数据，与长宽的二维数据相比多出一个时间维度，如何处理时间维度、更好的理解视频中的信息和变化也成为视频字幕生成的一个重大挑战。在视频表征中，这种基于时序的动态特征也被成为运动表征（Motion）。如图2-4所示，传统针对时序的方法利用循环时间网络和 LSTM 系列处理时序表征（如图2-4(a)），虽然却得了一定效果，但由于这样的框架使他相对三维卷积^[42,43] 具有几个缺陷：1）平行性：三维卷积网络具有高度的并行性，这意味着它们可以在 GPU 等并行计算硬件上进行训练，这可以大大加快训练时间。相比之下，RNNs 和 LSTMs 是顺序模型，依赖于网络的先前状态，这使得它们更难并行化。2）空间和时间特征：三维 CNN 通过在空间和时间维度上应用过滤器，可以同时捕获空间和时间特征。这使它们更适合于分析视频中的运动特征，这涉及到空间和时间的变化。3）内存要求：RNNs 和 LSTMs 可能是内存密集型的，因为它们要求网络保持关于以前状态的信息。相比之下，三维 CNN 不需要那么多内存来分析视频中的运动特征。4）对噪声的鲁棒性：三维 CNN 通常比 RNN 或 LSTM 对噪声更鲁棒，因为它们能够同时捕捉多种尺度的时空特征，这可以使它们更适合分析嘈杂的视频数据。因此三维卷积（如图2-4(b)）逐渐成为运动表征提取的主要范式。

I3D(Inflated 3D ConvNet)^[44] 和 C3D(卷积 3D)^[45] 是两种流行的视频识别和分析模型。

I3D 是膨胀 3D 卷积的缩写，是一种流行的 3D 卷积神经网络架构，由谷歌的研究人员于 2017 年推出。I3D 模型是流行的二维卷积神经网络架构 Inception V1

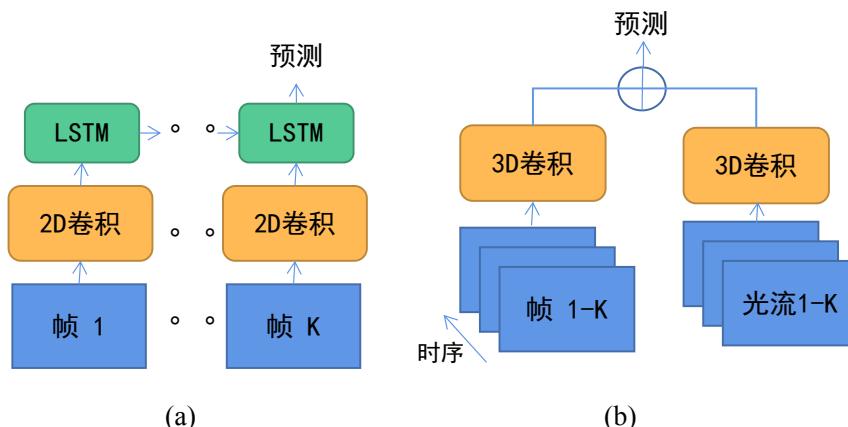


图 2-4 两种针对 3 维视频特征的范式。(a) 基于 LSTM 的视频处理; (b) 基于三维卷积的视频处理

的改进版本，它被扩展来操作视频等 3D 时空数据。I3D 背后的主要思想是在一个大型图像数据集上预训练 2D 卷积网络，然后使用预训练的 2D 卷积滤波器初始化 I3D 模型中 3D 卷积滤波器的权重。这个过程被称为“膨胀”，它允许 I3D 模型在相对较小的数据集上进行训练，而不会过度拟合。I3D 模型架构由几个 3D 卷积层块组成，每个块使用不同的过滤器大小和过滤器数量。该模型还包括最大池化层和批归一化层，以提高训练的稳定性和效率。I3D 的关键优势之一是它能够在视频中捕捉空间和时间信息。通过使用三维卷积滤波器，I3D 能够同时分析空间和时间特征，这对于许多视频分析任务是重要的，例如动作识别和视频分割。I3D 在许多视频分析基准上都取得了最先进的性能，包括 Kinetics 数据集，该数据集由 600 个动作类别的 60 多万视频剪辑组成。此外，I3D 已被证明可以有效地将知识从大型图像数据集的预训练转移到使用小型训练数据集的视频分析任务。

C3D 是最早为视频分类而设计的 3D 卷积神经网络架构之一，它是基于 2D CNN 架构 AlexNet，由 Facebook 和加州大学伯克利分校的研究人员在 2014 年推出。它将图像识别中的二维卷积滤波器扩展到三维空间，对视频数据进行时空处理。C3D 经过端到端训练，可以将视频分类为预定义的类别，例如动作或事件。C3D 模型架构由几层 3D 卷积过滤器组成，然后是最大池和全连接层。与 I3D 不同的是，I3D 使用预训练的二维卷积过滤器来初始化三维卷积过滤器，C3D 使用反向传播和随机梯度下降进行端到端的训练。C3D 的主要优势之一是它能够捕捉视频中的空间和时间信息，与 I3D 类似。然而，与 I3D ($1 \times 3 \times 3$, 3×1) 相比，C3D 使用更大的过滤器尺寸 ($3 \times 3 \times 3$)，这使它能够捕获更复杂的时空模式。C3D 已被证明对广泛的视频分析任务有效，包括动作识别、视频分割和物体检测。它还被用作其他视频分析任务的特征提取器，如视频字幕和视频检索。

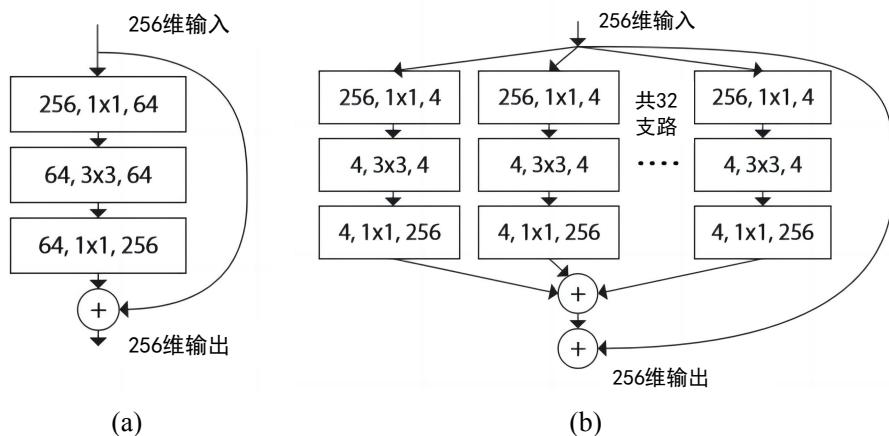


图 2-5 ResNeXt 基于 ResNet 的改进。(a)ResNet^[39] 示意图；(b)ResNeXt^[46] 示意图

现在的 3D 卷积也跟随了 2D 卷积的进步，现在最先进的模型已经被引入到了视频字幕生成领域的特征提取。下文将介绍 3D ResNeXt 模型，这一模型在现今视频字幕生成领域取得了很好的成果。

3D ResNeXt^[46] 是一种 3D 卷积神经网络架构，它基于 ResNeXt 架构，由 Facebook 的研究人员于 2018 年推出。ResNeXt 的主要思想是通过使用模块化方法构建深度网络来提高神经网络的表示能力，从而实现更好的并行化和泛化，如2-5所示。

三维 ResNeXt 架构扩展了 ResNeXt 方法，2D 概念被扩展到包括 3D 卷积，以从视频数据中学习空间-时间特征、操作三维时空数据。其贡献主要可以总结为瓶颈块和分组卷积。与 ResNeXt 类似，3D ResNeXt 采用模块化方法构建网络架构，每个模块由一组相同的“瓶颈”块组成。每个瓶颈块由三个卷积层组成，其中第一层降低输入的维度，第二层进行分组卷积操作，第三层恢复输出的维度。除了模块化方法外，3D ResNeXt 还使用了一个“基数”参数，控制分组卷积操作中的组数。ResNeXt 模型是基于使用多个平行的卷积路径的想法，使网络能够学习更多不同的特征，从而提高准确性。通过增加 cardinality，3D ResNeXt 可以在不大幅增加参数数量的情况下提高网络的表示能力。3D ResNeXt 的主要优势之一是它能够捕捉视频中的空间和时间信息，同时保持一个相对紧凑的网络结构。3D ResNeXt 能够在许多视频分析基准上实现最先进的性能，包括 Kinetics 数据集，对动作识别任务特别有效，在这些任务中，目标是将视频归入几个预定的动作或活动中。总的来说，3D ResNeXt 是一个强大的 3D 卷积神经网络架构，它建立在 ResNeXt 方法的基础上，可以操作 3D 时空数据，如视频。它的模块化方法和分组卷积操作允

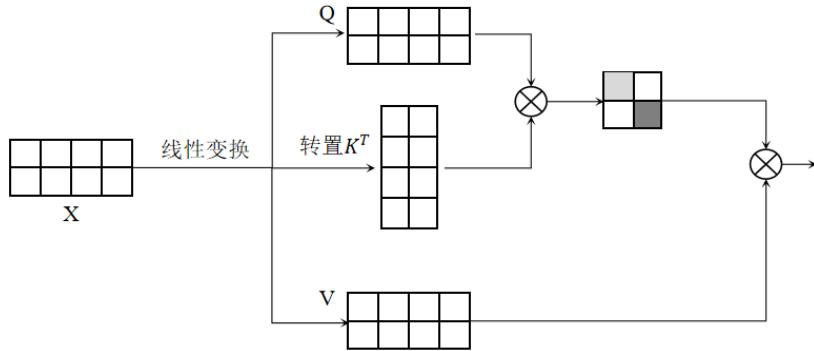


图 2-6 自注意力机制示意图

许更好的并行化和泛化，同时保持相对紧凑的网络架构。

2.1.2 注意力机制

在视频字幕生成领域，注意力机制^[47]是在信息理解和多模态融合中使用的一种技术，能够在处理输入序列时有选择地关注其某些部分。它首先被引入机器翻译等文本生成任务中，通过允许它们在生成输出序列时关注输入序列的相关部分来提高序列到序列模型的性能。如图2-6所示，注意力机制的工作原理是根据输入序列中的每个元素与当前语境的相关性来计算其权重。然后，这些权重被用来计算输入元素的加权和，作为当前处理步骤的语境向量。然后，这个上下文向量与输入序列和先前的隐藏状态一起被送入下一步的处理。给出一个 query 矩阵 \mathbf{q} ，一系列 key 矩阵 $\mathbf{K} = \mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_n$ ，和一系列 value 矩阵 $\mathbf{V} = \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ ，注意机制计算值向量的加权和，如下式（2-4）所示：

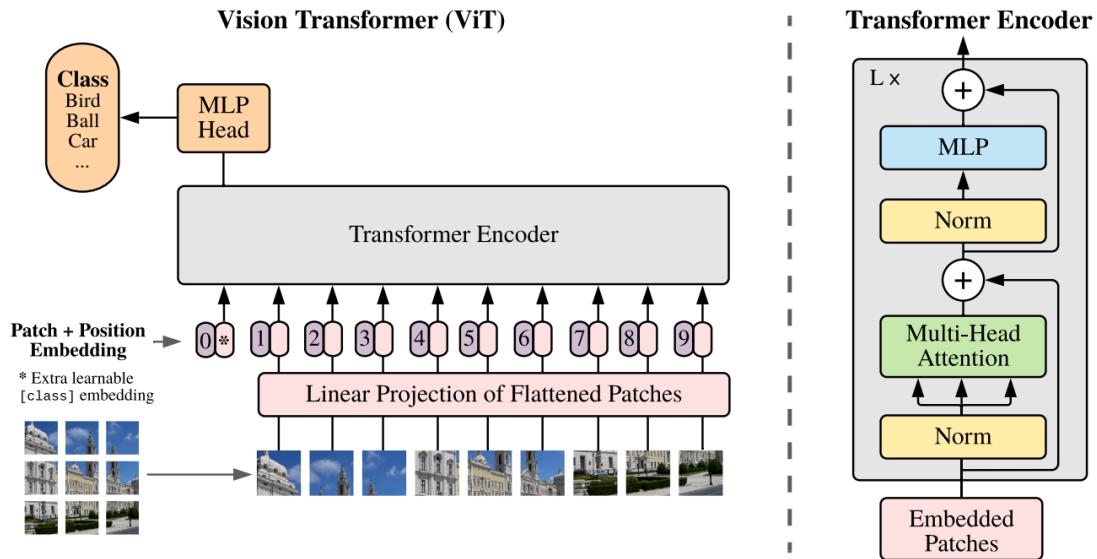
$$\text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \sum_{i=1}^n \alpha_i \mathbf{v}_i, \quad (2-4)$$

其中注意力权重 α 计算为查询向量和每个关键向量之间的点积的 softmax^[48]，按常数因子 $\frac{1}{\sqrt{d}}$ 缩放来避免梯度消失和梯度爆炸，如式（2-5）所示：

$$\alpha_i = \frac{\exp(\frac{1}{\sqrt{d}} \mathbf{q} \cdot \mathbf{k}_i)}{\sum_{j=1}^n \exp(\frac{1}{\sqrt{d}} \mathbf{q} \cdot \mathbf{k}_j)}, \quad (2-5)$$

其中 \cdot 代表点乘； d 表示 query 和 key 的维度。query 矩阵 \mathbf{q} 可以是表示输入的任何向量，而键和值向量通常是在不同的表示空间中表示输入的相同维度的向量（即 $\mathbf{K} = \mathbf{V}$ ）。

Transformer^[18] 是一种完全基于注意力机制的神经网络架构。它是在机器翻译的背景下引入的，后来成为各种自然语言处理任务的流行架构，包括语言建模、情

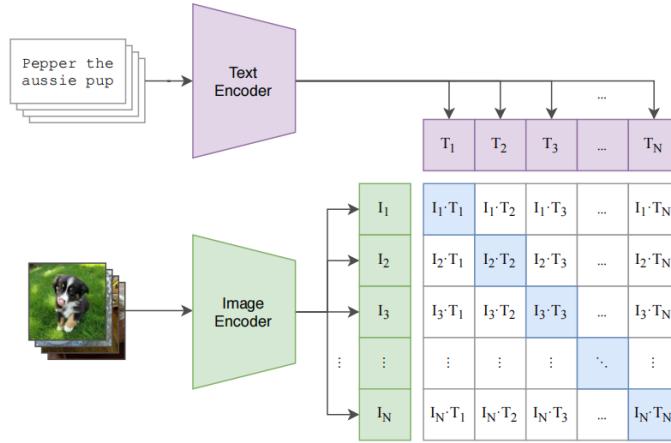
图 2-7 ViT 网络架构^[49]

感分析和问题回答。其强大的建模能力已经使它成为深度学习领域最为重要的推动力之一。Transformer 架构由一个编码器和一个解码器组成，每个编码器由多层自我注意和前馈神经网络组成。编码器中的自我注意层使其能够有选择地关注输入序列的不同部分，而解码器中的自我注意层使其能够有选择地关注输出序列的不同部分。前馈神经网络提供了注意力层之间隐藏状态的非线性转换。

与传统的递归神经网络架构相比，Transformer 架构有几个优点，包括更快的训练时间、并行处理序列的能力，以及更有效地处理长序列的能力。它已成为广泛的自然语言处理任务的流行架构，并已被用于许多最先进的模型中。

在 Transfomer 的基础之上，自然语言处理领域先一步进行了巨大的变革，大规模数据集以及端到端预训练成为当今深度学习发展的主流。视觉算法科研人员也由此引发思考，视觉相关任务是否可以摆脱两阶段的特征提取-下游解码器范式，实现端到端的，大规模的数据处理，由此，ViT^[49] 被提出并进入人们的视野。

ViT (Vision Transformer) 是一个深度学习模型架构，它使用自我注意机制和变换器进行图像识别任务。它在谷歌人工智能于 2020 年发表的一篇论文中被介绍，并因在几个图像分类基准上取得有竞争力的结果而获得关注。传统上，卷积神经网络 (CNN) 一直被作为是图像嵌入学习的基础框架，然而，如图2-7所示，ViT 采取了一种不同的方法，使用最初为自然语言处理任务设计的 Transformer 来处理图像数据。这是通过将图像视为一个补丁 (patch) 序列来实现的，然后将其压扁拉伸为一维序列并送入基于 Transformer 的网络。它由一个多层的变换器编码器组成，它类似于自然语言处理中使用的变换器架构的编码器组件。变换器编码器将

图 2-8 CLIP 网络架构^[50]

扁平化的图像补丁序列作为输入，并通过多个注意力层对其进行处理，每个注意力层都涉及计算所有补丁的自我注意力分数，以捕捉全局和局部的依赖关系。所得的特征表示然后被送入多层感知器以产生最终的分类结果。ViT 的主要优势之一是它可以在大规模数据集上进行端到端的训练，这使得它可以学习有效的视觉表征，而不需要在辅助任务上进行预训练。此外，ViT 被证明是高度可扩展的，使其能够处理大型图像，并在几个具有挑战性的图像分类基准上取得最先进的性能。

由 ViT 开始，基于注意力机制和 Transfomer，计算机视觉领域的特征提取以及训练范式被不断更新，继续涌现了更多的先进模型，例如 CLIP^[50]，SwinTransformer^[51]等，为下游任务如视频字幕生成的发展开拓了光明的前景并给予了有力的支撑。

CLIP^[50]（对比性语言-图像预训练）是一个由 OpenAI 开发的大规模跨模态图像-文本预训练模型，由 4 亿个通过社交媒体、新闻文章和网页搜集的图像-文本对组成的大规模数据集进行训练，这使它能够学习一套多样化的视觉和语言概念。如图2-8所示，CLIP 使用一个基于 Transfomer 的架构构建图像和文本编码器，进行特征嵌入，并通过最小化对比损失来学习关联图像和它们的描述，这鼓励模型给正确的图像-文本对分配更高的分数，并且通过赋予更低分数拉远负样图像-文本对之间的距离。通过其独立的图像编码器和文本编码器，CLIP 可以执行广泛的任务，包括图像分类、对象检测、图像说明和基于文本的图像检索，而不需要为每个任务进行专门训练。更重要的一点是，该架构可以将图像和文本编码到一个共享的嵌入空间，共享相近的语义，这使得该模型能够以一种无缝和有效的方式比较和匹配图像和文本，在零样本和少样本任务中体现了巨大的潜能。

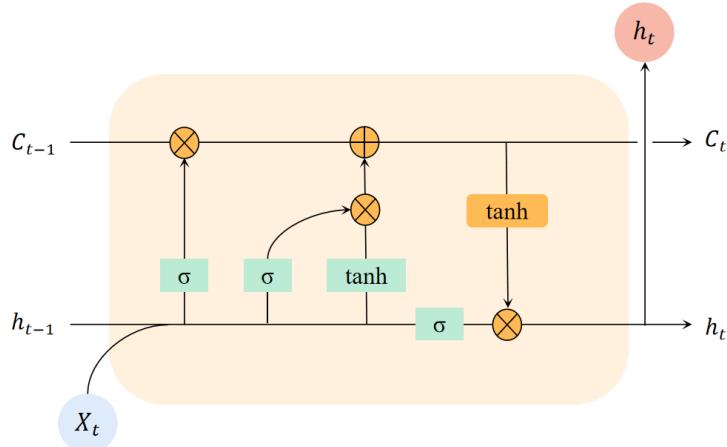


图 2-9 LSTM 网络结构示意图

2.1.3 循环神经网络

循环神经网络（RNN）^[52]是一种可以处理具有时间序列特征数据的神经网络，在视频字幕生成领域常被用于解码器设计。它被设计为保持对过去输入的“记忆”，并向网络添加反馈回路，允许将一个时间步的输出反馈到网络中，作为下一个时间步的输入，这使网络能够保持一个内部状态，以捕捉输入元素之间的长期依赖关系。在具体设计中，与处理固定长度输入的前馈神经网络不同，RNN 有一个隐藏状态，使其能够捕捉到连续数据的时间依赖性，隐藏状态在每个时间步中使用当前输入和前一个隐藏状态进行更新。RNN 模型主要靠隐藏层来进行数据传输。在 t 时刻的隐藏层状态可被表示为式 (2-6)：

$$\mathbf{h}_t = \sigma(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{b}_h), \quad (2-6)$$

其中 \mathbf{x}_t 是 t 时刻的输入； \mathbf{W}_{xh} 和 \mathbf{W}_{hh} 是输入和隐藏层的权重矩阵； \mathbf{b}_h 是偏差， σ 是激活函数。然而，标准的 RNN 可能遭受梯度消失的问题，即用于更新网络权重的梯度随着时间的推移变得非常小或为零，这使得它难以学习长期依赖关系。为了解决这个问题，长短期记忆（LSTM）网络^[53]被引入。

LSTM 网络是一种特殊的 RNN，如图2-9所示，它结合了专门的记忆单元和门控机制来选择性地记住或忘记以前的时间步骤的信息。每个 LSTM 单元都有一个记忆状态，可以通过三个门进行修改：一个输入门控制新信息加入记忆的程度，一个遗忘门控制旧信息从记忆中丢弃的程度，一个输出门控制记忆对输出的影响程度。LSTM 中的门控机制使它们能够在较长的时间范围内有选择地存储或检索信息，使它们在处理具有长期依赖性的数据序列时特别有效。由此，LSTM 能够解决传统 RNN 中梯度消失的问题：梯度消失会使网络难以学习连续数据中的长期依赖

关系，LSTM 使用门控机制来控制信息在网络中的流动，并根据当前的输入和之前的隐藏状态选择性地更新隐藏状态。具体来讲，更新门、输出门和遗忘门分别由以下（2-7）公式计算：

$$\begin{aligned}\mathbf{i}_t &= \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i), \\ \mathbf{f}_t &= \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f), \\ \mathbf{o}_t &= \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o),\end{aligned}\quad (2-7)$$

其中 \mathbf{x}_t 是 t 时刻的输入； \mathbf{h}_{t-1} 是之前的隐藏状态； \mathbf{W}_{xi} , \mathbf{W}_{hi} , \mathbf{b}_i , \mathbf{W}_{xf} , \mathbf{W}_{hf} , \mathbf{b}_f , \mathbf{W}_{xo} , \mathbf{W}_{ho} 和 \mathbf{b}_o 是分别的权重矩阵和偏差项； σ 是 sigmoid 激活函数。新的单元格状态可如下计算为如下式（2-8）：

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c), \quad (2-8)$$

其中 \mathbf{W}_{xc} , \mathbf{W}_{hc} 和 \mathbf{b}_c 是权重矩阵和偏差项。最终的单元格状态和隐藏状态计算如下式（2-9）：

$$\begin{aligned}\mathbf{c}_t &= \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{c}}_t, \\ \mathbf{h}_t &= \mathbf{o}_t \circ \tanh(\mathbf{c}_t),\end{aligned}\quad (2-9)$$

其中 \circ 代表点乘操作。

尽管有其局限性，RNN 和 LSTM 在深度学习的发展中发挥了重要作用，并成为机器学习领域许多重要应用的基础。近年来，人们对 RNN 和 LSTM 重新产生了兴趣，研究人员探索新的方法来提高它们的性能并扩展其能力。在视频字幕生成领域，循环神经网络也仍继续其使命和发展。

2.2 本章小结

在本章，主要介绍了与视频字幕生成相关的多项经典、先进研究成果。首先针对视频特征提取介绍了视频特有的两类特征提取网络：外观表征提取网络和运动表征提取网络，详细介绍了其中较为突出的工作如 ResNet、InceptionNet, I3D、C3D 和 3D ResNeXt。接下来从模态处理出发，介绍了注意力机制以及基于此的 Transformer 模型及其最新进展（ViT、CLIP），这些模型展现了其在多模态领域及计算机视觉领域的优异表现和广阔前景。最后，针对视频字幕生成的解码器部分，详细介绍了循环神经网络的基础和数据传输过程，概述了从最早的 RNN 出发到 LSTM 门控机制的变化。这些网络都构成了视频字幕生成的基础，为视频字幕生成的进一步发展提供了有力保障。

第三章 基于支持集视觉表达增强的视频字幕生成

3.1 动机

视频字幕生成的目的是为视频自动生成丰富生动、准确扼要的语言描述。与图像字幕生成任务相比，由于视频包含更多的时空信息、更多的细节和更丰富的语义，需要的视觉知识更加广阔，视频字幕生成也更为困难。面对视频这一表达形式，如何正确处理这些丰富的潜在语义，挖掘其中蕴含的信息和知识，提升语义表示的水平，是本课题面临的重大挑战。

为了获得更优秀的语义表达，已经有丰富的工作从不同角度进行了尝试。部分工作通过引入网格特征获取更多的空间位置信息，并在此基础上进一步进行构建帧内的消息交互^[26,27]：这些网格特征实际上是更细粒度的视频表示，使得特征包含了更多粒度的表达，再结合注意力机制的参与，捕捉到更多的视频细节。还有部分工作是通过引入一个预训练的目标检测器，获取视频中的目标物体信息的指导从而进行特征融合^[28,29]：通过这样的方法，在帧内能够建立起更多依赖目标实体的联系，使得模型特征更加丰富和具象。然而，如图3-1左侧所示，这些典型的方法在训练阶段遵循了严格的从视频到标题的一对一映射，这种严格的映射过程使得参与训练的样本都专注于自己的样本空间，从而忽略了样本之间的内在语义联系（例如，在训练视频集中可能包含许多表达类似语义的材料）。这种一对一的映射会导致模型表达能力受限，并且没有充分利用已有数据中的内部知识，对内部信息的挖掘不足，如左侧中“输出”所示，即使这对视频在语义上高度相关，其对应的生成文本也仅限于自己样本的注释文本，缺乏了表达的生动性。

为了解决这一内部信息挖掘不充分的问题，本工作引入了一个“支持集”的概念。这一概念灵感来自跨模态检索等多模态表示学习任务^[54]，通过构造一组视觉的辅助信息并捕获内部细节的连接，以此参与到模型的训练和优化。本工作将此概念用于视频字幕生成领域，用于捕捉语义子空间中与某样本语义相关的视觉元素，构成支持集，并以此为基础，在生成过程中创建灵活的映射关系，从而让模型通过更丰富的映射，学到更生动多样的字幕表达。

示意图3-1展示了传统方法和本工作的方法在训练规则和结果方面的差异。GT表示对于样本的人工标注字幕。典型的方法遵循一对一映射，会获得基于自身样本空间的普通语句，而本工作的方法建立了更灵活的映射方法，从而会产生更丰富的语义表达式：如图3-1右侧所示，本工作的方法针对语义相关的视觉元素补充了更多样的映射关系，使得生成的句子能够从其他语义相关视频的标注中获得更



图 3-1 基于支持集的视觉表达增强动机示意图

丰富的文本表达。从图中可以看出，“切”操作产生了新的表达“切片”，与此同时，“切肉”元素衍生出了“在厨房”的更高层次语义。

本工作提出了一种基于支持集的视觉表达增强 (SMRE) 模型，这一模型利用样本间共享的语义子空间中的信息，增强了对内部信息的挖掘程度，通过多样的映射方式学到更丰富的表达。具体来讲，本工作的模型是基于一个典型的编码器-解码器框架，在此基础上，本工作通过支持集构建 (SC) 模块学习获取一个支持集，并将其馈送到与原始路径共享的编码器-解码器的新分支中，以构建对应关系，建立灵活的映射。值得注意的是，由于这一过程在训练阶段直接优化了模型的表达，在推断阶段不需要应用支持集。此外，为了在共享子空间中建模这一复杂的语义关系，保证支持集正确行使职责，本工作引入了语义空间转换 (SST) 模块，在多模态语义空间中约束各个表达的相对距离，并分别从模态间和模态内两个角度通过损失设计实现了自监督过程的表达约束。基于以上设计，编码器-解码器结构可以学习更好的语义表示，并生成语义丰富的字幕。综上所述，本工作的贡献有三点：

- 1) 与传统的一对一映射相比，本工作通过学习构建支持集，应用一种新颖灵活的映射框架 (SMRE)，从而捕捉样本内部细节的联系，进一步挖掘样本内部知识；

2) 为了进一步约束语义关系, 本工作从模态间和模态内两个角度提出了自监督过程的语义空间转换 (SST) 模块;

3) 在 MSVD 和 MSR-VTT 两个基准数据集上, 本工作的达到了最先进的性能。针对 MSVD 数据集, SMRE 模型比 SOTA 在 BLEU-4 和 CIDEr 上分别高出 5.3% 和 1.1%, 在 MSR-VTT 上分别高出了 1.5% 和 0.4%, 证明了本工作的有效性和先进性。

3.2 基于支持集的视觉表达增强

在本节将介绍一个基于支持集的视觉表示增强方法 (SMRE), 以挖掘共享语义子空间中的内部信息, 通过建立灵活的映射关系, 得到表示增强和更好的语义。图3-2描述了 SMRE 的基础框架。首先, 学习构造支持集并学习内部知识, 再在原有编码器-解码器的基础上, 加入一个新的支持集输入分支, 并且这两条路径共享相同的编码器-解码器参数, 建立新的映射关系。在编码器部分, 本工作添加了支持集构建 (SC) 模块 (章节 3.2.2) 和语义空间转换 (SST) 模块 (章节 3.2.3), 并且, 在该部分使用了两种对比学习方法: 用于模态间交互的三元组损失^[55]以及模态内交互的对比学习损失^[56]。解码器部分, 保留了典型的双层结构: 注意力-语言 LSTM 结构, 并使用交叉熵损失来约束整个生成过程。以下将进行方法的详细阐述。

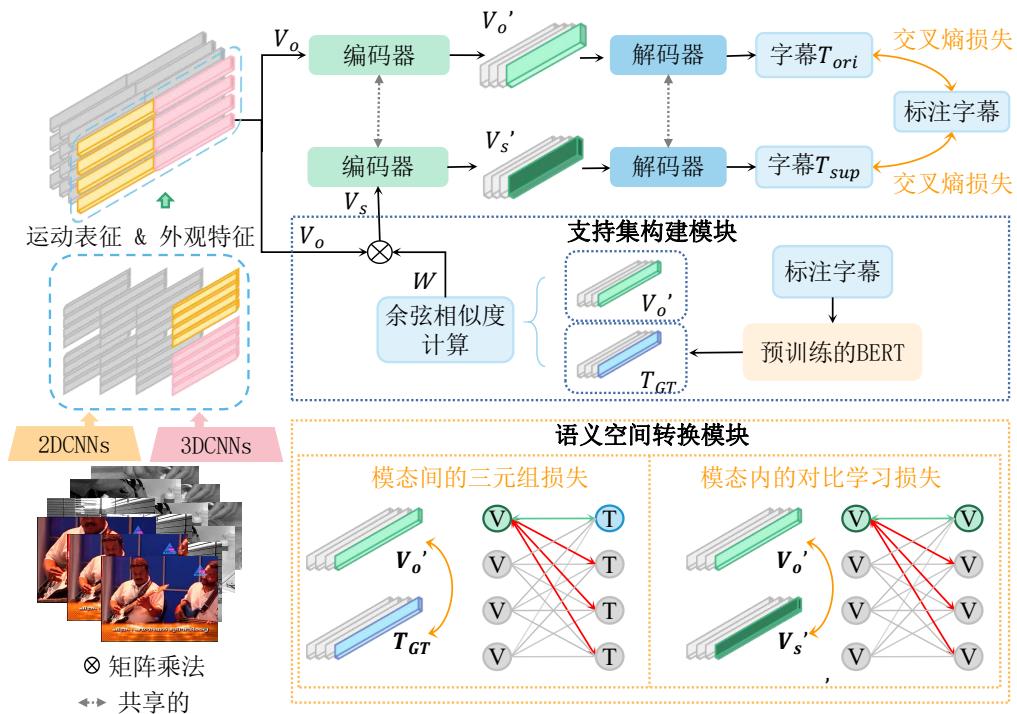


图 3-2 基于支持集的视觉表达增强的框架示意图

3.2.1 模型框架设计

模型框架部分，本方法针对内部信息挖掘不充分，会导致生成文本受限不生动的问题，优化了传统的一对一映射的范式：通过学习构造支持集捕捉到与样本相关的视觉元素，将其作为一个新的输入支路映射到该样本的文本，经这一表达增强过程形成具有灵活映射关系的新型范式。

传统的范式将一个视频分别经过二维和三维的特征提取，得到运动表征和外观表征并进行合并，将得到的部分做为视觉的完整表征，首先输入到编码器进行一定的空间变化，再进入解码器生成句子，在此过程中通过交叉熵损失的约束实现了一对一的映射关系。

本方法基于此范式进行了扩展。如图3-2所示，在单流网络的基础上加入了另外一条输入通路，通过支持集构建模块构建得到能够捕捉语义相关的视觉元素支持集，将其作为附加的视觉特征送入共享的编码器-解码器支路，得到一个双流的网络框架。这一支路经交叉熵损失的约束，能够基于新建立的映射关系使模型学到更丰富的语义表达，由此得到表达的增强。

具体而言，在编码器模块中，支持集构建模块首先学习构造一个支持集，并用新的分支将支持集映射到真实的标注字幕，在此过程中，应用语义空间转换模块中的两种对比学习方法进行模态间和模态内的交互，捕捉共享多模态语义子空间中的微妙语义关系，学习挖掘已有样本之间的内部知识。在解码器模块中，将沿用经典的双层 LSTM 对视觉特征进行解码，生成相关文本描述。

模型设计以及训练的具体细节将在下文中进行展示。

3.2.2 支持集构建

出于对样本内部知识挖掘的需要，本文提出学习构建支持集来捕捉样本内部相似的视觉元素，并将其组织起来构成支持集，通过新增的输入通路建立灵活多样的映射关系。在本节将介绍支持集构建的出发点和具体实施细节。

首先，出于对支持集的需求，本文认为支持集应符合以下两个基本观点：1) 支持集应该拥有语义丰富的视觉信息，这些视觉信息需要能够与原始样本相关，表达相似的视觉元素，因此可以从建立的新映射中对原本的学习过程进行补充和表达增强；2) 支持集需要从现有资源中获取，深入挖掘数据集内部的细节联系和视觉知识，无需额外视觉数据或是视觉特征提取器作为补充。

由此，为了使支持集获取符合上文描述的精确的语义关系，该模块在支持集构建过程中应用了相应的文本标注作为特征构建的学习指导，通过对文本特征以及视觉特征表达相似度的衡量，会获取到这一批次样本视觉特征中与查询样本视

觉相近似的视觉元素，并通过提取该元素将其组织成新的支持集输入，使其能够通过共享的编码器-解码器，学习到更丰富的内部知识，优化表达空间，具体细节如下。

由 B 表示训练时每个批次内样本的数量，在支持集构建模块内，首先计算原通路编码器的输出即 \mathbf{V}_o' 与该批次对应的文本标注的嵌入 \mathbf{T}_{GT} 之间的余弦相似度 $\mathbf{S} \in \mathbb{R}^{B \times B}$ ，这一过程中文本的嵌入使用了预训练的语言 BERT 模型^[57]，表示如下式 (3-1)：

$$\mathbf{S} = \text{cosine_similarity}(\mathbf{T}_{GT}, \mathbf{V}_o'). \quad (3-1)$$

下一步，如式 (3-2)，经 softmax 得到结果后，将这一矩阵作为视觉-文本对之间的权重参数 $\mathbf{W} \in \mathbb{R}^{B \times B}$ ：

$$\mathbf{W} = \text{softmax}(\theta_{scale}\mathbf{S}), \quad (3-2)$$

其中 θ_{scale} 是一个控制范围的常量，能够使得到的数据更加分散，区分度更高，在实际操作过程中，本工作将其取 100。然后，将 \mathbf{W} 与原视频特征输入 \mathbf{V}_o 相乘，得到包含新建视觉输入的支持集 \mathbf{V}_s ，这一支持集可以被认为是输入的视觉和语言对相似性的加权和。这一整体流程可以由公式表示如式 (3-3)：

$$\mathbf{V}_s = \mathbf{W} \otimes \mathbf{V}_o. \quad (3-3)$$

在根据上述操作得到支持集之后，将其 \mathbf{V}_s 送入到与原有通路共享的同一编码器中（在本方案仅引入简单的线性层作为空间变换），获得基于新通路的支持集中间状态 \mathbf{V}_s' 。

3.2.3 语义空间转换

通过 3.2.2 支持集的构建和双流网络的框架设计，能够在原有框架的基础上建立起丰富灵活的映射关系，但是，出于语义空间的复杂性，对于支持集的学习和网络的优化仍需要进一步约束的支持。如图3-3所示，在原本的样本空间，每个样本的语义信息较为分割，虽然在组成上含有相似的关联语义，但由于一对一的映射关系难以被模型捕捉，因此不能被模型学习到其内部联系，对于内部信息的挖掘力度不足。如3-3所示，进行丰富的映射关系之后，希望使得在语义空间距离较为接近的语义，如“肉”和“冻牛肉”，“切片”、“切”和“刀”，“鸡肉”和烹饪等关联能够被建立，帮助模型具有相关语义的拓展能力，使得生成的视频字幕能够具有更加丰富的表达。

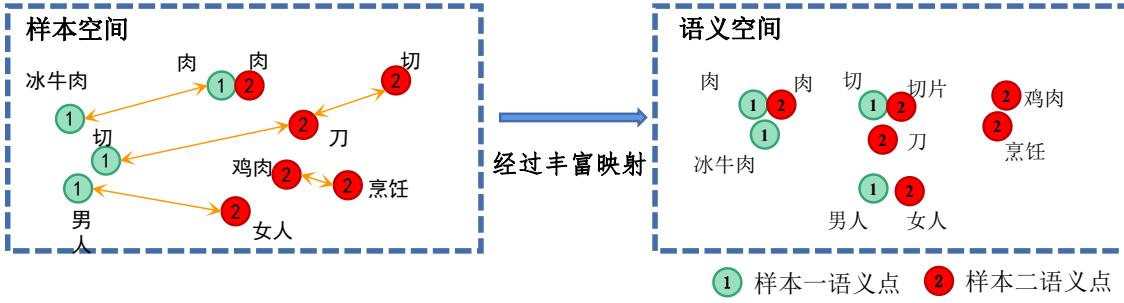


图 3-3 使用丰富映射后的语义空间示意图

为实现这一过程，本工作设计了语义空间转换模块，从模态内和模态间角度共同探索，进一步约束双流网络中语义空间关系，帮助编码器-解码器的表达增强。

3.2.3.1 模态间的交互

对于模态间的相互作用关系，本工作采用了最困难负样本的三元组损失函数^[58]。为了能够使约束过程对编码器的学习提供正反馈，本工作将这个损失应用于编码器输出的视觉嵌入 \mathbf{V}_o' 和经过预训练模型 BERT 得到的文本嵌入 \mathbf{T}_{GT} 之间，来拉近这两种形式之间的语义空间距离。

最困难负样本的三元组损失关注训练中的最难负样本来进行正负样本对之间的空间约束。让 $\mathbf{P} = (\mathbf{v}_i, \mathbf{t}_i)_{i=1}^B$ 代表一组视觉-文本对， $s(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$ 是 \mathbf{a} 和 \mathbf{b} 的余弦相似度。如果有一组正样本对 (\mathbf{v}, \mathbf{t}) ，那么在样本批次中对于视觉和文本的最难负样本即为 $\mathbf{v}' = argmax_{m \neq v} s(\mathbf{m}, \mathbf{t})$ 和 $\mathbf{t}' = argmax_{n \neq t} s(\mathbf{v}, \mathbf{n})$ 。

遵循三元组损失的基本设定，模态间交互的损失设计可以表示为如下式 (3-4)：

$$\begin{aligned} \mathcal{L}_{inter} = & \max_{\mathbf{t}'} [\alpha + s(\mathbf{v}, \mathbf{t}') - s(\mathbf{v}, \mathbf{t})]_+ \\ & + \max_{\mathbf{v}'} [\alpha + s(\mathbf{v}', \mathbf{t}) - s(\mathbf{v}, \mathbf{t})]_+, \end{aligned} \quad (3-4)$$

其中 $[\cdot]_+ = max(0, \cdot)$ 是零截断操作；边界值 α 被设定为 0.2。

由此，通过该模块，可以减小正样本视觉-文本对之间的语义距离，与此同时扩大负样本视觉-文本对之间的语义距离，直到这一距离到达边界值 α 。采用这样的设计，通过约束编码器输出的中间状态的 \mathbf{V}_o' ，能够将学到的空间信息反馈到编码器的学习之中，缩小视觉特征空间和文本特征空间之间的语义鸿沟，让编码器在空间变换中学到更好的视觉表达。

3.2.3.2 模态内的交互

对于模态内的交互关系，本工作采用对比学习损失来捕捉编码器输出两个中间状态 \mathbf{v}_o' 和 \mathbf{v}_s' 之间的关系，并且用此约束保持未处理视频集和支持集之间的不同，使得支持集能够充分发挥其作用。

理论上，在基于支持集的双流模型框架中，如果给出的约束仅仅是上文模态间的约束，通过构造正负样本对让自身文本-视觉对的空间表达相近，推远不同样本之间的文本-视觉对，试图让 \mathbf{v}_o' 和 \mathbf{T}_{GT} 无限的相似，那么，编码器将无限缩小这两个特征之间的距离，这显然与构建支持集作为视觉补充的出发点造成了偏差，会导致得到的支持集失真。

直觉上，本工作期望达到： \mathbf{v}_o' 和 \mathbf{T}_{GT} 在共享语义子空间中足够相似，以允许支持集这一附加分支能够经过文本的指导反映相似的语义含义。但与此同时，作为新的视觉输入，本工作希望支持集能够在向量的表现上具有独特性，不应与原本的视频集合近乎一致，从而才能为模型的学习提供额外的视觉补充指导。

因此，在模态内加入约束，引入了对比学习损失来约束语义表达。设定 $D = 1 - s(\mathbf{v}_o', \mathbf{v}_s')$ 为 \mathbf{v}_o' 和 \mathbf{v}_s' 两个中间状态之间的余弦距离。该对比学习损失 \mathcal{L}_{intra} 的计算函数可以被表示如下式（3-5）：

$$\begin{aligned}\mathcal{L}_{intra} &= (1 - Y)D^2 + Y\{\max(0, m - D)\}^2, \\ D &= 1 - s(\mathbf{v}_o', \mathbf{v}_s'),\end{aligned}\tag{3-5}$$

其中 m 代表损失 \mathcal{L}_{intra} 的边界值，在本文中被设定为 0.2；控制信号表示为 Y ， $1 - Y$ 约束了正样本对之间的距离， Y 约束了负样本对之间的距离。 Y 越接近于 1，该对比学习公式更倾向于关注负样本之间的语义空间距离。

具体来讲，在本工作的设置中，倾向于设置控制信号 Y 近似于 1.0，这是由于本工作的支持集来自于一个小批量中原始视频表达矩阵根据相似度计算的加权和， \mathbf{V}_s 从构造过程来分析本就与 \mathbf{V}_o 有非常强烈的联系，在共享的语义空间中距离非常相近。与其更多关注正样本之间本就非常相近的语义距离，把他们拉得更近，更应该以负样本为重点，使其更具可区分性和补充能力，以确保支持集能够真实有效地优化编码器参数。

通过模态内和模态间语义空间距离的约束，对支持集的构建和编码器-解码器的优化都提供了更具体的学习方向，优化了视觉的表达。

3.2.3.3 解码器组成

本工作的解码器遵循来自经典网络的双层 LSTM 的分层结构^[59]，分别包括一个注意力 LSTM 和一个语言 LSTM，这一模块整体在文中被表示为“deLSTM”。在 deLSTM 中，在每个时间步，通过对编码器得到中间状态和上一个生成单词的解码，进一步计算 softmax 后的最终单词概率，生成相应的单词。从两个分支最终分别生成的视频字幕预测可以表示为 \mathbf{T}_{ori} 和 \mathbf{T}_{sup} 。

公式简化如下式 (3-6) :

$$\begin{aligned}\mathbf{T}_{ori} &= deLSTM(\mathbf{V}_o'), \\ \mathbf{T}_{sup} &= deLSTM(\mathbf{V}_s').\end{aligned}\quad (3-6)$$

与其他文本生成工作一样，本工作使用交叉熵损失来确保整个模型的正确推理过程。本文中基于支持集的视频字幕生成的损失函数可以定义如式 (3-7) :

$$\begin{aligned}\mathcal{L}_{ori_cap} &= CrossEntropyLoss(\mathbf{T}_{ori}, \mathbf{T}_{GT}), \\ \mathcal{L}_{sup_cap} &= CrossEntropyLoss(\mathbf{T}_{sup}, \mathbf{T}_{GT}).\end{aligned}\quad (3-7)$$

综上所述，总体损失包括分别限制两个通路的两个交叉熵损失、约束模态内交互的对比学习损失以及约束模态间交互的三元组损失。可以表示为如下的公式 (3-8) :

$$\mathcal{L}_{overall} = \lambda_1 \mathcal{L}_{inter} + \lambda_2 \mathcal{L}_{intra} + \lambda_3 \mathcal{L}_{sup_cap} + \mathcal{L}_{ori_cap}, \quad (3-8)$$

其中 λ_1 , λ_2 , λ_3 是三个可以调整的超参数，分别代表了三部分损失在总体损失中占据的重要性权重。在本工作的实验中，根据训练后得到的数据量级，一般将 λ_1 和 λ_2 设置为 50，将 λ_3 设置为 0.5。

3.3 实验结果

3.3.1 数据集

MSVD ^[60] 由微软研究院和德克萨斯大学奥斯汀分校提出，它由 1970 个来自 Youtube 的短视频片段组成。每个视频片段平均包含大约 41 个英文描述。其描述较为简单，数据集标注的文本平均在 10-12 词。根据之前的研究的惯例，在训练中基于交叉验证^[61]，将数据集分为三个子组：将其中的 1200 个样本用于训练，100 个样本用于验证阶段，最终的 670 个样本被用于测试。

MSR-VTT ^[62] 是用于开放域视频字幕的大规模数据集，其通过商业视频搜索引擎，汇总了 257 个热门查询并进行收集来实现数据获取，其中每个查询包含了

118 个视频，共计 41.2 小时。这一数据集涵盖的类别非常全面，并在标注语句的复杂度、语义内容丰富度和词汇量方面占据较大的优势，是文本生成领域最广泛使用的数据集之一。它包括来自 20 个类别的 10000 个视频剪辑，每个视频段由标注人员配有 20 个英语句子，共 20 万标注文本。在本文中使用标准的数据集分割方式，将其中的 6513 个样本用于训练，497 个用于验证，余下的 2990 个样本构成测试集。

3.3.2 评价指标

本工作使用常用的自动评估指标来评估生成的字幕的质量，即 BLEU-4^[63], METEOR^[64], CIDEr^[65], ROUGE-L^[66]。分数越高，说明图片质量越好。以下详细介绍这些评价指标的不同点。

BLEU-4: BLEU 是机器翻译中使用的一种评估指标，用来衡量机器生成的标题和一个或多个参考标题之间的相似性的指标。它基于机器生成的标题和参考标题之间的不同粒度（在这里被称为 n-gram）的重叠程度。BLEU-4 指标的取值范围为 0 到 1 之间，当为 1 时表示两个对比项完全相似。BLEU-1、BLEU-2、BLEU-3 和 BLEU-4 分别指以长度 n 为 1、2、3 和 4 个单词的短语作为粒度的指标。具体的准确率的计算方法是计算每个 n-gram 在标准翻译中出现的次数，然后除以划分粒度的总数。较小的 n 值可以更好的捕捉翻译的准确程度，较大的 n 值可以衡量生成句子是否流畅。然而，BLEU 也存在缺点：无法应对某些特殊情况，如仅有一个恰在参考句子里的单词，那么仅仅这一个单词的生成句子就会获得满分的指标，这显然是不合理的。为了解决这个问题，BLEU 包括一个长度惩罚因子，对太短的译文进行惩罚。但一般来说，BLEU 这一指标仍倾向于短的生成文本，短的译文仍能得到较高的分数，这显然并不是字幕生成任务所希望的，故要结合其他方法进行衡量。

METEOR: 该方法提出的出发点在于，在文本生成任务中，有时候模型生产的结果是正确的，但仅仅因为与参考的标注信息不匹配（例如使用了一个同义词），而使得性能下降很多，这显然不符合逻辑。为了解决这个问题，该指标提出使用外部知识源来构造扩展了一个同义词的集合，并同时考虑单词的形态变换（例如，动词有过去式和第三人称单数，与其他完全不相关的词语相比，这些词本质上还是与正确词语是一致的语义，不该完全视为生成错误）。此外在评估句子的流畅度时，引入 chunk 的概念（将能够对齐的连续单词形成的序列称为 chunk），并使用启发式算法 beam search 对齐生成文本和参考文本。最后，同时考虑召回率和准确率，使用 F 分数作为最终评价指标。

CIDEr: CIDEr 结合了 BLEU 系列和向量空间的思想，它将每个句子视为一个文档，并采用 TF-IDF 的思想计算其向量的余弦夹角，用这种方法衡量生成句子与标注句子的相似程度。该方法计算生成的标题和一组参考标题之间的共识，同时也考虑到生成的标题的多样性。CIDEr 的工作方式是首先计算生成的标题和参考标题之间的 n-gram 重叠度，然后，它通过比较每个 n-gram 在生成的标题中的频率和它在参考标题中的频率，计算出一个共识分数。这个共识分数反映了生成的字幕在多大程度上捕捉了参考字幕的意义和内容。除了共识分数，CIDEr 还计算了一个多样性分数，以衡量生成的标题的多样性。多样性得分反映了所生成的标题彼此不同的程度，并有助于确保机器生成的标题不会过度冗余。依靠以上两个分数设计，能够实现对生成句子质量的综合评估。

ROUGE-L: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) 是自然语言处理中用来评估文本总结和机器生成的文本质量的一组指标。它最初是由 Chin-Yew Lin 在 2004 年提出的，此后在研究界被广泛使用。ROUGE 测量机器生成的摘要或文本与一个或多个参考摘要或文本之间的相似度，由召回率衡量。它通过计算机器生成的文本和参考文本中的 n-grams (连续的单词序列) 之间的重叠来实现这一目的。最常用的 ROUGE 变体是 ROUGE-N 和 ROUGE-L。ROUGE-N 测量机器生成的文本和参考文本中特定长度的 n-grams 之间的重叠程度。N 最常见的值是 2，但也可以设置为 1、3 或其他值。ROUGE-L 测量机器生成的文本和参考文本之间最长的共同子序列 (LCS)，即在机器生成的文本和参考文本中以相同顺序出现的最长的单词序列。ROUGE-L 对于评估摘要或不一定是参考文本的逐字复制的文本的质量非常有用。在视频字幕生成领域，通常使用 ROUGE-L 作为评价指标。与 BLEU 系列不同的是，由于其统计召回率的逻辑，该指标对长文本是更有力的。

3.3.3 执行细节

在本工作的实验中，使用 InceptionResNetV2(IRV2)^[41] 来提取外观特征，使用 I3D^[44] 作为 3D 卷积网络部分来提取运动特征。为了更好地与最新的工作进行比较，本工作还使用了另一组特征对：ResNet-101^[39] 提取外观特征，3D-ResNext-101 来提取动作特征。本工作将每个视频都平均地截取 26 帧，使用取帧平均方法来获得视频的全局特征。LSTM 模型隐藏层维度为 1024。

本工作采用了 Teacher-Enforced Learning(TEL)^[59] 方法来在每个训练步骤中，提示字幕生成模型以一定的概率学习标注中的单词，从而使训练更为快速。本文模型使用了 Adam 优化器，初始学习率被设定为 1e-4 来保障文中使用的三元组损

表 3-1 在 MSVD 和 MSR-VTT 数据集上的对比实验结果。IRv2、Iv3、Res 和 3D-r 分别代表 Inception-ResNet-V2、Inception-V3、ResNet 和 3D ResNeXt-101 的基础框架。此外，B@4, M, R, C 表示 BLEU-4, METEOR, ROUGE-L 和 CIDEr 四种评价指标。表中的 O/G 表示使用额外的物体/网格特征,w/o 代表未引入上述两种特征。

Method	Feature	O/G	MSVD				MSR-VTT			
			B@4	M	R	C	B@4	M	R	C
MGSA ^[26]	IRv2+C3D	G	53.4	35.0		86.7	42.4	27.6		47.5
SAAT ^[29]	IRv2+C3D	O	46.5	33.5	69.4	81.0	40.5	28.2	60.9	49.1
RMN ^[30]	IRv2+I3D	O	54.6	36.5	73.4	94.4	42.5	28.4	61.6	49.6
MGRMP ^[27]	IRv2+C3D	G	53.2	35.4	73.5	90.7	42.1	28.8	61.4	50.1
M ³ ^[19]	VGG+C3D	w/o	51.8	32.5			38.1	26.6		
GRU-EVE ^[21]	IRv2+C3D	w/o	47.9	35.0	71.5	78.1	38.3	28.4	60.7	48.1
POS-CG ^[22]	IRv2+I3D	w/o	52.5	34.1	71.3	88.7	42.0	28.2	61.6	48.7
Ours	IRv2+I3D	w/o	55.4	35.3	72.6	92.9	42.2	28.1	61.3	49.5
MARN ^[20]	Res-101+3D-R	w/o	48.6	35.1	71.9	92.2	40.4	28.1	60.7	47.1
MDT ^[23]	Res-101+3D-R	w/o	49.0	35.3	72.2	92.5	40.2	28.2	61.1	47.3
SGN ^[24]	Res-101+3D-R	w/o	52.8	35.5	72.9	94.3	40.8	28.3	60.8	49.5
Ours	Res-101+3D-R	w/o	55.5	35.6	72.6	95.2	41.4	28.1	61.4	49.7

失能够顺利收敛。本文训练使用的词汇表由至少出现两次的单词构成，这能够在一定程度上保证模型更好地学习和生成字幕。整体训练过程为 20 个 epoch。另外，beam search 方法在模型推理阶段被应用，同时，beam search 中使用的束大小在 MSR-VTT 中被设置成 2，在 MSVD 中被设置成 5 以更好的匹配数据集特点。

3.3.4 定量结果对比

如表3-1所示，表中对比了本工作提出的模型和其他最先进模型的性能，展示了基于数据集 MSVD 和 MSR-VTT 的对比实验结果。为了进行公平的比较，本工作没有将提出模型与利用目标检测器或网格特征的方法进行比较，因为这些模型引入了更多的特征。表3-1展示了对比实验的结果。对于这两个数据集，SMRE 超过了大多数具有基础框架的最先进方法，特别是在 BLEU 和 CIDEr 指标上。POS-CG^[22] 和 SGN^[24] 在该部分由于是具有与本工作相同特征的最新工作被选为对照组。

与 POS-CG 相比，本工作的模型有明显的改进：在 MSVD 数据及上上 BLEU-4

指标提升了 5.5%，CIDEr 提升了 4.7%；在 MSR-VTT 数据集上 BLEU-4 提升了 0.7%，CIDEr 提升了 1.8%。与 SGN 模型相比，本工作的模型在 MSVD 上的 BLEU-4 和 CIDEr 指标分别高出 5.3% 和 1.1%；在 MSR-VTT 数据集上的 BLEU-4 高出 1.5%，在 CIDEr 指标上高出 0.4%。

3.3.5 消融实验

在本节中，本工作进行了消融实验，以调查提出模型中每个设计对 MSVD 数据集上性能提升的贡献。在本部分采用特征对 IRv2 和 I3D 进行实验。

3.3.5.1 每个模块的作用

表3-2展示了本工作提出模型中每个模块的作用。实验使用了 5 种不同配置：

- (1) 仅有一条编码器-解码器的基准模型，体现基于此模型框架和特征抽取模块的基准性能；
- (2) 引入支持集和 $\mathcal{L}_{sup} = \mathcal{L}_{ori_cap} + \mathcal{L}_{sup_cap}$ 两个交叉熵损失函数的双流模型，体现不受约束的支持集展现的性能结果；
- (3) 引入支持集合和模态间语义约束的损失函数 \mathcal{L}_{inter} ，体现仅有模态间约束的支持集结果；
- (4) 引入支持集合和模态内语义约束的损失函数 \mathcal{L}_{intra} ，体现仅有模态内约束的支持集结果；
- (5) 具备 4 种损失设计的完整模型，体现 SMRE 框架表达增强的性能。

表 3-2 关于各个模块作用的消融实验

Function	MSVD			
	B@4	M	R	C
baseline	52.3	34.4	71.0	88.6
\mathcal{L}_{sup}	51.2	34.6	71.4	89.0
$+ \mathcal{L}_{inter}$	55.4	35.1	72.5	90.4
$+ \mathcal{L}_{intra}$	53.8	35.3	72.0	88.7
$\mathcal{L}_{overall}$	55.4	35.3	72.6	92.9

从表3-2中，可以看到结构的每一部分都是不可或缺的。特别是，与基线模型(1)相比，(2)的 \mathcal{L}_{sup} 取得了一定的提升，这说明了支持集可以获得额外的补充信息，使得建立的灵活映射的双流框架能够为模型学到更好的表达。为了进一步约束原始视频特征和支持集之间的相对关系，本工作通过在支持集使用期间分别

从模态内和模态间两个角度引入的两种语义空间空间约束损失函数：三元组损失 \mathcal{L}_{inter} 和对比学习损失 \mathcal{L}_{intra} ，对 SMRE 框架中的表达语义进行约束，通过（3）和（4）可以看出，这一手段获取了一定的收益。此外，当将这三个部分添加的损失函数与基准模型相结合时，整体 SMRE 模型的表现是最好的，这表明：支持集应该同时平衡这些表达角度之间的相关性，在支持集的学习构建过程中增强模型表达。

因此，上述结果表明，支持集构建模块、模态内语义空间交互、模态间语义空间交互三部分的协作和约束可以更好的支持编码器-解码器结构，学习更好的语义表示。

3.3.5.2 \mathcal{L}_{intra} 中比例 Y 的影响分析

本工作进一步探讨了模态内部信息交互中控制信号 Y 的作用。结果如表3-3所示，从表中可以发现：对于 MSVD 数据集，当控制信号 Y 接近于 1.0 的时候模型性能会更好。

表 3-3 模态内损失中比例 Y 的影响分析

Y	MSVD			
	B@4	M	R	C
Y=0.0	53.2	35.0	72.2	90.6
Y=0.2	53.9	34.3	70.3	87.7
Y=0.5	53.0	34.4	70.8	89.2
Y=0.8	53.7	34.8	71.6	90.9
Y=1.0	55.4	35.3	72.6	92.9

正如前文语义空间变换部分所提到的，猜测原因如下：可以知道当控制信号 Y 比较小时， \mathcal{L}_{intra} 将更多的关注将正样本拉近的操作，当控制信号 Y 比较大时， \mathcal{L}_{intra} 将更多的关注将负样本拉远至一定阈值的操作。然而，支持集本质上具有与原始视频相似的天然属性，与其再关注本已经较为相似的部分，不如将权重偏向于将负样本之间的距离推远一些的部分，这样较大的 Y 由于其让构建生成的支持集与原有的视频输入在语义相近的前提下保留一定的差异性，能够让模型真正从支持集中学到不同的表达，实际上对于本工作预设的支持集构想会更有帮助。

由此，本工作对于 MSVD 数据集设定 $Y=1.0$ ，对 MSR-VTT 模型设定 $Y=0.8$ 。都是较为接近 1 的比例。

3.3.5.3 Beam Search 中束大小针对不同数据集的影响

Beam Search 是一种启发式的搜索算法，用于包括文本生成的自然语言处理任务。Beam Search 的目标是在给定一组候选序列的情况下，生成使整个序列的可能

性最大化的最有可能的词语序列，可以根据束大小将其生成过程形象化为一个不断扩展的多叉树。

Beam Search 算法从一个初始输入序列开始，通过基于语言模型预测下一个词，生成一组候选序列，并使用概率指标给每个候选序列打分，如序列的对数可能性或复杂度，进而选择分数最高的 k 个候选序列。其中 k 是束大小，这些前 k 个候选序列被称为“束”。Beam Search 通过在每一步考虑多个可能的候选词，而不是在每一步只选择最可能的词，从而允许生成更长、更连贯的文本。在字幕生成任务的推理阶段被广泛应用。

束大小是 Beam Search 中一个重要的参数，决定了搜索的宽度。但束大小并不遵从一个越大越好的规则，为一个文本生成任务选择适当的束大小取决于几个因素，如数据集的大小和复杂性，生成文本的长度和复杂性，以及文本质量和计算效率之间的理想平衡。由此，在该部分对 Beam Search 中束大小针对不同数据集的影响做出分析实验。

表 3-4 束大小在 MSVD 数据集上的影响

Beam Size	MSVD			
	B@4	M	R	C
2	52.7	35.2	72.7	91.9
3	53.5	35.2	72.4	92.0
4	54.5	35.4	72.5	92.5
5	55.4	35.4	72.7	92.9
6	55.9	35.5	72.7	92.6

表 3-5 束大小在 MSR-VTT 数据集上的影响

Beam Size	MSR-VTT			
	B@4	M	R	C
2	41.4	28.1	61.4	49.7
3	41.2	27.7	60.8	48.9
5	41.6	27.6	60.9	48.9

表3-4为束大小在 MSVD 数据集上的影响。在此数据集上本部分进行了当 Beam Size 为 2、3、4、5 和 6 时的实验，对结果分析如下：从表中可以看到当 Beam Size 逐渐变大时，两个指标 METEOR 和 ROUGE-L 波动较小，而字幕生成的两

一个重要指标 BLEU-4 和 CIDEr 均都有稳步提升，直到束大小为 6 时达到该设定下 BLEU-4 性能的巅峰，甚至有进一步提升的可能；对于 CIDEr 指标在束大小为 5 时达到巅峰，又在进一步变大时饱和甚至回落。猜测这一现象是由于 MSVD 数据集是一个相对较小的数据集，其文本构成也较为简单，在此基础上使用更大的束大小使得搜索空间的宽度变大，能够使该数据集的模型表达能够被更好的探索。综合计算效率和生成质量，在本文中针对 MSVD 数据集采用束大小为 5 的设定进行实验。

表3-5体现的是在另一数据集 MSR-VTT 上的实验结果。MSR-VTT 是一个较大的数据集，其数据量约为 MSVD 的五倍，且其视频内容和与其对应的标注文本都较 MSVD 更为复杂。在此基础上对该数据集进行了束大小为 2、3、5 的消融试验。从表中可以看到对于 MSR-VTT 数据集而言，整体的性能变动都并不显著，但当束大小为 2 时其性能基本达到了最优，之后随着束的增大，BLEU-4 出现了向上的波动但其他指标都出现了下滑。猜测原因主要在于该数据集的复杂程度较高，仅是对搜索宽度的增加很难对复杂的表达进行捕捉，其他对于模型内部学习的优化会较为有效。出于对计算效率的考虑，在 MSR-VTT 数据集上本工作采用了束大小为 2 的设定。

综合上述分析，可以看到在不同数据集上束大小对性能的影响是较为不同的，因此需要进行根据实际情况的分析设定

3.3.6 可视化展示

图3-4为本工作结果的可视化展示，如图中所示，首先可以看到上半部分样例：标注文本主要展示了“一个女人和两个男人正坐在椅子上交谈”这一事件，基础模型生成的主要内容为：“两个男人在谈论一些事情”，而经过支持集表达增强的本模型表达了“两个男人和一个女人坐在沙发上在电视节目中交谈”，其中“在电视节目”和“坐在沙发上”都为根据支持集灵活映射学到的新表达，体现了支持集能够带来的效果提升。

下半部分的标注文本主要展示了：“一个人正在金属碗里拌匀调料”和“一个人正在向料理中加入很多颜色”，基础模型生成的主要内容为：“一个人正在做饭”，而本工作的方法能够生成：“一个女人正在碗里拌匀材料”，其中“在碗里拌匀”则是学到的丰富表达。

3.3.7 存在的不足以及未来工作

虽然以上消融实验和可视化证明了模型的有效性，但从生成文本看仍然存在一定不足，也将成为未来的工作方向。



GT_1: a woman and two man are sitting on the chair and discussing about something

GT_2: two men are talking about something

Baseline: a man is talking to a woman

Ours: two men are **sitting on a couch** talking to a woman **on a tv show**



GT_1: a man mixes various spices **in a metal bowl**

GT_2: there is a man is adding colors to a dish

Baseline: a man is cooking

Ours: **a woman** is **mixing** some ingredients **in a bowl**

图 3-4 本工作生成字幕的可视化展示

例如图3-4中下半部分的实验结果中，通过支持集的灵活映射和内部知识的挖掘，将在“碗里拌匀材料”这一事件类似于“料理”的事件与“女性”结合，某种意义上继承了属于数据集内部知识的一定“偏见”(bias)，即使视频中并未有对料理人任何性别信息的展示，但综合内部信息仍然得到了基于这一事件对“女性”较高的预测概率。这说明对于数据本身内部的联系和知识需要有一定的约束手段，来自数据内部的“偏见”需要被处理，由此，未来的工作可以将其给予重视。

另外，从图3-5中可以看到，由于基于视觉输入和上一个单词嵌入的预测概率，有些生成句子里存在一定的重复语句，并不符合语言的逻辑。如图3-5上半部分：预测至“一个人正在展示怎样制作一小片”后面应该接入一个物体，但是概率预测使其生成又一个“一小片”。图3-5下半也展示了类似的问题，当“一个人正在唱歌”后接上了“并且”后，模型倾向于继续预测，并继续生成了“一个人正在唱歌”。这种的重复问题可以通过在生成模块加入一定的掩码机制进行解决，在未来工作也可探索其他解决方法。



Ours: a person is showing how to make a small piece of a small piece



Ours: a man is singing and a man is singing

图 3-5 句子产生重复部分的可视化展示

3.4 本章小结

在本章中，提出了一个基于支持集的视觉表达增强 (SMRE) 视频字幕模型，以建立灵活的映射关系，并在样本之间共享的语义子空间中挖掘信息。具体来说，本工作设计了支持集构建 (SC) 模块和语义空间转换 (SST) 模块来捕捉多模态语义空间中的微妙连接，从而获得更好的语义表示。在 MSVD 和 MSR-VTT 上的实验结果证明了该方法的有效性和先进性。

第四章 基于预训练模型视觉-语言知识挖掘的视频字幕生成

4.1 动机

视频字幕生成是一个跨模态生成任务，随着如今网络速度的高速发展和社交媒体上视频数据的增多，对于视频内容理解解释意的需求逐渐增加，视频字幕生成愈发重要。

视频字幕生成领域已经从多个方面产生了许多优秀的探索成果，这些方法主要集中在学习视频的时空表示来充分挖掘视觉信息，挖掘视觉的多粒度表达或针对特定领域需求发展出新的解决方案。例如，HMN^[67] 提出了一个分层模块化网络，从三个层次连接视频表示和语言语义：实体层、动作层和句子层，学习整体-局部表达，得到了更好的字幕生成效果；FLIP^[68] 结合工业界实际生产需要，为视频字幕生成限定了用户指向的需求，并利用大规模预训练的视觉模型及文本模型进行字幕生成，针对多样性和场景适配度获得提升。

这些方法虽然进行了成功的探索，但也仍然存在某些问题：如图4-1左图所示：首先，由于视频字幕生成是一个从单一视觉模态开始的生成过程，这一过程缺乏适当指导，这会导致模型倾向于生成较为普适的句子，这与获得生动描述文本的期望显然是不符合的；其次，由于训练的数据是固定的，模型没有办法扩展探索其他资源，导致对于数据的表现形式极为有限，这会导致生成文本也被限制。

针对以上问题，如图4-1右图所示，本文提出了引入外部知识来进行补充，通过对大规模预训练模型中知识的挖掘，得到更加生动多样的文本描述。具体来讲：视频字幕生成技术一般基于编码器-解码器构造，由编码器完成特征嵌入和表达优化，由解码器进行文本的生成。本工作的框架将传统方法的输入进行拓展，通过CLIP 跨模态预训练成果，无需任何调优流程，经零样本预测得到针对每个视频的辅助关键词描述，将其作为文本模态与视觉模态共同输入模型，得到一个基于大规模预训练模型外部知识的多模态的输入。此外，针对加入的多模态数据在解码器部分提出多注意力的双层解码器，让文本参与语言解码器的特征变换，更好地融合利用视觉-语言双模态的信息。最后，基于文本模态输入，在传统的生成模型上引入指针网络，使其能够动态在视觉输入产生的输出和文本输出产生的输出之间选择切换，使模型既能利用检索到的辅助关键词中的不同表达，又能生成自然准确的视频内容。

综上所述，本工作的贡献有三点：

(1) 针对视频字幕生成单一的视觉模态输入通过引入大规模跨模态预训练模

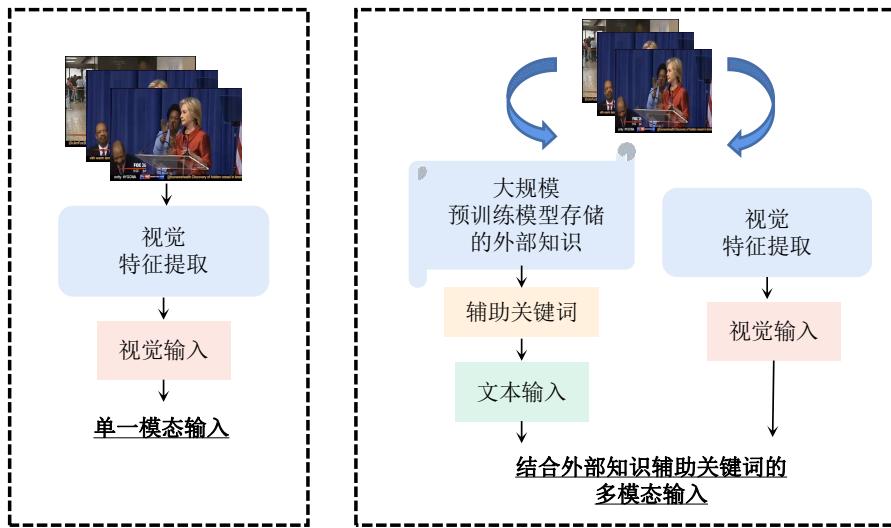


图 4-1 基于预训练模型视觉-语言知识挖掘的视频字幕生成动机示意图

型中的外部知识，得到辅助关键词的指导，这种多模态的视频输入有助于生成更为丰富多样的字幕，提高其质量；

(2) 针对文本模态，设计多注意力的双层解码器，并引入指针网络，使解码器能够动态在视觉输入产生的输出和文本输入产生的输出之间选择切换，充分利用多模态的输入信息；

(3) 在视频字幕生成数据集 MSVD 和 MSR-VTT 上的大量实验证明了本工作模型的有效性：本工作的模型基于基线在四个经典指标：BLEU-4, CIDEr, METEOR, ROUGE-L 都有显著的性能提升。

4.2 基于预训练模型的视觉-语言知识挖掘

以下将通过三个部分：辅助关键词词提取（4.2.1），多注意力的双层解码器（4.2.2）以及结合指针网络的生成模块（4.2.3）介绍本文提出的基于预训练模型视觉-语言知识挖掘的视频字幕生成。如图4-2框架图所示，针对需要描述的视频，本工作的模型先将其使用冻结的 CLIP 视觉特征提取器进行嵌入，视觉表征进一步被送入辅助关键词提取模块进行辅助关键词的生成，通过同样冻结的 CLIP 文本特征提取器得到文本表达，将视觉-语言模态特征送入多注意力的双层解码器得到注意力融合后的隐层状态，最后通过结合指针网络的生成模块得到文本预测。接下来将进行详细介绍。

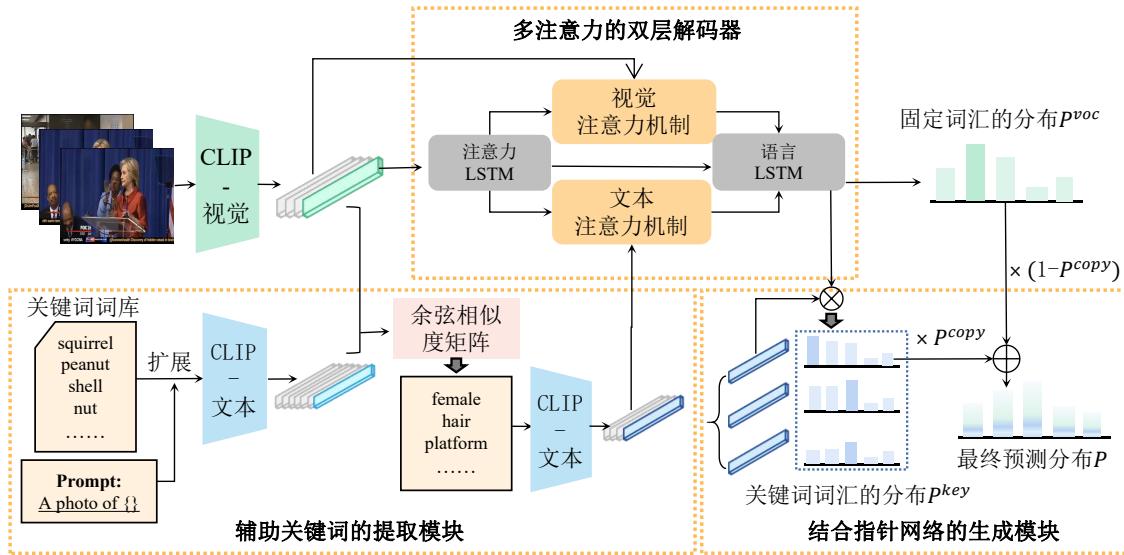


图 4-2 基于预训练模型视觉-语言知识挖掘的视频字幕生成框架示意图

4.2.1 辅助关键词的提取

CLIP 来自于文章“从自然语言监督中学习可转换的视觉模型”^[50]，是一个跨文本-视觉模态的大规模预训练模型，在其中存储了自然界非常丰富的知识。与此同时，由于其在训练过程中使用了图像-文本的跨模态对比学习训练，使得模型能够同时具有处理文本和视觉的能力，并且，得到的文本和视觉嵌入还因其独特的训练方式能够具有零样本预测的能力。综合以上特性，本文采用 CLIP 作为视觉特征提取器，并利用其零样本预测能力进行辅助关键词的提取。

首先，需要处理得到辅助关键词的词库。具体操作上，将视频字幕生成的两个数据集分别对应的总词库进行细化处理：使用 nltk 库将每个词标注词性，进一步只保留整个词库中意义较为明确的一部分词语：名词、形容词和动词，将这一细化词库成为辅助关键词词库。对每个关键词进行扩展，补充其形式为 CLIP 建议的语句格式：“一张关于 {} 的图像”，括号内为关键词，并使用 CLIP 进行句子的嵌入，得到辅助关键词矩阵 \mathbf{T}_{key} 。

得到辅助关键词的词库后，进一步抽取样本对应的关键词。针对每一个训练和测试视频，首先均等抽取 30 帧，计算每一帧 i 经过 CLIP 嵌入后的视觉表征 \mathbf{V}_{fi} 和辅助关键词矩阵 \mathbf{T}_{key} 的余弦相似度，并取相似度最高的 2 个单词，最终通过合并不去重操作得到针对每个视频的一系列辅助关键词。

通过 CLIP 零样本预测辅助关键词，实现了对 CLIP 中存储的大量外部知识的引入。最终，将样本对应的辅助关键词进行随机抽取、特征嵌入，得到了针对视频字幕生成任务一个新的文本输入 \mathbf{T} 。

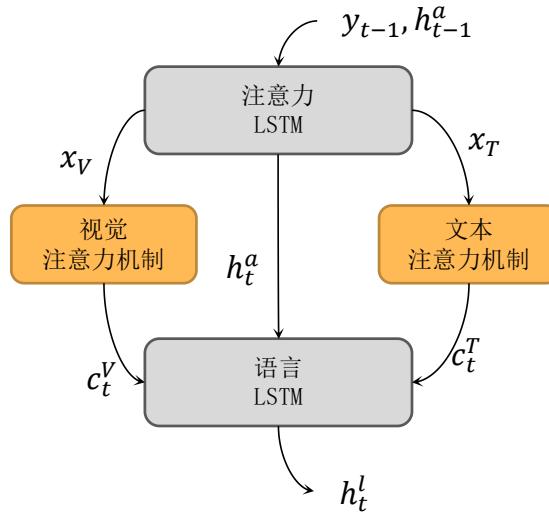


图 4-3 多注意力的双层解码器示意图

4.2.2 多注意力的双层解码器

基于引入外部知识得到的辅助关键词经过嵌入作为文本模态 \mathbf{T} 被输入到网络，由此结合视觉-语言多模态输入的特点，在解码器阶段设计了多注意力的双层解码器，如图4-3所示。视觉字幕生成任务中解码器的经典形式为双层解码器：将视觉输入的全局表达 $\mathbf{V}_{global} = MeanPool(\mathbf{V}_i)$ ，其中 $i=1$ 至 30 ，先送入注意力 LSTM，经过针对视觉的注意力机制，最后输入语言 LSTM 得到生成文本概率预测。

针对语言-文本多模态输入，多注意力的双层解码器扩展了其形式，使其能够更好地适配任务，捕捉到来自外部知识辅助关键词提供的指导。具体来讲，注意力 LSTM 根据此刻的隐藏层状态 \mathbf{h}_t^a 关注不同的视觉特征 \mathbf{x}_V ，得到关注的视觉上下文 \mathbf{c}_t^v ；此刻注意力 LSTM 的隐藏层状态根据之前的隐藏层状态 \mathbf{h}_{t-1}^a 和前一个生成单词 y_{t-1} 更新，如式 (4-1) 所示：

$$\begin{aligned} \mathbf{h}_t^a &= LSTM(\mathbf{W}_e y_{t-1}, \mathbf{h}_{t-1}^a; \theta_a), \\ \mathbf{c}_t^v &= Attention(\mathbf{h}_T^a, \mathbf{x}_V, \mathbf{x}_V; \theta_V), \end{aligned} \quad (4-1)$$

其中 W_e 是将之前生成的单词进行嵌入的矩阵。在此基础上，根据此刻注意力 LSTM 的隐藏层状态，通过关注不同的文本特征 \mathbf{x}_T 得到语言上下文 \mathbf{c}_t^T ，如式 (4-2) 所示：

$$\mathbf{c}_t^T = Attention(\mathbf{h}_T^a, \mathbf{x}_T, \mathbf{x}_T; \theta_T), \quad (4-2)$$

其中， $Attention(query, key, value; \theta)$ 为标准的注意力机制模块，使用超参数 θ 来简

化。

然后，语言 LSTM 进一步聚合注意力 LSTM 的隐藏层状态 \mathbf{h}_t^a 和视觉上下文 \mathbf{c}_t^V 、文本上下文 \mathbf{c}_t^T ，来生成得到在每个时间步固定词库的概率分布 \mathbf{p}_{voc} ，计算过程如式 (4-3) 所示：

$$\begin{aligned}\mathbf{h}_t^l &= LSTM([\mathbf{h}_t^a, \mathbf{c}_t^V, \mathbf{c}_t^T], \mathbf{h}_{t-1}^l; \theta_l), \\ \mathbf{p}_{voc} &= softmax(\mathbf{W}_{voc}\mathbf{h}_t^l + \mathbf{b}_{voc}),\end{aligned}\quad (4-3)$$

其中 $\theta_a, \theta_V, \theta_T, \theta_l$ 是分层 LSTM 和注意力模块的超参数； \mathbf{W}_{voc} 和 \mathbf{b}_{voc} 是可学习参数。

4.2.3 结合指针网络的生成模块

在 4.2.1 模块，本工作通过 CLIP 的零样本预测能力得到了辅助关键词及其嵌入，同时，利用 4.2.2 的多注意力双层解码器捕捉文本上下文，参与固定词库的预测。为了更好的利用辅助关键词，本工作结合指针网络^[69]设计了可以动态考虑固定词库和辅助词库概率的字幕生成模块。

在每个时间步 t ，该模块对于检索到的每个辅助关键词的嵌入 \mathbf{x}_{Ti} 进行注意力机制的计算，使用语言 LSTM 的隐藏层状态 \mathbf{h}_t^l 作为 query，得到针对每个词的概率 $\mathbf{p}_{key,i}$ ，如式 (4-4) 所示：

$$\mathbf{p}_{key,i}, \mathbf{c}_{i,t}^r = Attention(\mathbf{h}_t^l, \mathbf{x}_{Ti}, \mathbf{x}_{Ti}; \theta_r), \quad (4-4)$$

其中 $Attention(*)$ 是注意力模块，其参数为 θ_r ； $\mathbf{c}_{i,t}^r$ 是辅助关键词以 $\mathbf{p}_{key,i}$ 为权重得到的加权和，表示上下文关系。模型需要决定是复制辅助关键词还是根据固定词库概率来动态生成，这一决定 \mathbf{p}_{copy} 由上下文关系 $\mathbf{c}_{i,t}^r$ 和语言 LSTM 的隐藏层状态 \mathbf{h}_t^l 进行调控，如式 (4-5) 所示：

$$\mathbf{p}_{copy} = \sigma(\mathbf{W}_r \mathbf{c}_{i,t}^r + \mathbf{W}_l \mathbf{h}_t^l). \quad (4-5)$$

最终，本工作经过 \mathbf{p}_{copy} 对上下文关系的动态调控，结合固定词库的概率 \mathbf{p}_{voc} 以及扩展成固定词库长度的辅助关键词的概率分布 \mathbf{p}_{key} ，得到了结合辅助关键词次的概率分布 \mathbf{p} 。其过程如下式 (4-6) 所示：

$$\mathbf{p} = (1 - \mathbf{p}_{copy})\mathbf{p}_{voc} + \mathbf{p}_{copy}\mathbf{p}_{key}. \quad (4-6)$$

最终使用概率 \mathbf{p} 进行输出单词的预测，实现了结合指针网络的生成模块。

4.3 实验结果

4.3.1 数据集及评价指标

与第三章所示工作相同，本工作采用数据集 MSVD 和 MSR-VTT 进行实验，这两个数据集各有长短，可以对实验设计进行具体全面的结果展示。

与第三章所示工作相同，本工作采用 BLEU-4, METEOR, ROUGE-L,CIDEr 作为评价指标，来展现本文方案设计的先进性以及有效性。

4.3.2 执行细节

在实践中，本方案完善了如下细节：

首先是辅助关键词提取模块。本章工作的词库采用了从标注文本中进行分词，保留出现词频在两次以上的词汇，将其作为字幕生成的总词库。在此基础上，使用 nltk 库对词汇进行词性分类，保留其中的“名词”、“动词”和“形容词”，将保留的词汇作为关键词的词库。具体来讲，针对 MSR-VTT 数据集，总词库共有 16242 个词，辅助词库单词数量为 9616，其中包含名词 8842 个，形容词 720 个，动词 54 个；针对 MSVD 数据集，总词库共有 5295 个单词，辅助词库单词数量为 3070，其中包含名词 2847 个，形容词 192 个和动词 31 个。

在该模块，本文通过计算视觉-关键词库的余弦相似度，得到针对每一帧的关键词预测概率，对于每一帧，本方案取前两个关键词进行输出。对于每一个视频，将抽取 30 帧图像，对于这 30 帧图像得到的 60 个关键词进行去重，将得到对于单个视频的指导关键词数据。在实践中，每个视频的关键词数量最终在 7-20 的范围内。

另外，为了消融辅助关键词生成过程中基于 CLIP 模型零样本预测能力的影响，本方法还进行了引入 GT 的实验。即将辅助关键词替换为标注文本中的关键词，进行字幕生成。具体操作上，将每个标注文本使用 nltk 库进行分词，将得到的每个标注句子的词语中存在于辅助关键词库里的词语保留存储，得到 GT 情况下的辅助关键词，同样采用随机抽取将其加入到文本模态的输入中。

其次，在本文实验中，使用 CLIP 的 B-32 规格来提取视频的外观特征和文本的嵌入特征，每个视频都平均地截取 30 帧。与此同时为了缩减其他影响因素对实验结果的影响，本文框架没有引入动作特征，仅使用外观特征作为视频表达。

本工作同样采用了 Teacher-Enforced Learning(TEL)^[59] 来指导模型的学习，使其收敛更快。另外，beam search 方法在模型推理阶段被应用，使用的束大小在 MSR-VTT 中被设置成 2，在 MSVD 中被设置成 5。优化器方面，本文使用了 Adam 优化器^[70]，初始学习率被设定为 1e-4，同时设置了一定的学习率衰减。本文训练

表 4-1 在 MSVD 和 MSR-VTT 数据集上的对比实验结果。B@4, M, R, C 表示 BLEU-4, METEOR, ROUGE-L 和 CIDEr 四种评价指标，带 “*” 表示该实验进行了端到端设计并对特征提取网络进行了调优，GT 代表将辅助关键词替换为真实标注文本中提取的关键词得到的结果，可以用以展现当特征提取器的零样本预测能力提高时得到性能表现

Method	MSR-VTT				MSVD			
	B@4	M	R	C	B@4	M	R	C
SAAT ^[29]	40.5	28.2	60.9	49.1	46.5	33.5	69.4	81.0
RMN ^[30]	42.5	28.4	61.6	49.6	54.6	36.5	73.4	94.4
MGRMP ^[27]	42.1	28.8	61.4	50.1	53.2	35.4	73.5	90.7
M ³ ^[19]	38.1	26.6			51.8	32.5		
GRU-EVE ^[21]	38.3	28.4	60.7	48.1	47.9	35.0	71.5	78.1
SGN ^[24]	40.8	28.3	60.8	49.5	52.8	35.5	72.9	94.3
MGRMP ^[27]	41.7	28.9	62.1	51.4	55.8	36.9	74.5	98.5
ORG-TRL ^[59]	43.6	28.8	62.1	50.9	54.3	36.4	73.9	95.2
MDT ^[23]	40.2	28.2	61.1	47.3	49.0	35.3	72.2	92.5
Clip4Caption* ^[71]	46.1	30.7	63.7	57.7				
HMN ^[67]	43.5	29.0	62.7	51.5	59.2	37.7	75.1	104.0
Ours	46.3	30.4	64.1	57.6	61.8	40.0	76.7	107.7
Ours(GT)	47.1	30.6	64.5	59.3	64.5	40.4	76.9	111.3

使用的固定词汇表由至少出现两次的单词构成，这能够一定程度上保证模型更好的学习。

4.3.3 定量结果对比

为了验证本文提出方案的有效性，将经典工作、现阶段最先进工作与本文提出方法进行对比实验，结果如表4-1所示：本工作提出模型在两个数据集和各个指标上都取得了优秀的效果，将 Clip4Caption 和 HMN 作为实验的对照组进行对比。首先 HMN 结合了多层次语义信息，提出了一个分层模块化网络，在生成字幕之前从实体、动作和句子三个层次连接视频表示和语言语义，在实验中采用了 IRv2、C3D 和 Faster-RCNN 提取的外观特征、动作特征和物体信息。从对比中可以看到，相比传统模型，本文模型在四个指标和两个数据集上都取得了更优秀的成果。但实际上，由于经过大规模预训练的 CLIP 作为特征提取器相比传统卷积神经网络有一定的优势，与传统方法较难比较。因此，Clip4Caption 这篇文章被选为对照组。该 Clip4Caption 模型可以分为两阶段：第一阶段按照 CLIP 的预训练范式在下游数

据集 MSR-VTT 上进行微调，第二阶段将微调后的模型作为特征提取器进入网络生成视频对应的视频描述。这一方法虽然取得了一定的成果，但是实际上在微调操作中带来了非常大的计算需求，这使得这一方法需要的资源和时间较长。即使如此，从表中可以看到，本章模型在没有对预训练模型引入任何调优手段的前提下，仅使用其零样本预测能力达到了优秀的效果，由此可以证明本章中提到方法的有效性和优越性。

4.3.4 消融实验

在对比实验的基础上，为了继续验证各个模块的影响、辅助关键词的相关细节等具体信息，进一步进行消融实验，分析模型细节。

4.3.4.1 各个模块的影响

首先为了验证各个模块设计对于模型性能的影响，设计了如下消融试验：

- (1) 第一部分为基础模型，即仅仅使用 CLIP 作为特征提取器对视频进行特征提取，采用经典的双层 LSTM 解码器进行文本生成得到的结果；
- (2) 第二部分为在基础模型的基础上加入了辅助关键词提取并指针网络设计，即引入 4.2.1 辅助关键词提取和 4.2.3 结合指针网络的生成模块得到的结果。
- (3) 第三部分为在基础模型上加入辅助关键词提取、对文本的注意力模块以及指针网络设计，即引入 4.2.1 辅助关键词提取、4.2.2 多注意的双层解码器和 4.2.3 结合指针网络的生成模块得到的结果。

表 4-2 MSVD 数据集上消融实验

		MSVD			
		B@4	M	R	C
(I)		59.0	39.0	75.7	103.7
(II)		61.6	39.3	75.7	105.1
(III)		61.8	40.0	76.7	107.7

如表4-2所示，可以看到 (II) 相对 (I) 有较大的性能提升，这说明辅助关键词结合指针网络的生成模块确实能为生成文本引入一定的丰富性。并且，(III) 相对于 (II) 同样有显著提升，这说明对于文本的多注意力机制设计能够基于文本更多的关注，使得模型的生成部分能够学到更加丰富优秀的表达，给予生成文本更优的性能。

4.3.4.2 辅助关键词相关细节

辅助关键词是本方法最重要的组成部分之一，通过辅助关键词的抽取，可以利用挖掘来自 CLIP 预训练模型中存储的外部知识，为视频字幕生成这一下游任务增添新的驱动力。本节探究辅助关键词数量和质量对不同数据集的影响，并分析如下。

表 4-3 MSVD 数据集上辅助关键词抽取数量 k 以及质量的影响

	MSVD			
	B@4	M	R	C
k=3	61.8	40.0	76.7	107.7
k=5	61.5	39.2	75.9	107.4
k=10	62.3	39.8	76.2	107.6
GT k=3	62.8	39.7	76.5	105.9
GT k=5	63.6	40.2	77.1	108.8
GT k=10	64.5	40.4	76.9	111.3

表4-3为在 MSVD 数据集上辅助关键词抽取数量 k 及质量影响的结果示意。前三行分别为辅助关键词数量 k 为 3、5、10 时的性能，可以看到在 MSVD 数据集上，这三者的性能差别实际并不悬殊：由上文对各个模块的消融可以看到辅助关键词可以带来生成文本的性能增益，因此可以得出，针对抽取的辅助关键词，加入模型的数量影响并不强烈，都能对模型带来一定的性能提升。

后三行则为在 MSVD 数据集上引入了标注文本抽取的关键词作为辅助关键词的结果，分别为辅助关键词数量 k 为 3、5、10 时的性能。从后三行可以得出一个发现：当采用可靠的标注文本时，辅助关键词的数量越大，模型能够获得的外部信息就越多越可靠，生成文本质量越高。

由此可以得出，在辅助关键词的提取过程中，若能使用零样本预测能力更好的预训练模型或是在检索辅助关键词的过程中引入更多先进设计，模型性能仍能进一步提升，这也可作为针对该模型设计下一步的研究方向。

为了进一步探究辅助关键词带来的影响，进一步在 MSR-VTT 数据集上进行了实验，在表4-4中进行展示。在 MSR-VTT 上的实验进一步验证了上文中提到的结论：在表中的前三行性能差距并不显著，在表的后三行，随着引入的可靠的辅助关键词的增多，模型性能进一步提升。但是 MSR-VTT 数据集又与 MSVD 存在细微差异：MSR-VTT 的数据量近乎为 MSVD 的五倍，并且其内容信息含量更为丰富，标注文本平均长度也更长，这使得 MSR-VTT 上的预测较 MSVD 更为困难。即使如此，在零样本预测下的辅助关键词仍然获得了更好的性能，并且当辅助关

表 4-4 MSR-VTT 数据集上辅助关键词抽取数量 k 以及质量的影响

	MSR-VTT			
	B@4	M	R	C
k=3	46.3	30.4	64.1	57.6
k=5	45.5	30.2	63.7	58.1
k=10	46.0	30.3	63.9	57.7
GT k=3	46.4	30.5	64.3	58.4
GT k=5	47.1	30.6	64.5	59.3
GT k=10	47.1	30.6	64.5	59.3

关键词替换为更可靠的标注文本抽取的关键词时，模型性能获得了显著提升，也证明了本章多注意力的双层解码器以及结合指针网络的生成模块两者的有效性。

4.3.5 可视化展示

图4-4为本工作得到结果的可视化展示。

上下部分分别为两个样本。上半样本首先展示了原视频的三帧内容，可以看到主要为歌手“Lady Gaga”在进行表演，后面有一位舞者在拍手。下面第一部分附有通过本章提出的辅助关键词提取模块得到的关键词，其中内容为通过大规模预训练模型中外部知识引入的新概念，例如：歌手名“gaga”、“鼓掌”，“子标题”，“mtv”等有利于文本生成的概念。下面接有从标注文本中提取的关键词如“gaga”、“音乐”、“女性”、“黑色”和“白色”等。下一部分为与该视频对应的文本描述，“GT”代表人工标注的真实文本，“Baseline”代表基准模型，“Ours”代表本工作模型得到的文本。红色标注部分为突出显示的被认为对生成有辅助作用的关键词汇。从生成结果可以看到，相较于未引入外部知识的基准模型，本工作成功将外部知识歌手名“gaga”引入生成文本中，生成了“gaga 在演唱”而不是简单的“一个人在演唱”，优化了视频字幕的表达。

下半部分与其类似，同样是通过外部知识的引入为模型带来了“染”“头发”等重要的辅助关键词。可以看到辅助关键词中存在“染色”、“涂色”、“蓝色”等关键词，都能够非常形象的展示视频中的信息，使得在生成的文本中捕捉到了该语义，输出“一个女性正在展示如何染头发”的字幕。相比基准模型生成的字幕，因为引入了外部知识，能够得到更为丰富和准确的语义。

4.3.6 存在的不足以及未来工作

即使获得了优秀的表现，本工作仍然存在一定不足。当今深度学习领域有许多大模型的涌现，其发展日新月异，作为研究资源较为有限的科研学习者，调用数

以万计的计算资源显然并不现实，但运用大规模预训练打下的优良基础，在某些特定数据集上进行调优也不失为一个好的方式。本实验出于对影响因素的控制没有引入调优方法，未进行端到端的训练，但在未来工作的方面，可以基于本文提出的辅助关键词的思想，将端到端的训练范式以及调优手法引入到这一课题，便可以结合两者的优势，进行进一步发展。

其次，大模型的开源程度也将会是未来工作需要思考的一个问题。现今的大规模语言模型已经有了一定的商业化实现，其模型参数并不能直接获得，视觉大规模预训练模型和多模态预训练模型也无可避免的将涉及到这一问题。在不能将模型直接获取和参数优化的前提下，如何通过查询的设计得到一定的知识，优化自身的任务，也是未来工作需要进一步发展的方向。

4.4 本章小结

视频字幕生成是一个重要的一个跨模态的生成任务。本章针对传统方法由于固定的训练数据以及缺乏适当指导导致的描述匮乏不充分的问题，引入了 CLIP 预训练模型存储的外部知识。通过基于 CLIP 视觉-语言跨模态共享的语义空间及其零样本预测能力，获得了外部知识指导下的辅助关键词，并利用多注意力的双层解码器和结合指针网络的生成模块将辅助关键词和生成模型有机结合，得到了新的字幕生成方法。在 MSVD 和 MSR-VTT 上的大量实验证明了该模型的有效性和先进性。

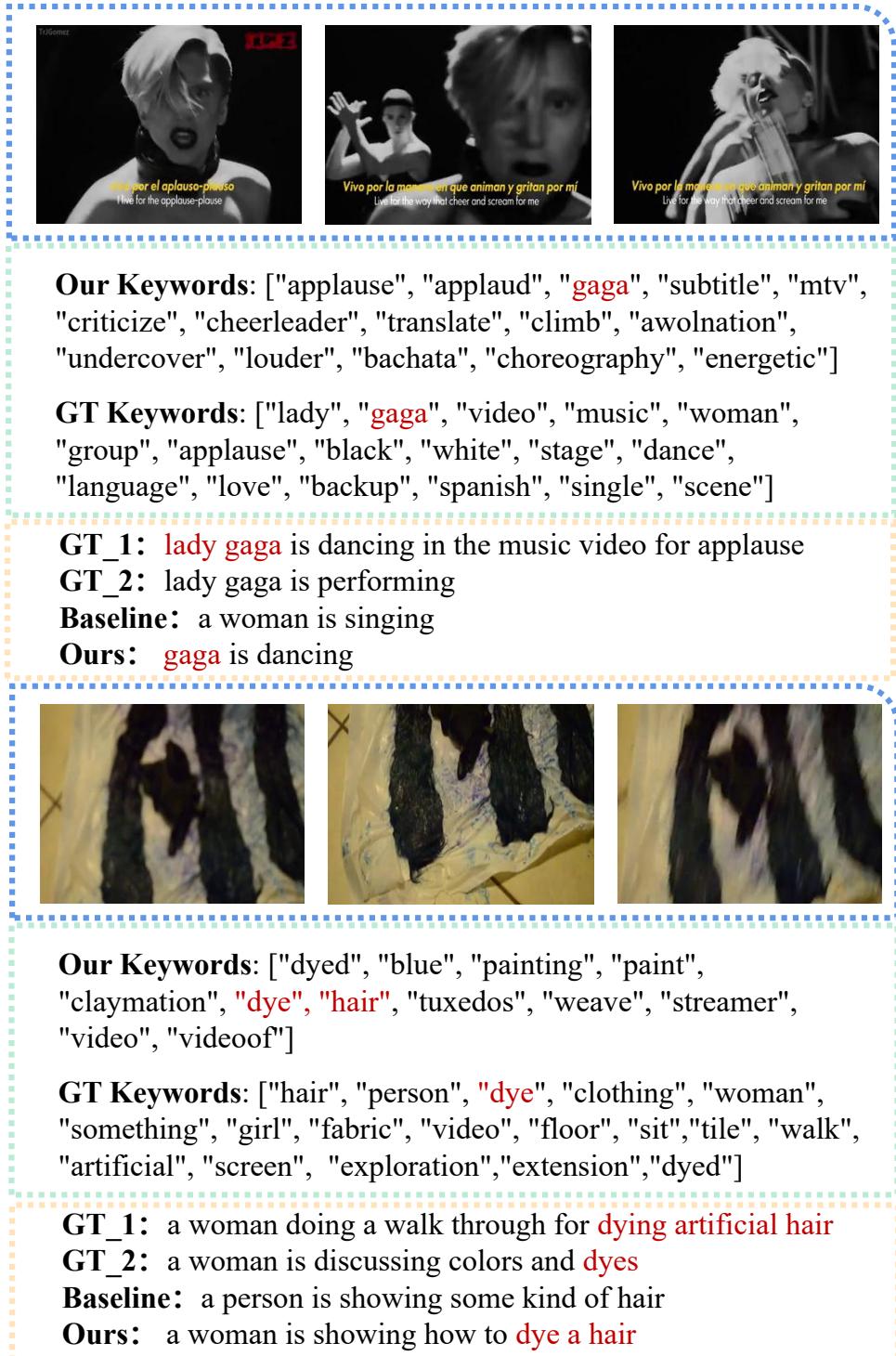


图 4-4 本工作生成字幕的可视化展示

第五章 全文总结与展望

5.1 全文总结

视频字幕生成是生动具体地描述视频中存在的主体及事件的任务，在盲人辅助，智能问答等丰富领域都有重要贡献。视频字幕生成作为跨模态领域最重要的任务之一，虽然在研究者的不断努力下得到了不断地发展和提升，但时至今日仍然面临着许多未解决的问题，例如：现阶段的模型存在信息挖掘不够充分的问题，单纯依靠引入更多的特征提取网络、引入更大的模型设计，反而会忽略对内部知识的挖掘，对现有资源没有充分利用；其次，现阶段的生成模型得到的文本描述仍然是匮乏、不生动的，虽然获取更多更全面的训练数据能够缓解这一问题，但数据的获得也面临着高昂的人力成本。对于这些问题的攻克能使视频字幕生成领域向着准确、详尽、生动的目标更进一步。

针对以上问题，本文分别提出了应对的方案：

(1) 针对对现有信息挖掘不充分的问题，本文提出了进行内部知识挖掘的“基于支持集的视觉表达增强模型”。该模型针对传统一对一映射带来的表达匮乏问题，构建了一种新颖灵活的映射框架，通过捕捉样本之间内部细节的联系，在不引入其他特征负担的情况下，构建支持集和语义空间转换模块，使得模型能够从丰富灵活的映射关系中学习增强视觉表达，获得更好的文本描述，推动了对内部知识的挖掘。

(2) 针对现有描述匮乏的问题，本文提出了引入外部知识的“基于预训练模型视觉-语言知识挖掘的视频字幕生成”模型。该模型从大规模预训练的先进成果出发，针对现有模型的生成过程缺乏适当指导和有限的训练数据的问题，引入了大规模预训练模型 CLIP 中的外部知识，通过对外部知识的挖掘，构建了辅助关键词，并通过多注意力的双层解码器和结合指针网络的生成模块将其与经典框架有机结合，在仅使用视频的外观表征的前提下，通过构建的视觉-语言模态输入获得了先进有效的性能。这一部分对外部知识进行了引入和挖掘。

5.2 后续工作展望

以上工作对视频字幕生成的发展进行了一定探索，但该领域仍有很多的问题值得继续探讨和展望。

首先，针对内部知识的挖掘仍然是一个重要的话题。以上工作中设计了基于支持集的视觉表达增强来优化模型的映射学习范式，得到了先进的性能，但这样

的挖掘仍然有进一步发展的空间，例如，在该工作中，每次训练获取的样本批次的数量影响着支持集的质量，为了应对该问题，支持集的构建可以设计其他更为准确的手段；另外，在语义空间变换的过程中可以加入更为具象，或者是多粒度的语义约束，能够进一步优化视觉表达。并且，针对第三章工作出现的问题，也可以在未来工作中进行展开：通过约束手段对数据集内部“偏见”进行约束，使生成文本更加符合无偏见的需求；并且，针对生成文本存在一定重复的问题，可以通过在生成部分加入重复掩码或是在未来工作中结合其他手段进行进一步探索。

其次，外部知识的引入也能够有更多的探索发展。以上工作设计了经过大规模预训练模型的知识挖掘，但出于不引入大量计算参数的考虑，仅仅简单使用辅助关键词设计，没有对预训练的特征提取器进行下游任务的调优，若能够应用端到端的生成范式进行调优，会得到适配领域和任务的更优结果。另外，随着大规模预训练模型的不断发展，越来越多的先进模型选择不开源其训练细节和权重，将其功能包装成一种服务供他人调用，这种模式在自然语言领域被称为 LMaaS^[72]（语言模型作为服务），跨模态领域也正走向如此道路。面对这样的趋势，为了更好地挖掘引入的外部知识的内容，应进一步设计调用服务接口的优化方法，将先进模型的知识引入到视频字幕生成课题中。

以上的实验和探索仅仅是视频字幕生成领域的一个部分，该领域的发展仍然需要研究者投入努力，得到更多的发现和证明。后续工作将继续探索对于视觉内容和语言信息的理解、分析和挖掘，推动该领域的进程。

致 谢

参考文献

- [1] Shin A, Ishii M, Narihira T. Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision[J]. International journal of computer vision, 2022, 130(2): 435-454.
- [2] Li S, Tao Z, Li K, et al. Visual to text: Survey of image and video captioning[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2019, 3(4): 297-312.
- [3] Shin A, Ishii M, Narihira T. Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision[J]. International journal of computer vision, 2022, 130(2): 435-454.
- [4] Sharma H, Jalal A S. Image captioning improved visual question answering[J]. Multimedia Tools and Applications, 2022, 81(24): 34775-34796.
- [5] 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, Boston, USA, 2015: 3156-3164.
- [6] Islam S, Dash A, Seum A, et al. Exploring video captioning techniques: A comprehensive survey on deep learning methods[J]. SN Computer Science, 2021, 2(2): 1-28.
- [7] Jain V, Al-Turjman F, Chaudhary G, et al. Video captioning: a review of theory, techniques and practices[J]. Multimedia Tools and Applications, 2022, 81(25): 35619-35653.
- [8] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, USA, 2018: 6077-6086.
- [9] Aafaq N, Mian A, Liu W, et al. Video description: A survey of methods, datasets, and evaluation metrics[J]. ACM Computing Surveys, 2019, 52(6): 1-37.
- [10] 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]. Proceedings of the IEEE conference on computer vision and pattern recognition, Portland, USA, 2013: 2634-2641.
- [11] Yu H, Siskind J M. Grounded language learning from video described with sentences[C]. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 2013: 53-63.

- [12] Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of actions[J]. International Journal of Computer Vision, 2002, 50(2): 171-184.
- [13] Guadarrama S, Krishnamoorthy N, Malkarnenkar G, et al. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition[C]. Proceedings of the IEEE international conference on computer vision, Sydney, Australia, 2013: 2712-2719.
- [14] Rohrbach M, Qiu W, Titov I, et al. Translating video content to natural language descriptions[C]. Proceedings of the IEEE international conference on computer vision, Sydney, Australia, 2013: 433-440.
- [15] Rohrbach A, Rohrbach M, Qiu W, et al. Coherent multi-sentence video description with variable level of detail[C]. Pattern Recognition: 36th German Conference, Münster,Germany, 2014: 184-195.
- [16] Baum L E, Petrie T, Soules G, et al. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains[J]. The annals of mathematical statistics, 1970, 41(1): 164-171.
- [17] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]. Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015: 2048-2057.
- [18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems, Long Beach, USA, 2017: 5998-6008.
- [19] Wang J, Wang W, Huang Y, et al. M3: Multimodal memory modelling for video captioning[C]. IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7512-7520.
- [20] Pei W, Zhang J, Wang X, et al. Memory-attended recurrent network for video captioning[C]. IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 8347-8356.
- [21] 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, Long Beach, USA, 2019: 12487-12496.
- [22] Wang B, Ma L, Zhang W, et al. Controllable video captioning with pos sequence guidance based on gated fusion network[C]. IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 2641-2650.

- [23] Zhao W, Wu X, Luo J. Multi-modal dependency tree for video captioning[C]. Advances in Neural Information Processing Systems, virtual, 2021: 6634-6645.
- [24] Ryu H, Kang S, Kang H, et al. Semantic grouping network for video captioning[C]. Proceedings of the AAAI Conference on Artificial Intelligence, virtual, 2021: 2514-2522.
- [25] Zhang Z, Qi Z, Yuan C, et al. Open-book video captioning with retrieve-copy-generate network[C]. IEEE Conference on Computer Vision and Pattern Recognition, virtual, 2021: 9837-9846.
- [26] Chen S, Jiang Y G. Motion guided spatial attention for video captioning[C]. Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, USA, 2019: 8191-8198.
- [27] Chen S, Jiang Y G. Motion guided region message passing for video captioning[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 1543-1552.
- [28] Pan B, Cai H, Huang D A, et al. Spatio-temporal graph for video captioning with knowledge distillation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 10870-10879.
- [29] Zheng Q, Wang C, Tao D. Syntax-aware action targeting for video captioning[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 13096-13105.
- [30] Tan G, Liu D, Wang M, et al. Learning to discretely compose reasoning module networks for video captioning[C]. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan, 2020: 745-752.
- [31] Jang E, Gu S, Poole B. Categorical reparameterization with gumbel-softmax[C]. International Conference on Learning Representations, Toulon, France, 2017: 1-12.
- [32] Gao L, Lei Y, Zeng P, et al. Hierarchical representation network with auxiliary tasks for video captioning and video question answering[J]. IEEE Trans. Image Process., 2022, 31: 202-215.
- [33] LeCun Y, Bengio Y, et al. Convolutional networks for images, speech, and time series[J]. The handbook of brain theory and neural networks, 1995, 3361(10): 1995.
- [34] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 248-255.
- [35] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

- [36] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
- [37] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. International Conference on Learning Representations, San Diego, USA, 2015: 1-14.
- [38] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, USA, 2015: 1-9.
- [39] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770-778.
- [40] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, USA, 2017: 1251-1258.
- [41] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]. Proceedings of the AAAI conference on artificial intelligence, San Francisco, USA, 2017: 1-1.
- [42] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6450-6459.
- [43] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?[C]. IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6546-6555.
- [44] 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, Honolulu, USA, 2017: 6299-6308.
- [45] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]. Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 2015: 4489-4497.
- [46] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, USA, 2017: 1492-1500.
- [47] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention[C]. Advances in Neural Information Processing Systems, Montreal, Canada, 2014: 2204-2212.
- [48] Ackley D H, Hinton G E, Sejnowski T J. A learning algorithm for boltzmann machines[J]. Cognitive science, 1985, 9(1): 147-169.

- [49] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]. 9th International Conference on Learning Representations, virtual, 2021: 1-22.
- [50] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]. International conference on machine learning, virtual, 2021: 8748-8763.
- [51] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF international conference on computer vision, Montreal, Canada, 2021: 10012-10022.
- [52] Elman J L. Finding structure in time[J]. Cognitive science, 1990, 14(2): 179-211.
- [53] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
- [54] Patrick M, Huang P, Asano Y M, et al. Support-set bottlenecks for video-text representation learning[C]. International Conference on Learning Representations, virtual, 2021: 1-18.
- [55] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]. IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 815-823.
- [56] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping[C]. IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, 2006: 1735-1742.
- [57] Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,, Minneapolis, USA, 2019: 4171-4186.
- [58] Faghri F, Fleet D J, Kiros J R, et al. VSE++: improving visual-semantic embeddings with hard negatives[C]. British Machine Vision Conference, Newcastle, UK, 2018: 12.
- [59] Zhang Z, Shi Y, Yuan C, et al. Object relational graph with teacher-recommended learning for video captioning[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, USA, 2020: 13278-13288.
- [60] 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, Portland, USA, 2011: 190-200.

- [61] Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation[J]. *The American Statistician*, 1983, 37(1): 36-48.
- [62] Xu J, Mei T, Yao T, et al. Msr-vtt: A large video description dataset for bridging video and language[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 5288-5296.
- [63] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]. Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, USA, 2002: 311-318.
- [64] Banerjee S, Lavie A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments[C]. Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Baltimore, USA, 2005: 65-72.
- [65] Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, USA, 2015: 4566-4575.
- [66] Lin C Y. Rouge: A package for automatic evaluation of summaries[C]. Text summarization branches out, Barcelona, Spain, 2004: 74-81.
- [67] Ye H, Li G, Qi Y, et al. Hierarchical modular network for video captioning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 17939-17948.
- [68] Nie L, Qu L, Meng D, et al. Search-oriented micro-video captioning[C]. Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 2022: 3234-3243.
- [69] Vinyals O, Fortunato M, Jaitly N. Pointer networks[C]. Advances in Neural Information Processing Systems, Montreal, Canada, 2015: 2692-2700.
- [70] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]. International Conference on Learning Representations, San Diego, USA, 2015: 1-15.
- [71] Tang M, Wang Z, Liu Z, et al. Clip4caption: Clip for video caption[C]. Proceedings of the 29th ACM International Conference on Multimedia, virtual, 2021: 4858-4862.
- [72] Sun T, Shao Y, Qian H, et al. Black-box tuning for language-model-as-a-service[C]. International Conference on Machine Learning, Baltimore, USA, 2022: 20841-20855.

攻读硕士学位期间取得的成果

- [1] **第一作者**. Support-set based multi-modal representation enhancement for video captioning[C]. IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, China, 2022: 1-6.