密 级_____

桂林電子科技大學
GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY

硕 士 学 位 论 文

（全日制专业学位硕士）

题目 基于深度学习的视频语义解析研究

（英文）Research on Video Semantic Parsing Based on Deep Learning

研 究 生 学 号: **20022303183**

研 究 生 姓 名: 周美欣

指导教师姓名、职称: 蔡晓东 研究员

申 请 学 位 类 别: 电子信息硕士

领 域: 电子信息

论 文 答 辩 日 期: 2023 年 5 月 23 日

摘要

视频描述作为视觉语义解析的高级表现形式，旨在用一句清晰明确的句子描述视频语义内容。近年，伴随着互联网技术的飞速发展，手机和面向用户的拍照设备的普遍出现，使得视频数据被快速生产、存储和上传。然而，这些数据数量巨大且杂乱无序，如何利用机器快速高效地分析视频所包含的语义信息，对庞大的视频数据进行有效的组织管理以及为分类、检索等任务提供有效的参考，已经成为计算机视觉领域亟待解决的问题。目前，视频内容描述任务仍然存在生成描述的语义内容不完整和不准确等问题。此外，中文训练样本的缺失也增加了中文描述的难度。针对以上问题，进行了基于深度学习的视频语义解析的研究。首先，针对英文生成描述语义信息不完整和不准确的问题，提出了一种基于句子语义与长度损失计算的视频描述的方法。然后在英文视频描述方法的基础上，提出了一种基于余弦注意力和语义选择的视频中文描述方法，解决了中文描述的语义冗余和语义不准确的问题。接着，在视频中文描述研究的基础上，又提出了一种基于自适应特征选择和融合的视频中文描述方法，进一步提升了生成的中文描述语义的质量。本文所做的工作内容与贡献如下：

1.针对生成的英文描述的语义不完整和不准确问题，提出了一种基于句子语义与长度损失计算的视频描述的方法。首先，设计了一个新的长度损失函数，通过度量预测句长与参考句长的距离来自适应调整错误惩罚，使得模型可以在高度相似的视觉内容中学习到最优的描述长度分布，从而提升生成描述语义信息的完整性。其次，设计了一个基于句子语义的描述生成损失函数，通过对预测与参考描述在句子级别的语义比较，使模型迭代获得最优的句子语义描述，从而提升生成描述语义信息的准确性。本方法在 MSVD 和 MSR-VTT 两个数据集上测试，各项性能指标显著提升，均优于目前先进的模型。其中 BLEU@4 和 METEOR 指标在两个数据集上提升尤为显著，说明该方法对于提升描述内容语义信息的完整性和准确性是十分有效的。

2.针对生成的中文描述语义不准确的问题，提出了一种基于余弦注意力与语义选择的视频中文描述方法。首先，设计一个缩放余弦注意力网络，先将查询和键矩阵进行余弦相似度计算，再通过可学习的参数进行放大，使得模型能自适应地关注正确的视觉语义特征，提升生成语义正确的描述。其次，在解码阶段，设计了一种语义选择网络，过滤由视觉语义特征与句子语义特征融合产生的冗余信息，减少干扰，提升模型语义的准确性。最后，将视频英文描述数据集 MSVD 扩展成中文数据集 MSVD-C，并在该数据集上进行实验，结果表明各项指标和实际描述内容都显著优于其他先进模型，说明本方法不仅能准确地关注视觉语义特征，还能过滤冗余信息。

3.针对生成的中文描述语义不准确的问题，还提出了一个基于自适应特征选择与融合的视频中文描述方法。首先，设计一个特征选择网络，先使用注意力机制关注重

要的特征，再使用门控机制选择性地保留或丢弃特征，从而提升模型语义的准确性。其次，设计一种自适应动态融合机制，通过计算运动特征的权重系数将视觉和运动特征向量进行动态融合，减少冗余信息的干扰，从而提升模型语义的准确性。最后，还在中文数据集 MSVD-C 上进行上测试，结果表明各项指标和实际描述内容都显著优于其他先进模型，其中 BLEU@4 和 CIDEr-D 指标提升尤为显著，说明本方法在降维时不仅能避免丢失重要信息，还能自适应融合视觉特征和运动特征避免产生冗余信息。

综上所述，本文成功解决了视频描述任务中存在的语义不完整和不准确问题，并且在中英文语言环境下均获得了显著的效果。本研究的成果为视频语义解析领域的技术发展提供了有力的支持，同时也有望促进多语言自然语言处理技术的发展。

关键字：视频描述；句子语义；句子长度损失计算；缩放余弦注意力；语义选择；冗余信息；特征选择；

Abstract

As an advanced representation of visual semantic parsing, video description aims to describe the semantic content of a video with a clear and unambiguous sentence. In recent years, with the rapid development of Internet technology, the widespread emergence of mobile phones and user-oriented camera equipment has enabled the rapid production, storage and upload of video data. However, the amount of these data is huge and disorderly. How to use machines to quickly and efficiently analyze the semantic information contained in the video, effectively organize and manage the huge video data, and provide effective reference for tasks such as classification and retrieval has become a problem for computers. Problems that need to be solved urgently in the field of vision. At present, the video content description task still has the problem of incomplete and inaccurate semantic content of the generated description. In addition, the lack of Chinese training samples also increases the difficulty of Chinese description. Aiming at the above problems, research on video semantic analysis based on deep learning is carried out. First, to solve the problem of incomplete and inaccurate semantic information in English generated descriptions, a video description method based on sentence semantics and length loss calculation is proposed. Then, based on the English video description method, a Chinese video description method based on cosine attention and semantic selection is proposed, which solves the problem of semantic redundancy and semantic inaccuracy in Chinese description. Then, based on the research on video Chinese description, a Chinese video description method based on adaptive feature selection and fusion is proposed, which further improves the semantic quality of the generated Chinese description. The content and contributions of this paper are as follows:

1. Aiming at the semantic incompleteness and inaccuracy of the generated English description, a method of video description based on sentence semantics and length loss calculation is proposed. First, a new length loss function is designed to adaptively adjust the error penalty by measuring the distance between the predicted sentence length and the reference sentence length, so that the model can learn the optimal description length distribution in highly similar visual content, thereby improving the generation Describe the completeness of semantic information. Secondly, a description generation loss function based on sentence semantics is designed. By comparing the prediction and reference descriptions at the sentence level, the model iteratively obtains the optimal sentence

semantic description, thereby improving the accuracy of generating description semantic information. This method is tested on two data sets, MSVD and MSR-VTT, and the performance indicators are significantly improved, all of which are better than the current advanced models. Among them, the BLEU@4 and METEOR indicators are particularly improved on the two data sets, indicating that this method is very effective in improving the completeness and accuracy of the semantic information of the description content.

2. Aiming at the problem of inaccurate semantics in the generated Chinese description, a video Chinese description method based on cosine attention and semantic selection is proposed. First of all, a scaled cosine attention network is designed to calculate the cosine similarity between the query and the key matrix, and then zoom in through learnable parameters, so that the model can adaptively focus on the correct visual semantic features and improve the generation of semantically correct descriptions. . Secondly, in the decoding stage, a semantic selection network is designed to filter redundant information generated by the fusion of visual semantic features and sentence semantic features, reduce interference, and improve the accuracy of model semantics. Finally, the video English description data set MSVD is extended to the Chinese data set MSVD-C, and experiments are carried out on this data set. The results show that the indicators and actual description content are significantly better than other advanced models, which shows that this method can not only accurately Focusing on visual semantic features, it can also filter redundant information.

3. Aiming at the inaccurate semantics of the generated Chinese description, a video Chinese description method based on adaptive feature selection and fusion is also proposed. First, design a feature selection network, first use the attention mechanism to focus on important features, and then use the gating mechanism to selectively retain or discard features, thereby improving the accuracy of the model semantics. Secondly, an adaptive dynamic fusion mechanism is designed to dynamically fuse the visual and motion feature vectors by calculating the weight coefficients of the motion features to reduce the interference of redundant information, thereby improving the accuracy of the model semantics. Finally, the upper test was also carried out on the Chinese data set MSVD-C, and the results showed that the indicators and actual description content were significantly better than other advanced models, among which the improvement of BLEU@4 and CIDEr-D indicators was particularly significant, indicating that this method is reducing Time-dimensioning can not only avoid losing important information, but also adaptively fuse visual features and motion features to avoid redundant information.

To sum up, this paper successfully solves the problem of semantic incompleteness and

inaccuracy in the video description task, and achieves remarkable results in both Chinese and English language environments. The results of this research provide strong support for the technological development in the field of video semantic analysis, and are also expected to promote the development of multilingual natural language processing technology.

Keywords: video description; sentence semantics; sentence length loss computation; scaled cosine attention; semantic selection; redundant information; feature selection;

目 录

第一章 绪论

§1.1 研究背景与意义

随着互联网技术的高速发展和多媒体智能设备（手机、平板、摄像机等）的大规模普及，文字、图像和视频等多媒体数据开始呈现爆炸式增长。在当今以互联网占据主流的时代背景下，每天都有许多人在网上工作、享受娱乐、接受教育、购物以及通过手机 APP（b 站、快手、抖音、微信等)分享生活的日常等。据数据统计，每分钟 YouTube 用户上传的视频时长超过 500 小时，而快手、抖音等短视频平台每日的短视频上传数量已经超过千万级别。中国互联网络信息中心的《第 48 次中国互联网络发展状况统计报告》显示，中国网民规模高达 10 亿，网络视频（包括短视频）用户数量达到 9.44 亿，用户使用率已超 90%[1]。其中，短视频用户占网民总数的三分之二以上，高达 8.88 亿的规模，成为互联网移动时代的一种主流表达方式。

随着短视频等新型媒体的快速发展和普及，视频数据已经成为网络中占据重要地位的数据类型之一。视频数据所包含的丰富信息不仅给人类带来了娱乐和信息获取的便利，还被广泛应用于视频监控、视频会议、自动驾驶等各种领域。然而，由于视频数据的高维和复杂性，视频语义解析一直是计算机视觉领域的重要研究方向之一。在视频内容理解中，视频语义解析是一项具有挑战性的任务[1]。传统的视频分类、检索等方法通常只能对视频进行表面的特征提取，无法对视频进行深度的语义分析。因此，视频语义解析的研究旨在通过自然语言描述来准确地表达视频的语义信息。

目前，随着深度学习和自然语言处理技术的快速发展，视频描述任务也取得了很大的进展，已经成为计算机视觉和自然语言处理领域的研究热点之一[3,4]。视频描述任务作为视频语义解析的高阶形式不仅可以提高视频检索、视频推荐、视频自动生成等应用的效率，还可以为人类和机器之间的交互提供更加自然和高效的方式。因此，研究基于深度学习的视频语义解析课题具有实际的应用价值，此外，视频语义理解在学术上也是一个前沿研究课题，非常具有挑战性。综上所述，视频语义解析课的研究具有巨大的研究价值和应用前景。

§1.2 国内外研究历史与现状

视频描述作为一种人工智能技术，需要深入挖掘视频内容的语义信息，将视频内容转化为自然语言描述来提高视频的可理解性。近年来，随着视频内容的迅速发展，越来越多的研究者开始重视视频描述任务，并提出了多种有效的方法来生成描述视频

内容的句子。这些方法可分为基于模板和基于深度学习的两种类型。目前，基于编码解码结构的方法是视频描述领域的主流方法，主要得益于基于深度神经网络的编码解码模型在机器翻译领域取得突破进展的启发[4]。接下来将介绍国内外视频描述方法的研究现状以及现有视频描述方法的不足。

§1.2.1 基于模板的视频描述方法

由于早期硬件设备落后、数据处理能力弱，研究人员采用基于模板的视频描述生成方法。该方法需先通过计算机视觉技术检测出视频中目标的属性、动作、场景以及目标间的交互关系，然后将检测到的语义转化为词，再使用规定的模板进行单词填充生成描述句子。2002 年，Kojima 等人[6]首次提出视频描述方法，该方法侧重描述单人单动作的视频。首先通过颜色和形状分布区分出画面中的目标和背景，再利用检测技术进一步识别视频中目标姿态，然后将识别的动作定义为谓语，而不同的目标特征分别对应描述语句中的主语和宾语，即对语句模板进行填空以获得完整的描述语句。但由于该方法模板固定和词汇量小，会导致描述语句可读性差和语义不丰富。Rohrbach 等人[7]提出一种采用条件随机场的模型，该模型学习视频内容的目标及目标属性并转化为自然语言描述，使用最大后验估计学习语句中的谓语，再生成语义丰富的描述语句。Thomason 等人[8]提出了一种图形化模型，首先获取视频中目标的相关信息的置信度，然后使用图形化模型将获得的信息与语料库中的概率知识进行融合以获得描述语句的主语、谓语、宾语，最后将不同部分填入模板得到完整描述语句。由于基于模板的视频描述生成方法的描述模板和识别视觉信息固定，所以采用基于模板的视频描述方法会导致生成描述的句式不够灵活，因此基于模板的视频描述方法存在一定的局限性。

§1.2.2 基于深度学习的视频描述方法

由上节可知，基于模板的视频描述方法存在一定的局限性，人类语言丰富灵活，而基于模板的视频描述方法根本无法生成令人满意的描述。随着深度学习在图像、语音等领域取得不错成果，基于深度学习的描述生成方法开始日益受到人们的关注。基于深度学习的描述生成方法通常采用基于序列到序列（编码器-解码器）模型[9-12]，即先使用编码器提取视觉语义特征，再使用解码器将来自编码后的视觉语义特征进行解码，最后根据解码后的特征生成准确的描述语句。序列到序列模型最先应用于自然语言处理的机器翻译任务，该方法将输入序列转换为特征向量的编码器和再将特征向量解码为目标语言序列的解码器组成了基于编码解码结构的方法，可实现跨语言翻译。

视频描述的本质是将视觉信息转为自然语言信息，因此可以将视频描述比作机器

翻译任务的输入序列，生成的描述句子比作特定语种的输出序列。2015 年，Venugopalan 等人[13]首次提出基于编码-解码器的视频描述模型，该模型首先采用卷积神经网络作为编码器提取视频所有帧的视觉信息再做平均，然后通过长短期记忆网络将均值后的视觉特征生成自然语言描述语句。由于对帧级别的特征平均容易忽略视频目标的动作变化和帧的顺序。为了解决该问题，Wu 等人[14]也提出了一种轨迹结构化注意力编码器-解码器框架，结合注意力建模方案自适应学习视频的视觉内容中句子结构与运动目标的相关性，从而在解码阶段生成更准确、更细致的语句描述。Rothenpieler 等人[15]提出了一种新的双重方法。首先，利用一个奖励引导的 KL 分歧来训练一个对词排列具有弹性的视频字描述模型。其次，利用双模分层强化学习（BMHRL）转换器架构来捕获输入数据的长期时间依赖关系以作为分层描述模块的基础。Xu 等人[16]针对不同模态特征对生成描述重要性不同的情况，使用视频的帧、运动、音频三种不同类型特征进行有效融合来增强描述语句。并且为了有效融合特征，还设计了一个能决定融合时不同类型特征之间贡献度的融合单元。接着，Xu 等人[17]提出了一种新的具有多模态预训练的模块化设计，可以从模态协作中受益，同时解决模态纠缠问题。最近，Ghaderi 等人[18]还引入了自适应帧选择方案以减少所需的传入帧数，同时在训练两个转换器时保持相关内容。此外，还通过聚合每个样本的所有真实字描述来估计与视频描述相关的语义概念。虽然上述使用多模态特征能更完整地表达视频中丰富的视觉信息，但如何充分地融合多模态特征使得不同类型特征相互影响等问题还尚未解决。为了更好地解决该问题，Monfort 等人[19]提出了 500 万个口语字幕的口语时刻数据集，该数据描绘了各种不同的事件。Tanaka 等人[20]为了利用该数据集，他们还提出了一种新的自适应平均边际方法来对比学习，并评估了在多个数据集上的视频/描述检索模型。同时，等人还提出了一个用于电子竞技视频描述的大规模数据集，该新数据集提出了多个新的视频字幕挑战，例如大量特定于领域的词汇、具有重要意义的微妙动作以及大多数描述与发生的事件之间的时间差距。为了解决词汇问题，他们屏蔽特定领域的单词，并为此提供额外的注释，通过实验结果表明了该数据集对现有的视频描述方法构成了挑战，并且遮罩可以显着提高性能。

尽管现阶段的视频描述算法主要采用编码器-解码器架构，但许多研究者认为该结构并不能有效处理视频的丰富信息，并且会随着描述句子长度的增加，降低描述生成的质量。因此 Zhang 等人[21]提出一种基于双向时间图（OA-BTG）的对象感知聚合的视频描述方法，模型可以捕获视频中突出对象的详细时间动态，并通过对检测到的对象区域进行对象感知局部特征聚合来学习判别性时空表示。为了解决用自然语言描述视频序列的视觉内容的问题，Wang 等人[22]提出了一种具有重构器架构的重建网（RecNet），该网络利用双向流（视频到句子和句子到视频）方法对视频进行解析。基于非编码器-解码器的视频描述方法[23,24]的核心思想是通过引入额外的机制来改善视频描述的质量和流畅性，而不是直接从视觉特征中生成描述。

§1.3 本文研究目标和内容

本课题旨在基于深度学习来解析视频的语义内容，并使用不同语言进行描述。目前视频描述任务的大量研究都是针对英文描述方法，较少研究视频中文描述方法，主要是因为缺少中文训练样本。因此，本文先基于英文描述来研究视频语义内容解析，解决描述语义信息不完整和不准确的问题。接着，在视频英文内容描述方法的基础上，提出一种基于余弦注意力与语义选择的视频中文描述方法，以解决中文描述的语义冗余和语义不准确的问题。为了解决中文样本缺少的问题，将视频英文描述数据集 MSVD 扩展成中文数据集 MSVD-C。本文主要研究具体工作如下：

1.基于句子语义与长度损失计算的视频描述方法

在研究视频英文语义解析问题中，现有模型通常使用的 KL 散度损失函数可以通过学习特定的视觉内容特征来表示长度分布的单一映射关系，但无法在一对多的映射关系中准确预测句子的长度分布，导致描述语义信息不完整。其次，使用字符级别交叉熵函数计算损失，忽略了句子级别的语义，导致生成描述的语义信息不准确。首先，设计了一个新的长度损失函数，通过度量预测句长与参考句长的距离来自适应调整错误惩罚，使得模型可以在高度相似的视觉内容中学习到最优的描述长度分布。其次，通过解码层的语义隐藏状态计算，设计了一个基于句子语义的描述生成损失函数，通过对预测与参考描述在句子级别的语义比较，迭代获得最优的句子语义描述。并分别在 MSVD 和 MSR-VTT 数据集上测试，实验结果表明各项指标显著提升，均优于现有模型。

2.基于余弦注意力与语义选择的视频中文描述方法

在研究视频中文语义解析问题中，现有模型的自注意力机制使用查询和键矩阵的点积计算会导致 Softmax 函数出现饱和的问题。解码器使用高级视觉语义特征与句子语义特征融合来防止网络退化，但会导致大量冗余信息的产生，影响模型生成描述的语义准确性。首先，设计了一个缩放余弦注意力网络，先将查询和键矩阵进行余弦相似度计算，再通过可学习的参数进行放大，使得模型能自适应地关注正确的视觉语义特征，从而生成语义正确的描述。然后在解码阶段，设计了一种语义选择网络，过滤由视觉语义特征与句子语义特征融合产生的冗余信息，减少干扰，从而提升模型语义的准确性。将视频英文描述数据集 MSVD 扩展成中文数据集 MSVD-C，并在该数据集上进行实验，结果表明效果显著优于其他先进模型。

3.基于自适应特征选择与融合的视频中文描述方法

在研究视频中文语义解析问题中，现有视频中文描述模型通常使用线性变换方法将高维的视频特征映射到低维空间来减少特征的维度和冗余性，但会丢失重要信息。此外，将视觉和运动特征拼接可以获取更丰富的视觉语义信息，但会产生冗余信息。

首先，设计一种特征选择网络，先使用注意力机制关注重要的特征，再使用门控机制选择性地保留或丢弃特征。其次，设计一种自适应动态融合机制，通过计算运动特征的权重系数将视觉和运动特征向量进行动态融合，减少冗余信息的干扰，从而提升模型语义的准确性。本文在中文数据集 MSVD-C 上进行测试，结果表明本文方法显著优于其他先进模型。

§1.4 论文的结构与安排

本文包括五个章节，第一章为绪论部分，首先基于深度学习的视频语义解析的现实意义着手，对国内外研究现状进行介绍，为后面的视频描述任务的相关研究打下基础。为了生成语义信息完整和准确的英文描述，第二章提出了一种基于句子语义与长度损失计算的视频描述方法。为了生成准确的中文描述，第三章提出了一种基于余弦注意力与语义选择的视频中文描述方法，第四章提出了一种基于自适应特征选择与融合的视频中文描述方法。

第一章，绪论部分。首先介绍了网络短视频和视觉语义解析的现状，分析了视频语义解析研究的挑战和意义，进而探讨了视频描述算法的发展历程和不足。同时，针对该任务的技术难点提出了解决思路，并引出了本文的研究目标、内容和结构安排，以突显研究该课题的重要性。

第二章，基于句子语义与长度损失计算的视频描述方法。首先，对视频语义解析的相关技术进行研究，包括视频信息提取方法的基础、序列到序列模型、注意力机制等，为本文的第二、三和第四章的研究做铺垫。并对相关研究工作进行介绍。然后，针对现有方法存在的问题，提出基于句子语义与长度损失计算的视频描述方法。针对现有基于视频英文描述算法生成的描述不准确和不完整问题，设计了一种基于句子语义与长度损失计算的视频描述算法，并对算法流程框架进行了介绍。接着，详细介绍了该算法中的长度预测网络、长度损失函数以及双向自注意力的解码器的原理和结构。最后，再根据在数据集上的测试结果进行实验分析，证明该方法能提高生成描述准确性和完整性问题

第三章，基于余弦注意力与语义选择的视频中文描述方法。首先，针对视频中文描述算法生成的描述语义不准确问题，设计了一个基于余弦注意力与语义选择的视频中文描述算法，并对该算法的流程进行了详细的阐述。其次，详细介绍了缩放余弦注意力网络的设计、解码阶段的语义选择网络设计过程，最后将视频英文描述数据集 MSVD 扩展成中文数据集 MSVD-C，并在该数据集上进行实验，结果表明该算法效果显著优于其他先进模型。

第四章，基于自适应特征选择与融合的视频中文描述方法。首先，针对视频中文描述算法生成的描述语义不准确问题，设计一个基于自适应特征选择与融合的视频中

文描述算法，并对该算法的流程进行了详细地阐述。其次，介绍了特征选择网络、自适应动态融合机制的设计过程。最后，还在中文数据集 MSVD-C 上进行上测试，效果显著优于其他先进模型，其中 BLEU@4 和 CIDEr-D 指标提升尤为显著.

第五章 总结与展望。主要对基于深度学习的视频语义解析的研究内容进行总结，并分析了视频语义解析的研究内容中存在的不足和需要进一步完善的地方，并提出了未来工作的展望。

§1.5 本章小结

本章首先介绍了短视频和视频语义解析的现状，分析了视频语义解析研究的挑战和意义，进而探讨了视频描述算法的发展历程和不足。同时，提出了解决该任务技术难点的思路，并引出本文的研究目标、内容和结构安排。最后，绘制了如图 1-1 所示的总体安排结构图来清晰地展示章与章之间的关系。

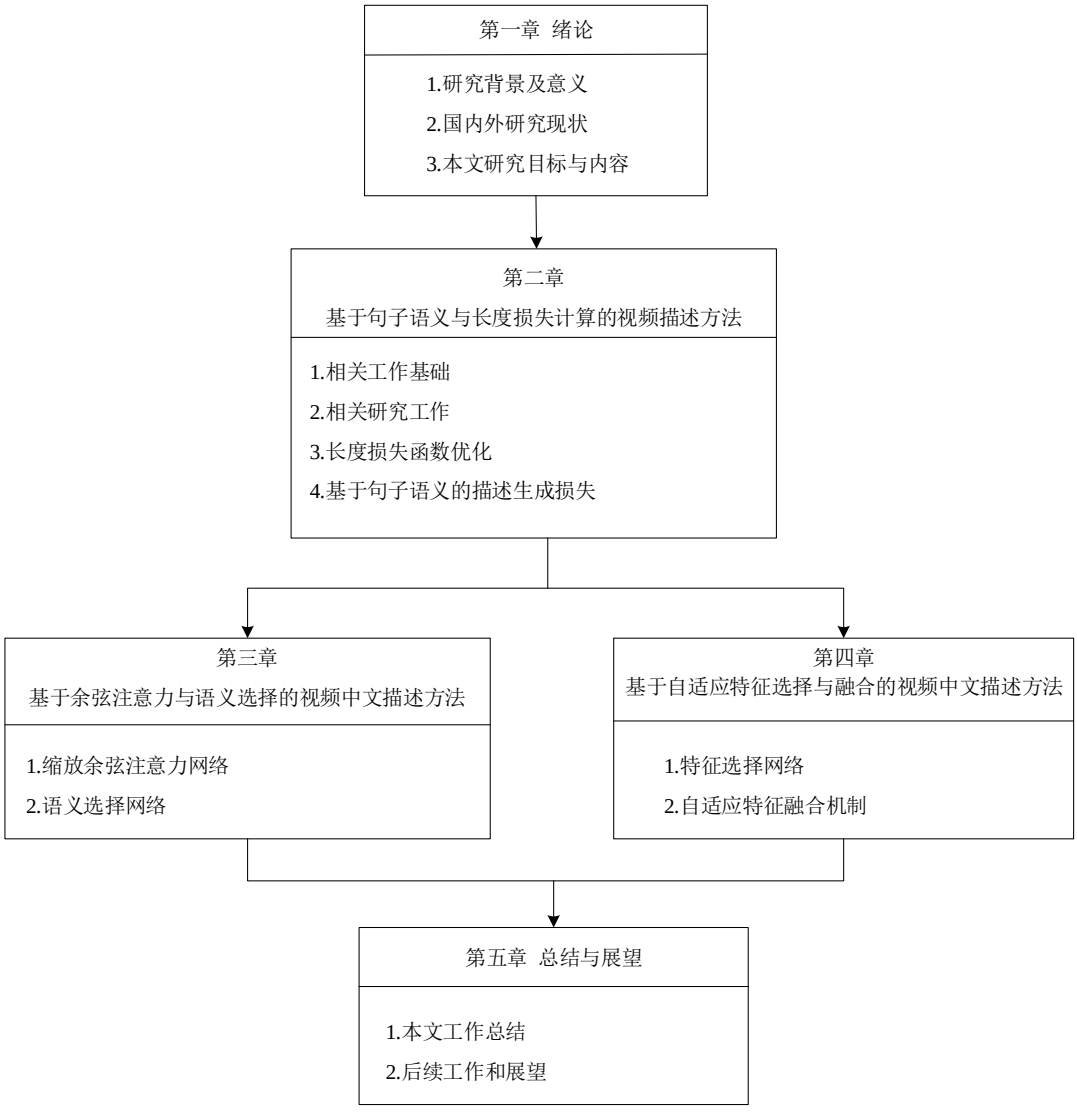

图 1-1 总体安排结构图

第二章 基于句子语义与长度损失计算的视频描述方法

视频描述任务是自然语言和计算机视觉感知两大领域的交叉融合课题，旨在用一句简短的句子自动描述视频中丰富的内容。随着深度学习技术的发展，基于深度学习的视频描述方法逐渐受到关注，但是现有的方法仍然存在以下问题，如使用善于学习单一映射关系的 KL 散度损失函数来对模型优化，在面对高度相似视觉内容的一对多映射关系时，会使模型无法准确预测句子的长度分布，导致生成描述内容不完整。此外，使用仅在单词级别的交叉熵函数计算句子语义，忽略了句子级别语义信息，导致模型生成描述的语义信息存在偏差。针对上述存在的问题，提出一种基于句子语义与长度损失计算的视频描述方法。首先，为了使模型可以在高度相似的视觉内容中学习到最优的描述长度分布，设计了一个新的长度损失函数，通过度量预测句长与参考句长的距离来自适应调整错误惩罚。其次，为了使模型准确学习句子语义信息，设计了一个基于句子语义的描述生成损失函数，通过对预测与参考描述在句子级别的语义比较，迭代获得最优的句子语义描述。

§2.1 相关工作基础

§2.1.1 视频特征提取方法

1.基于 2D-CNN 的特征提取

近年来，2D 卷积被广泛应用于目标识别、语义分割和图像分类等 2D 视觉任务[25,26]中。由于卷积神经网络[27,28]在图像处理任务中表现出色，常被用于提取视频单帧特征信息，例如 AlexNet[29]、VGG[30]。2D-CNN 单通道常指 2D-CNN 中输入数据只有单个通道的情况。在图像处理任务中，一般的彩色图像包含三个通道（红、绿、蓝），因此彩图的输入数据通常是三通道。但在有些情况下，输入数据可能只包含一个通道，比如灰度图像，这时候的输入数据就可以使用单通道的 2D-CNN 来处理。在深度学习中，卷积旨在提取有用的特征，而基于深度学习的卷积是通过将卷积核与输入信号进行元素相乘和累加的过程来实现的。图 2-1 展示了单通道的一次 2D 卷积过程，由图可知，左边是一个 4*4 的像素矩阵，中间是一个 3*3 的卷积核矩阵，如将中间的 3*3 卷积核矩阵分别乘左边的像素矩阵的不同区域，可获得右边的 2*2 矩阵的黄色区域。

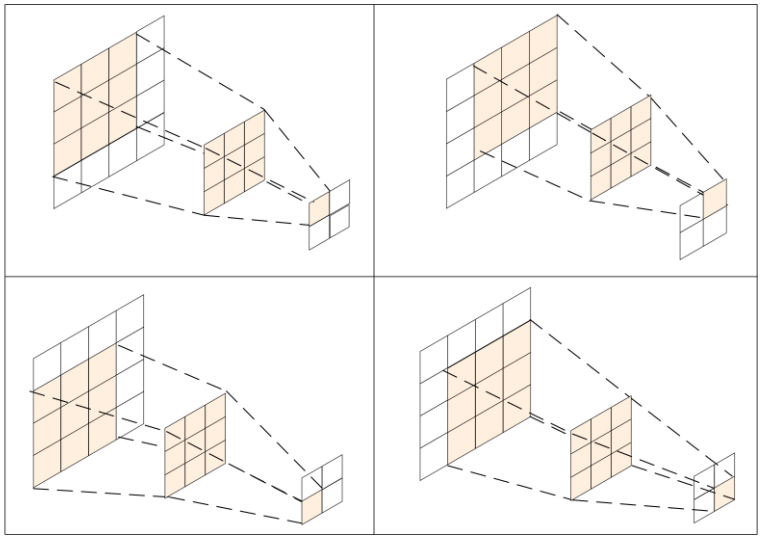

图 2-1 单通道 2D 卷积

2D-CNN 多通道是 2D-CNN 使用多个通道进行计算的情况，它通常被用于提取比单通道更丰富的特征信息。每个通道可以看作是一个不同的滤波器，它可以捕捉到图像中的不同特征[31]。通过使用多个通道，网络可以同时学习到多个特征，从而提高模型的准确性。因此，本文使用 2D-CNN 多通道提取运动特征。假如将一张 RGB 图像分为三个通道，每个通道使用不同的卷积核进行卷积计算，再将三个通道的结果合并起来，便可捕捉图像中更丰富的视觉语义信息，达到提高识别的准确率。下图 2-2 演示了多通道 2D 卷积运算过程：对于一个 5*5*3 的 RGB 图像输入，可以使用一个 3*3*3 的卷积核（filter）对其进行卷积，得到 3 个 3*3 的特征图，然后将这 3 个特征图逐元素相加得到一个单通道的特征图。

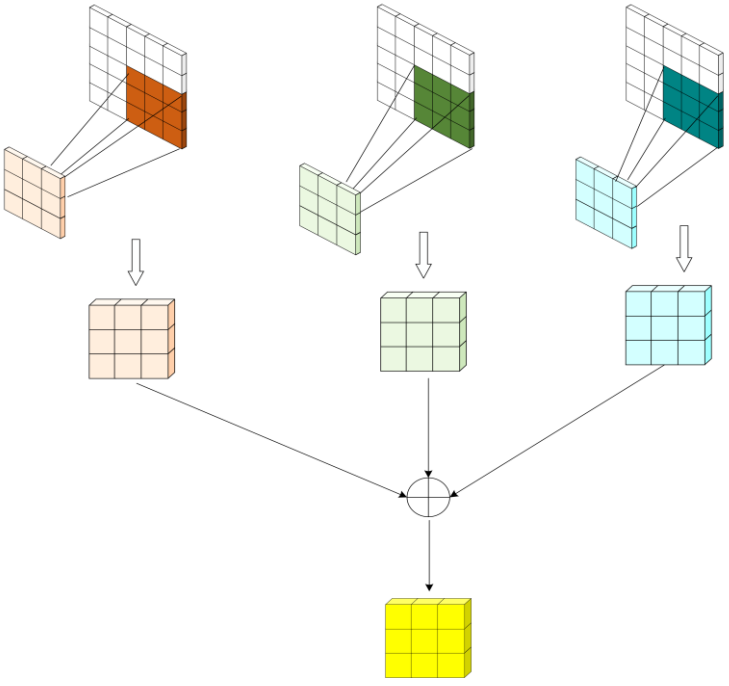

图 2-2 多通道 2D 卷积

2. 基于 3D-CNN 的特征提取

虽然 2D 卷积网络也能作用于多张图片，但需将多张图片变为多通道的单张图像再叠放在一起。叠放操作很容易忽略图像在事件维度的依赖关系，使模型难以有效的理解时空特征。因此，许多研究人员开始研究可以在视频等三维数据上进行特征提取和分类。在 3D-CNN 中，卷积核不仅考虑了空间上的信息，还考虑了时间上的信息，因此可以更好地捕捉视频中物体的运动信息[32]。3D-CNN 的特征提取一般分为输入数据预处理、卷积层提取特征、展平数据、全连接层提取特征等几个步骤。图 2-3 为单个通道的 3D 卷积过程，左侧为输入的视频像素，大小为 5*5*T，其中 T 表示时间维度大小。中间的立方体为 3*3 的卷积核，对输入的视频像素矩阵进行卷积，可以获得右侧输出的特征映射。与 2D-CNN 相比，3D 卷积能够更好地建模时间信息，在视频场景与行为的识别等任务上表现出了优异的效果。因此，本文使用 3D-CNN 提取运动特征。

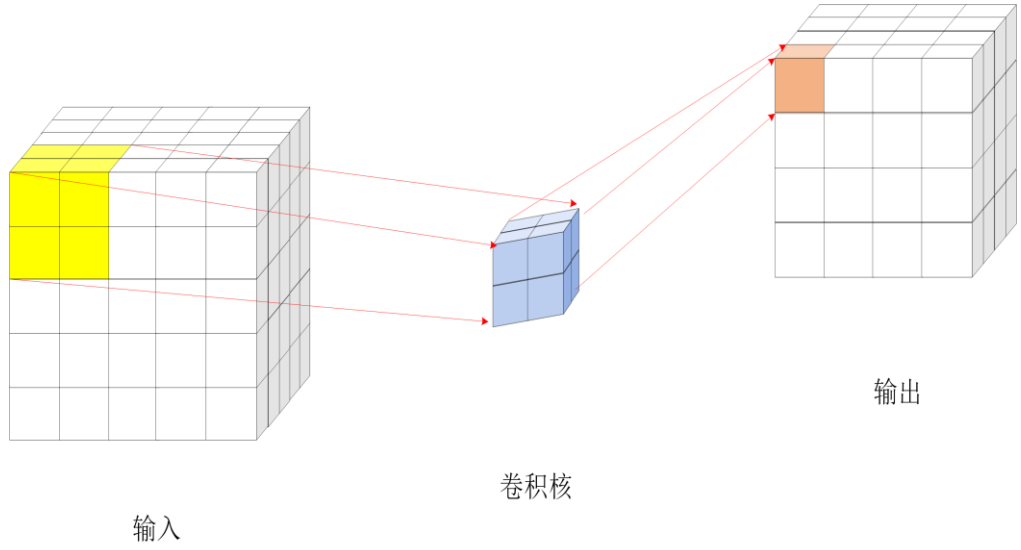

图 2-3 单通道 3D 卷积

§2.1.2 编码器-解码器框架

编码器-解码器模型是一种常见的深度学习模型，常被用于将输入数据映射到输出数据。编码器-解码器模型由两部分组成：编码器和解码器。编码器-解码器模型的基本工作原理是编码器将输入数据编码成低维度的特征表示，解码器将低维度特征表示解码为输出数据[33]。编码器通常采用循环神经网络（RNN）或卷积神经网络（CNN）对输入数据进行编码，而解码器则根据编码器生成的上下文信息和先前生成的输出序列逐步生成目标序列。图 2-4 为编码器-解码器框架。

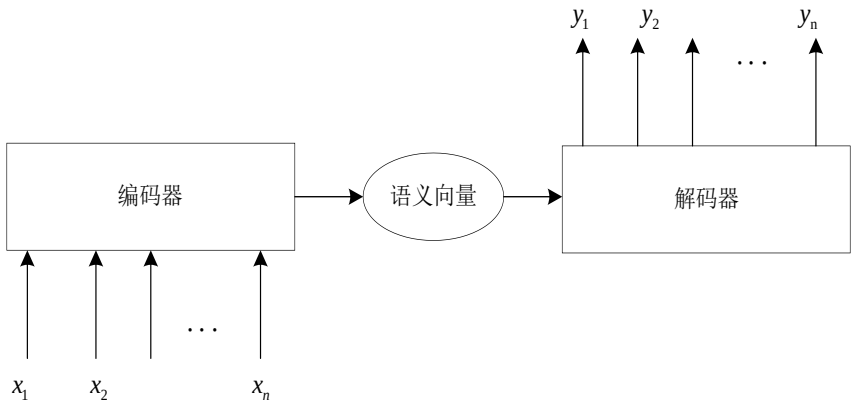

图 2-4 编码器-解码器框架

视频描述算法的编码器-解码器模型的基本工作原理是编码器将输入的视频数据进行处理，并将其映射成一个固定维度的向量表示，也称为“视觉语义向量”，而解码器则接收视觉语义向量，并将其解码为自然语言描述。解码器在每个时间步输出一个词，直到生成完整的描述。图 2-5 为视频描述算法的编码器-解码器模型的一般结构。

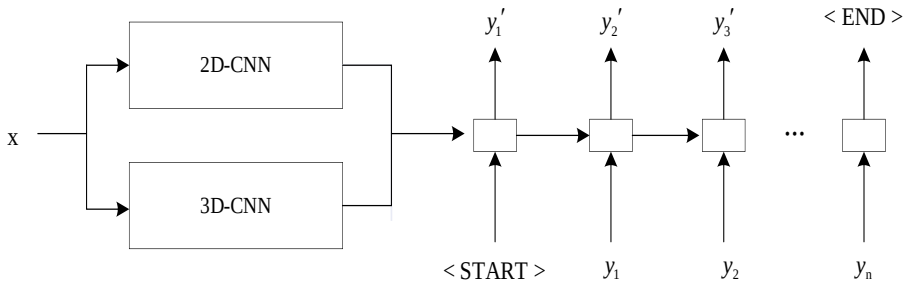

图 2-5 视频描述算法的编码器-解码器模型

由图 2-5 可以看出视频描述的整体过程，假设将一组给定的训练数据$<X,Y>$输入编码器 F，编码器分别提取视频的 2D 视觉特征和 3D 运动特征，再融合获得视觉语义特征 V，整个过程可以表示：

$$V=(F_{2D}(X):F_{3D}(X)) \tag{2-1}$$

其中，X 代表输入视频数据，：代表行拼接。

然后将视觉语义特征 V 输入解码器，编码器的第一个节点接受该特征并给出开始标志位<START>。在训练阶段，假设$Y=\{y_1,y_2,y_3...,y_n\}$为对应视频描述标签，n 为描述标签的序列长度，$Y'=\{y_1',y_2',y_3'...,y_n'\}$为预测输出的描述句子。解码器的节点$S_t$的输入为上一个节点$S_{t-1}$的隐层状态和真实序列在时刻 t-1 的值$y_{t-1}$，那么便可预测出该节点的输出$y_t'$。整个过程可以表示为：

$$h_t=\begin{cases}S(<START>,C),t=1\\S(y_{t-1},h_{t-1}),t\neq 1\end{cases} \tag{2-2}$$

$$y_t'=q(y_{t-1},h_t) \tag{2-3}$$

§2.1.3 注意力机制

注意力机制是基于人类的视觉系统，模仿人类在观察事物时将注意力集中在最为重要的部分，忽略掉不重要的信息。它的基本工作原理是通过对序列中每个元素进行加权，让模型注重关键信息来提高模型的准确性和性能。注意力机制在自然语言处理、语音识别、图像处理等领域都有广泛应用。在自然语言处理任务中，可以使用注意力机制来解决机器翻译、问答系统等任务；在语音识别任务中，可以使用注意力机制来提高语音识别的准确性；在图像处理任务中，可以使用注意力机制来识别图像中的物体，提高图像分类的准确性。

注意力机制基本原理如图 2-6 所示，常见的注意力机制包括“软”注意力机制和“硬”注意力机制。其中“硬”注意力机制则通过选择一个或多个输入特征来实现，而“软”注意力机制通过对每个输入特征的加权求和来实现，可以自适应地调节每个特征的权重，图 2-6 为“软”注意力机制，其计算整体计算过程如下：

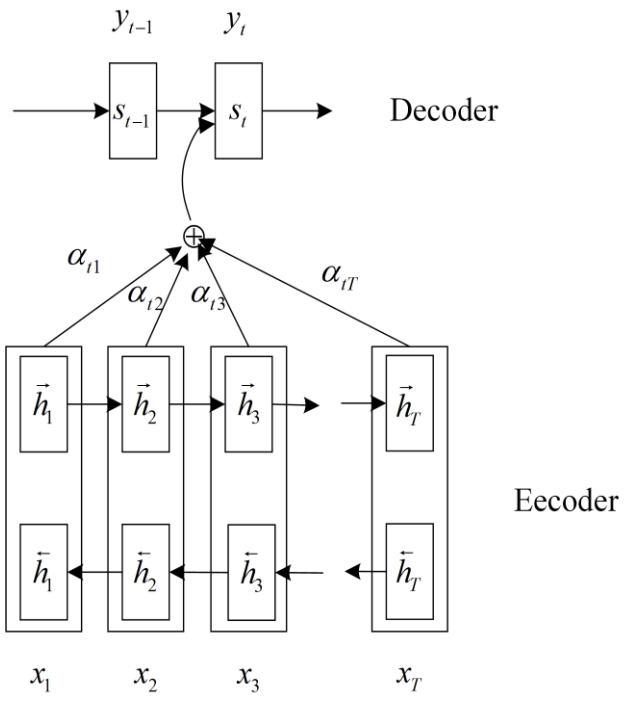

图 2-6 注意力机制

在视频描述任务中，假设$(x_1, x_2, ..., x_T)$为视频特征序列，那么预测第 i 个目标词y_i的条件概率为：

$$p(y_i \mid y_1, y_2, ..., y_{i-1}, x) = g(y_{i-1}, s_i, c_i) \tag{2-4}$$

其中，s_i为i时刻循环神经网络的隐藏状态信息，公式为：

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \tag{2-5}$$

其中，c_i 为编码后的视觉语义信息，它由整个视觉输入信息的编码端隐藏层状态 $(h_1,h_2,...,h_T)$ 决定。视频特征包含大量的冗余信息，因此需要使用注意力机制来获取重要特征，其计算过程公式为：

$$c_i = \sum_{j=1}^{T} \alpha_{ij} h_j \tag{2-6}$$

其中，α_{ij} 为隐藏状态 h_j 的注意力权重，可表示为：

$$\alpha_{\mathrm{ij}} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})} \tag{2-7}$$

近年，随着对注意力机制的研究不断深入，出现了许多不同的注意力机制的变种。2017 年，Google 机器翻译团队提出了包含自注意力机制[34]的 transformer 模型，与以往基于循环神经网络的序列到序列模型不同。它是由 6 个相同的网络构成，每层包括多头注意力机制和前向神经网络两个模块，模块间以残差网络的形式连接在一起。自注意力机制是多头注意力机制中的模块，是一种用于建模序列中不同元素之间相互关系的方法，它能够计算每个元素与序列中其他元素之间的相关性，并为每个元素分配一个权重，从而实现对序列的精细建模。自注意力机制可以看作是一种在输入序列中关注不同元素之间相互作用的方式。在自注意力机制中，每个元素通过查询(query)一系列键值(key-value)对进行交互。自注意力的结构如图 2-7。

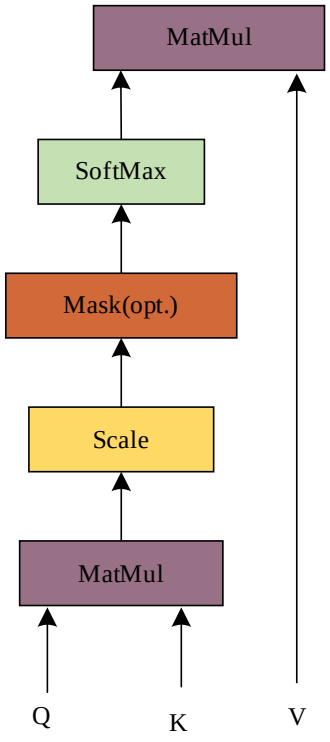

图 2-7 自注意力机制

§2.2 相关研究工作

近几年，基于深度学习的视频描述方法受到越来越多研究者的青睐，2016 年 Pan 等[35]首次将视频的视觉特征和语义特征映射到一个共同的嵌入空间。Pei 等人[36]通过训练数据来探索单词及各种视觉特征上下文之间的关系。Ryu 等人[37]利用解码后的描述对视频特征进行分组匹配，再结合未匹配的视频特征预测下一单词。Wang 等人[38]通过捕获语义上下文相关和视觉互补的信息来生成更好的描述。虽然上述方法能有效缩小视觉特征与高层语义之间的语义鸿沟，但是它们忽略了句子语法的表示。为了获得更加顺畅的描述，Wang 等人[39]将词性信息运用到视频描述任务，使用词性信息和视频片段的多重表示来生成描述内容。同年，Hou 等人[40]通过结合视觉表示和语法表示用于生成描述内容。此后，Zheng 等人[41]提出了一个语法感知的视频描述模型，学习视频中存在的动作及其主题，使其在描述中实现更好的语义一致性。

随着目标检测与动作识别技术的成熟，研究者们开始使用它们来获取更全面的视觉语义信息。2019 年 Aafaq 等人[42]通过短傅里叶变换捕获视频特征中的丰富时间动态，并从目标检测器中提取语义信息。接着，Tan 等人[43]利用时空注意力检测出所有目标，再利用帧间的差异信息进行动作推理。最后，Vaidya 等人[44]采用了一个自适应空间定位方法，不仅使模型聚焦于局部对象信息，而且减少了大量视频帧的时间冗余带来的时间和内存消耗。虽然这些方法提升了生成描述内容的完整性和语义的准确性，但它们的推理速度慢，需要耗费大量的时间和金钱成本。因此，Yang 等人[45]提出了一个基于非自回归解码模型，该模型设计了一种生成视觉词的机制，由于视觉词能决定描述内容语义的正确性，所以 NACF 模型在一定程度上解决了视频描述推理速度和描述质量不平衡的问题。

§2.3 基于句子语义与长度损失计算的视频描述模型

§2.3.1 视频描述模型的整体框架

视频描述数据集由 N 个数据样本组成，其中第 i 个数据样本表示为 (X_i, Y_i, B_i)，X_i 表示第 i 个视频，Y_i 表示第 i 个视频对应的参考描述，B_i 表示第 i 个视频对应的参考描述的单词。假设一个视频有 f 帧，其中对应 Y_i 的由 m 个参考描述组成，B_i 由 k 个单词组成，则有：

$$X_i = \{x_1^i, x_2^i, ..., x_f^i\} \tag{2-8}$$

$$Y_{\mathrm{i}} = \{y_1^i, y_2^i, ..., y_k^i\} \tag{2-9}$$

$$B_i = \{b_1^i, b_2^i, ..., b_k^i\} \tag{2-10}$$

本章提出的基于句子语义与长度损失计算的视频描述方法，简称 SSLL（Sentence Semantics and Length Loss）。整体框架如图 2-8 所示，主要由视频编码器（Encoder）、长度预测网络(Length Prediction Network)和基于双向自注意力（Bidirectional Self-attention）的解码器（Decoder）组成。首先，SSLL 模型利用视频编码器对视频进行编码，得到富有高级语义信息的特征表示，长度预测网络根据编码网络输出的特征去预测句子长度分布。然后，基于双向自注意力的解码器依据长度分布生成文本描述。最后，SSLL 模型再计算生成的文本描述和真实描述之间的差距，通过反复迭代获得最优的句子语义描述。在视频编码器中，首先采用已预训练好的 2D/3D 卷积 RetNet-101 模型[46]来提取富有高级语义信息的视觉和运动特征，然后将其进行融合来获得含有更丰富的语义信息视觉特征 V。

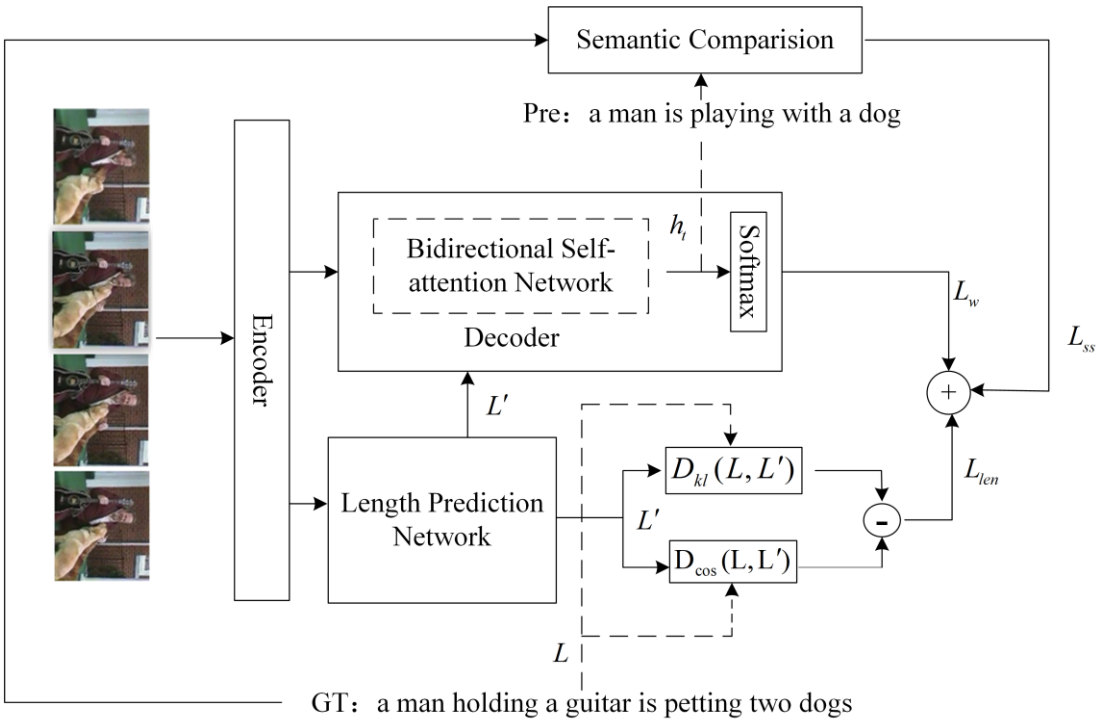

图 2-8 SSLL 模型的整体框架图

§2.3.2 长度预测网络

1.长度预测网络框架设计

为了解决自回归模型生成速度慢以及容易累积错误词的问题，本章采用了基于非自回归的视频描述生成方式[45]。与自回归解码模型根据预测句子最后一个词来确定句子长度不同，非自回归解码模型使用长度预测网络辅助生成描述。长度预测网络可以将视频的内容和复杂度转换为可预测的长度信息，然后将长度信息传递给解码器。解

码器一次性生成整个描述句子，避免了自回归模型中逐步生成导致的速度慢和错误累积词的问题。因此，将来自编码器输出特征V输入长度预测网络以获得预测句子长度分布L，计算过程如下所示：

$$L' = \text{Softmax}(\text{ReLU}(\text{MP}(V)W_{21})W_{22}) \tag{2-11}$$

其中，MP 为平均池化，ReLU 为激活函数，Softmax 为激活函数，$W_{21} \in R^{d_m \times d_m}$ 和 $W_{22} \in R^{d_m \times N_{\max}}$ 为权重参数。

2.长度损失函数优化设计

在视频描述任务中，现有模型通常使用擅长学习单一映射关系的 KL 散度损失函数来获得最优的描述长度分布。当每个输入具有单个正确参考时，即当每个视觉高度相似的视频都对应一个正确的参考长度分布时，KL 散度具有直观的正确推导，它将最大限度地提高单个正确参考的概率，使模型很容易学习到分布规律。但当每个输入具有多个正确参考时，KL 散度损失函数难以使模型收敛。因为 KL 散度损失函数要求模型为所有潜在的参考答案分配高概率，就无法从一对多的映射关系中学习到正确的长度分布规律。这就是 KL 散度损失函数作为对数损失具有的局限性。

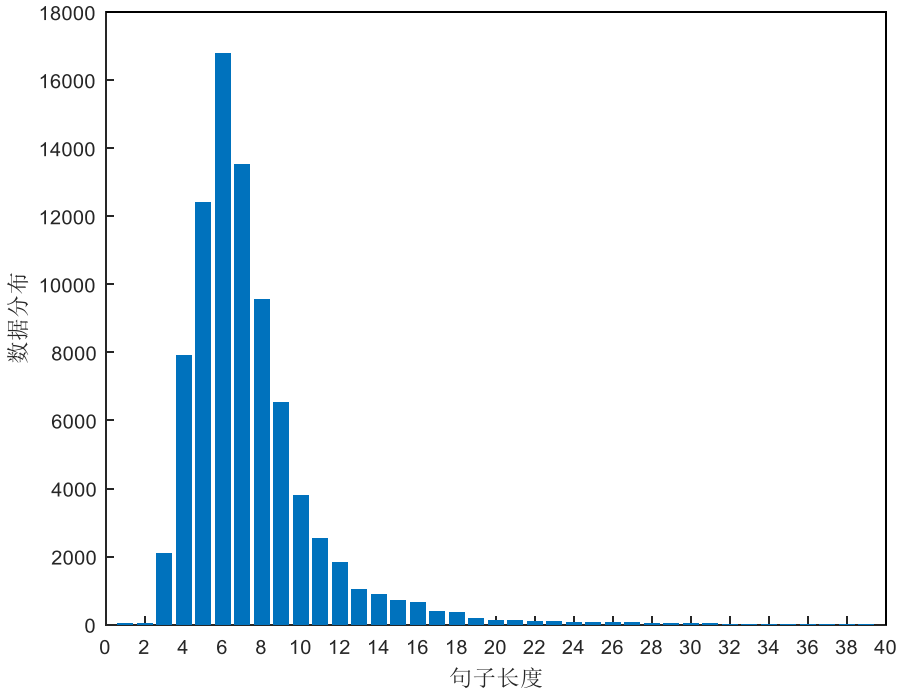

图 2-9 MSVD 数据集句子长度分布情况

如图 2-9 所示是 MSVD 数据集参考描述的长度分布情况，由图 2-9 可知句子长度分布范围为 1~40 字长。字长 1~3 的标签（如 man、a man、a man is）是由于人为的疏忽导致的标签噪音。这些标签噪音会扰乱长度分布的规律，从而进一步加大模型学习的难度。为了避免极端值或异常值对模型学习产生过度影响，Kang 等人[47]提出了截断损失。余弦损失被证明能够有效地度量生成的句子长度与期望长度之间的相似度

或距离。因此，受此启发提出了一个新的长度损失函数(Length Loss，LL)，该函数将余弦损失作为一种截断损失函数，通过惩罚过长或过短的句子长度，使模型能够在高度相似的视觉内容中学习最优的描述长度分布，从而提升生成描述语义信息的完整性。具体计算如下所示：

$$L_{len} = D_{KL}(L' \| L) - \beta \cos(L', L) = -\sum_{j=1}^{N_{\max}} l'_j \log \frac{l_j}{l'_j} - \beta \sum_{j=1}^{N_{\max}} \frac{l_j l'_j}{| l'_j \| l_j |} \tag{2-12}$$

其中，L 和 L' 分别为预测和参考描述的长度分布，l_j 为某个视频对应的参考描述长度为 j 的句子占的百分比，$N_{\max}$ 为最大序列长度，β 为调整损失值下降的比例系数，β 过大会使模型无法得到有效的训练，过小会使模型无法调整错误的惩罚。

当同一空间中的两个样本分布相差较大时，KL 散度损失值会较大，余弦损失函数给予的惩罚也越大；当同一空间中的两个样本分布相差较小时，KL 散度损失值会较小，余弦损失函数给予的惩罚也越小。因此，当面对一对多映射关系的样本时，模型会倾向学习出现概率最高的样本。从图 2-9 可知，MSVD 数据集的参考描述出现概率较高的句子长为度 5~8 字长，所以模型会更倾向生成字长为 5~8 的描述。

§2.3.3 基于双向自注意力的解码器

为了获得最终生成的描述，本文采用了文献[45]的解码框架。首先，将预测长度 L' 和编码后的视觉特征 V 输入到双向自注意网络来并行生成一个仅有视觉词（名词或者动词）的句子模板，然后再基于句子模板和编码后的视觉特征并行生成非视觉词来获得完整的句子模板。最后为了使模型获得更加流畅的描述，将获得的完整句子模板中评分较低的单词（不合适的词）输入基于双向自注意力的解码器，经过五次迭代替换为评分较高的单词。具体的计算过程如下：

$$p_\theta(\boldsymbol{y} | V, L) = F_{decoder}(V, L') \tag{2-13}$$

其中，$\boldsymbol{y}$ 代表生成描述，$F_{decoder}$ 代表解码器。

§2.3.4 基于句子语义的描述生成损失函数设计

在视频描述任务中，现有视频模型通常使用单词级别的交叉熵损失函数来学习句子单词的语序信息，容易忽略句子语义信息，因此模型会倾向生成句式相似但语义内容迥然不同的描述。Li 等人[48]为了缓解交叉熵损失函数的不足提出了句子级别损失。因此，在此启发下设计一个新的考虑句子语义的损失函数（Sentence Semantics，SS），该函数基于解码器的语义隐藏状态 h_t 获得预测的句子级别语义分数，然后通过对预测与参考描述在句子语义比较，迭代获得最优的句子语义描述。

为了获得句子级别的分数，需先从解码器层获取语义隐藏状态 h_t，再利用语义隐藏状态 h_t 在所有位置生成单词级得分 s_t，然后对 s_t 求和以获得句子级别语义分数，最后使用 Sigmoid 函数进行归一化以获得每个单词在句子中出现的概率 p_s， p_s 与句子的位置无关，计算过程如公式 2-14、2-15 和 2-16 所示：

$$P_s = \text{Sigmoid}(\sum_{t=1}^{K} s_t) \tag{2-14}$$

$$s_t = W_g h_t + b_g \tag{2-15}$$

其中，K 为参考描述的词数量，W_g 和 b_g 为可学习训练参数。

模型训练时，通过最小化生成句子和目标句子之间的词级损失来优化模型，使生成的句子和目标句子之间的语义相似度尽可能高。

$$L_{ss} = -\sum_{i=1}^{K} b_i \log p_s(b_i) \tag{2-16}$$

其中，K 为参考描述的词数量，b_i 目标句子的第 i 个单词。

由于 L_{ss} 关注的是整个句子的语义相似性未考虑词语之间的顺序和语义信息，而交叉熵损失强调的是每个单词的正确性，因此，本章还设计一个融合损失函数解决句子级别和单词级别损失函数的不足。如公式（2-18）所示，该损失函数将句子语义的损失函数 L_{ss} 与交叉熵损失函数 L_w 融合，使模型能充分考虑句子级别和单词级别的语义信息，进而更全面地评估模型的性能。两种损失结合除了可以使模型学习到更准确、更具有语义一致性的视频描述外，还可以缓解交叉熵损失中单词缺失对整个句子损失的影响，从而提高模型的鲁棒性。所以融合损失函数的整体设计如公式 2-17、2-18 和 2-19 所示：

$$L_{gen} = L_w + \alpha_i L_{ss} \tag{2-17}$$

$$L_w = -\sum_{y \in Y^{vis}} \log(y|V, L') - \sum_{y \in Y_{mask}} \log(y|Y_{obs}, V) \tag{2-18}$$

$$\alpha_i = \min(\alpha, k + \lambda i) \tag{2-19}$$

其中，Y^{vis} 代表生成的视觉词，Y_{obs} 和 Y_{mask} 分别代表模板未掩盖和迭代过程中被掩盖的词。其中 α_i 表示在第 i 个 epoch 平衡两个损失函数的系数，在训练初始阶段先分配一个较小的权重，然后根据 epoch 逐渐增大权重。在实验中，根据验证集将 α、k 和 λ 设置 1.0、0.1、0.1。

因此，基于句子语义与长度损失计算的视频描述模型的整体损失函数如下：

$$L_{SLESS} = L_{len} + L_{gen} \tag{2-20}$$

其中，L_{len} 为长度损失函数，L_{gen} 为融合损失函数。

§2.4 实验结果与分析

§2.4.1 实验数据与参数设置

1.数据集

本章分别通过在视频描述任务的两个常用数据集 MSVD 和 MSR-VTT 进行实验来验证该方法的有效性。前者收集了从 YouTube 网站上下载的 1970 个短视频，包含动物、玩耍和烹饪等主题，大部分视频时长在 9~12 秒之间，每个视频大约包含 28 到 60 条英文描述，句子长度集中分布在 6 到 12 个单词，总词汇表为 9562 个单词。训练、验证和测试集分别为 1200、100 和 670 个视频。后者共有 10000 个短视频，每个视频拥有 20 条人工标注的参考描述、类别标签和音频信息，大部分视频的时长在 15 秒左右，总词汇表为 23271 个单词。其中训练、验证和测试集分别为 6513、497 和 2990 个视频。如图 2-10 展示了 MSVD 和 MSR-VTT 两个数据集的部分视频及对应参考描述。其中图片部分为视频片段中的 3 帧，文字部分为视频对应的 4 个描述。

2.参数设置

本章实验设置如下：MSVD 和 MSR-VTT 数据集的最大序列长度分别设置为 20 和 30 字长。解码器隐藏层单元为 512，将每个模态设置为 8。Adam 优化器以 0.005 的初始学习速率分批训练 64 个视频，迭代次数 epoch 设置为 50，dropout 进行正则化，参数设置为 0.5，L2 权重衰减设置为 0.0005，训练模型时 batch size 设置为 128，测试模型时 batch size 设置为 32。

GT1:a man cuts a piece of paper

GT2:a man is cutting a paper by scissor

GT3:the man cut a piece of paper in half

GT4:the man is using scissors to cut paper

.....

GT1:a man is draining liquid out of a plastic container

GT2:a man is draining water from some cooked pasta

GT3:a man is draining pasta water into the sink

GT4:a man is emptying water from a container into the kitchen sink

……

GT1:the man is playing the piano

GT2:a man is playing a white grand piano

GT2:the man is playing music on the piano

GT4:a mand played on a white piano

……

图 2-10 MSVD 和 MS-VTT 数据集示例

3.评价指标

由于 SSLL 方法解决的语义信息完整性和准确性问题与句子长度、描述的语序和语义的正确性有关，因此本文采用视频描述生成任务的 BLEU@4[49]、METEOR[50]、ROUGE-L[51]和 CIDEr-D[52]四种评价指标作为衡量标准。

BLEU@4 是基于精度的相似度度量方法，用于分析目标与预测句子中的 1-4 元组词的重合度和衡量预测长度与标签长度的关系。计算公式如下：

$$BLEU=\mathrm{BP}*\exp(\sum_{n=1}^{N}W_{\mathrm{n}}*\log P_{\mathrm{n}}) \tag{2-21}$$

其中，BP 代表惩罚因子，P_{n} 代表 n-gram 的精准度，W_{n} 代表 n-gram 的权重。n-gram 是用于分析自然语言文本中的连续 n 个单词或字符序列，n 表示选择的连续单词或字符序列的长度。例如，本文采用 BLEU@4，即代表使用 4-gram(n=4)来计算准确度。

METEOR 是一种自动评估机器翻译（MT）和自然语言生成（NLG）系统的评价指标，它采用了一种基于单字召回率和精度的加权平均值的度量方法。METEOR 还

考虑了同义词、同义词组、相同词干和词根等因素，以确定预测句子和参考描述词与词组的匹配程度，并通过最小化块的数量准则来完成预测句子和标签的对齐。METEOR 是机器翻译和自然语言生成领域广泛使用的评价指标之一。与 BLEU 相比，它综合考虑了自动评估中最重要的两个因素：翻译的准确性和流畅性。该度量基于一元语法的精度和查全率的调和平均值 F_{mean}：

$$F_{\text{mean}}=\frac{10PR}{R+9P} \tag{2-22}$$

为了能够考虑更长的匹配，METEOR 方法将 unigram（单个词）拆分成尽可能少的块，块是由在候选描述和参考描述中相邻的 unigrams 组成的一组词。这样可以更准确地确定预测句子和参考描述之间的匹配程度。METEOR 会对更定的排列一个惩罚算方式如下：

$$\text{Penalty}=0.5*(\frac{chunk}{unigram\_matched}) \tag{2-23}$$

其中，unigram_matched 代表 unigrams 被映射的数量，chunk 是指块的数量。最后可以得到：

$$METEOR=F_{\text{mean}}*（1-\text{penalty}） \tag{2-24}$$

ROUGE-L 是基于精度的相似度度量方法，它衡量自动生成的预测描述与参考描述之间最长公共子序列的长度，然后将其除以参考描述的总长度，以此作为自动生成的摘要的质量得分。计算过程如下：

$$R_{\text{lcs}}=\max\frac{\text{lcs}(\text{c}_i,\text{s}_{i,j})}{\left|s_{i,j}\right|} \tag{2-25}$$

$$P_{\text{lcs}}=\max\frac{\text{lcs}(\text{c}_i,\text{s}_{i,j})}{\left|\text{c}_i\right|} \tag{2-26}$$

$$ROUGE_L(c_i,s)=\frac{(1+\beta^2)R_{lcs}p_{lcs}}{R_{lcs}+\beta^2P_{lcs}} \tag{2-27}$$

其中，$\text{lcs}(\text{c}_i,\text{s}_{i,j})$ 代表给定视频的标注语句与模型生成的描述语句之间最长公共子序列的长度，R_{lcs} 召回率，P_{lcs} 为 $\text{lcs}(\text{c}_i,\text{s}_{i,j})$ 的精确率，β 为需要设定的参数。

CIDEr 是一种用于评估自然语言描述生成模型的指标，它是对 BLEU 和 ROUGE 等指标的改进和扩展。CIDEr-D 在 CIDEr 的基础上增加了一个多样化采样的步骤，旨在减少模型生成的重复描述。CIDEr-D 在自然语言描述生成领域得到了广泛应用，尤其是在图像描述生成、视频描述生成等任务中。

§2.4.2 模型性能比较

为了验证提出的 SSLL 方法的有效性，本章将其与近三年先进的视频描述模型分别在 MSVD 和 MSR-VTT 数据集上进行了对比。实验结果如表 2-1 和 2-2 所示，其中最佳结果用黑色字突出显示。GRU-EVE[42]通过短傅里叶变换捕捉时空动态，POS-CG[41]将语句中的单词词性构成的词性序列作为规范句子的生成模板，MARN[36]利用内存网络捕获跨视频内容，SAAT[40]在每个解码步骤预测一组简洁的属性，SGN[37]利用外部语言模型解决长尾问题，NACF[45]利用非自回归模型生成由粗到细的模板。由于本章是基于 NACF 的框架来验证基于句子语义与长度损失计算的视频描述模型的有效性，因此选用 NACF 作为参考模型进行比较。由表 2-1 可知 SSLL 方法相较参考模型 NACF 均有提升，在 MSVD 数据集上的 BLEU@4、METEOR 和 ROUGE-L 指标分别提升了 2.51%、1.38%和 0.82%，在 MSR-VTT 数据集上的 BLEU@4、METEOR 和 ROUGE-L 指标分别提升了 0.95%、1.39%、0.97%和 0.94%。从历年的先进模型的性能提升趋势可以看出，SSLL 方法对于提升描述内容语义信息的完整性和准确性是十分有效的。

表 2-1 在 MSVD 测试集上 SSLL 模型和其他模型的对比结果

模型	BLEU@4	METEOR	ROUGE-L	CIDEr-D
GRU-EVE	47.9	35.0	71.5	78.1
POS-CG	53.9	34.9	72.1	88.7
MARN	48.6	35.1	71.9	92.2
SAAT	46.5	33.5	69.4	81.0
SGN	52.8	35.5	72.9	94.3
NACF	55.6	36.2	73.1	96.3
SSLL（ours）	**57.0**	**36.7**	**73.7**	**96.3**

表 2-2 在 MSR-VTT 测试集上 SSLL 模型和其他模型的对比结果

模型	BLEU@4	METEOR	ROUGE-L	CIDEr-D
GRU-EVE	38.3	28.4	60.7	48.1
POS-CG	42.0	28.2	61.6	48.7
MARN	40.4	28.1	60.7	47.1
SAAT	39.9	27.7	61.2	49.1
SGN	40.8	28.3	60.8	49.5
NACF	42.0	28.2	61.6	51.4
SSLL（ours）	**42.2**	**29.0**	**62.2**	**51.9**

§2.4.3 消融实验

1.SSLL 中 SS 和 LL 的有效性

如表 2-3 和表 2-4 所示，本文通过控制变量法分别在 MSVD 和 MSR-VTT 数据集上进行消融实验，验证文中所提组件句子长度损失函数（LL）和基于句子语义的描述生成损失函数（SS）的有效性。在 MSVD 数据集上测试时，LL 模型的 CIDEr-D 指标下降，是因为 LL 模型虽然能够生成更完整的句子并提供更多的描述信息，但是生成的语义信息不一定正确。CIDEr-D 指标是基于 n-gram 匹配来评估描述和参考描述之间的相似度，包括描述中的词汇和短语。因此，如果 LL 能够使模型生成的描述更加完整，但同时生成的语义信息不正确，那么这些描述可能会包含更多的 n-gram 不匹配，进而导致 CIDEr-D 指标评分下降。LL 在 MSR-VTT 数据集的各个指标都优于基线 NACF 模型，LL 在 BLEU@4 和 ROUGE-L 指标提升尤为明显，意味着调整句子长度使得生成描述的语义信息更完整。SS 在 MSVD 和 MSR-VTT 数据集的各个指标都优于基线 NACF 模型，而 SS 在各个指标得到提升，意味着它通过纠正语义内容使生成描述的更准确。为了进一步考察上述指标的提升在实际应用中的效果，将 LL 和 SS 与 NACF 模型的描述生成结果进行对比分析，后者的 100 条测试样本生成的结果大约有 60 条样本存在语义信息不完整和不准确问题，引入本文所提升的 SS 和 LL 方法后分别有 6 和 8 条样本获得明显改进。提升幅度大约达 10%和 13.3%，实际应用的效果得到显著提升。部分实例可参考图 2-11。

表 2-3 在 MSVD 数据集上的 LL 与 SS 方法有效性对比

模型	BLEU@4	METEOR	ROUGE-L	CIDEr-D
NACF	55.6	36.2	73.1	96.3
-SS	55.8	36.3	73.4	94.1
-LL	56.1	36.5	73.5	96.4
SSLL	57.0	36.7	73.7	96.3

表 2-4 在 MSR-VTT 数据集上的 LL 与 SS 方法有效性对比

模型	BLEU@4	METEOR	ROUGE-L	CIDEr-D
NACF	41.7	28.7	61.6	51.4
-SS	42.0	28.9	62.1	51.8
-LL	42.0	28.8	62.1	51.5
SSLL	42.1	29.1	62.2	51.9

GT: a man is playing with a dog

NACF: a man is plying

LL: a man is sitting on a char

SSLL: a man is plying with a dog

图 2-11 实例分析(a)

GT: a man holding a guitar is petting two dogs

NACF: a dog is plying

LL: a dog is plying with a dog

SSLL: a man is plying with a dog

图 2-11 实例分析(b)

GT: a woman is mixing ingredients in a bowl

NACF: a woman is cooking

LL: a woman is mixing ingredients in a bowl

SSLL: a woman is mixing flour in a bowl

图 2-11 实例分析(c)

图 2-11 展示了 MSVD 和 MSRVTT 数据集的部分视频和分别使用 NACF、LL 和 SSLL 模型测试生成的描述结果，其中 GT 代表参考描述。从下面的图 2-11 实例分析(a)可以看出，NACF 模型生成的描述会缺少 with a dog，引入 LL 模型后能使生成描述增加了 on a chair，SSLL 模型的生成结果增加了 with a dog，说明 SS 能更正了 LL

生成不准确的语义信息。从图 2-11 的实例分析可以看出，LL 模型的生成描述句子的成分比 NACF 模型生成描述句子的成分更完整长度，证明 LL 能使预测的句子长度更接近参考标签的句子长度，从而使生成描述的语义信息更加完整。SSLL 是在 LL 的基础上加入纠正句子语义信息的 SS，实例表明两个方法一起可以使得生成的描述内容和语义信息更完整和更准确。从图 2-11 实例分析可以看出现有模型在 MSVD 数据集上测试更倾向生成 4 字长的描述，说明 LL 可以使模型生成语义信息更完整的描述。

2.LL 中采样比例系数对模型性能的影响

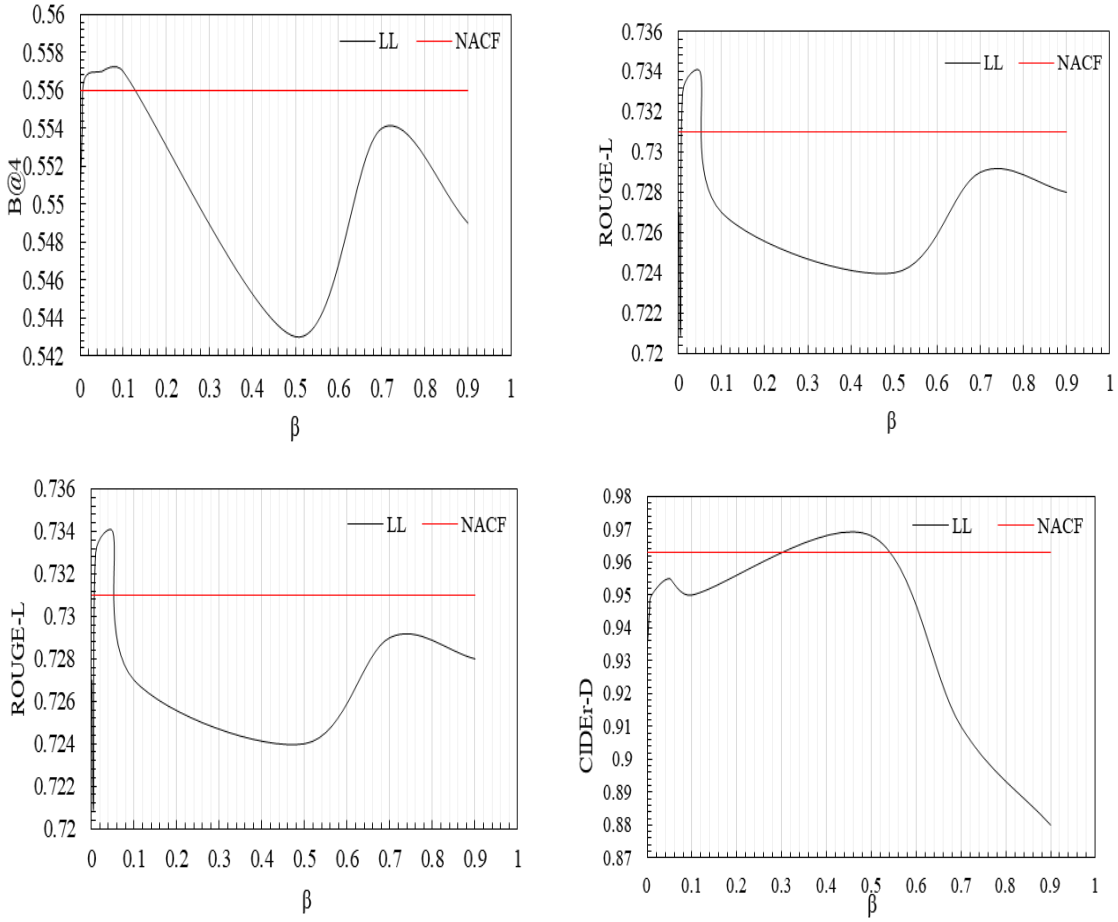

图 2-12 在 MSVD 数据集上不同的采样比例系数 β 对模型性能的影响

本节通过设置超参数 β 的不同权重比例系数来比较句子长度损失（LL）优化前后的表现，从而验证优化后的句子长度损失的有效性。图 2-12 和 2-13 分别展示了设置不同的参数 β 对评估的指标 BLEU@4、METEOR、ROUGE-L 和 CIDEr-D 产生的影响。从图 2-12 看出，在 MSVD 数据集的超参数 β 调整至 0.01~0.1 左右时，LL 模型的性能优于 NACF 模型的性能，其中调至 0.05 时句子长度损失优化函数的性能最好。从图 2-13 看出，对于 MSR-VTT 数据集将超参数 β 调整至 0.0005~0.1 左右时，LL 模型的性能优于 NACF 模型的性能，而调至 0.01 时句子长度损失优化达到最好的表现。从图 2-12 以及图 2-13 的实验结果证明了引入新的句子长度损失（LL）提高模型生成描述的准确性。

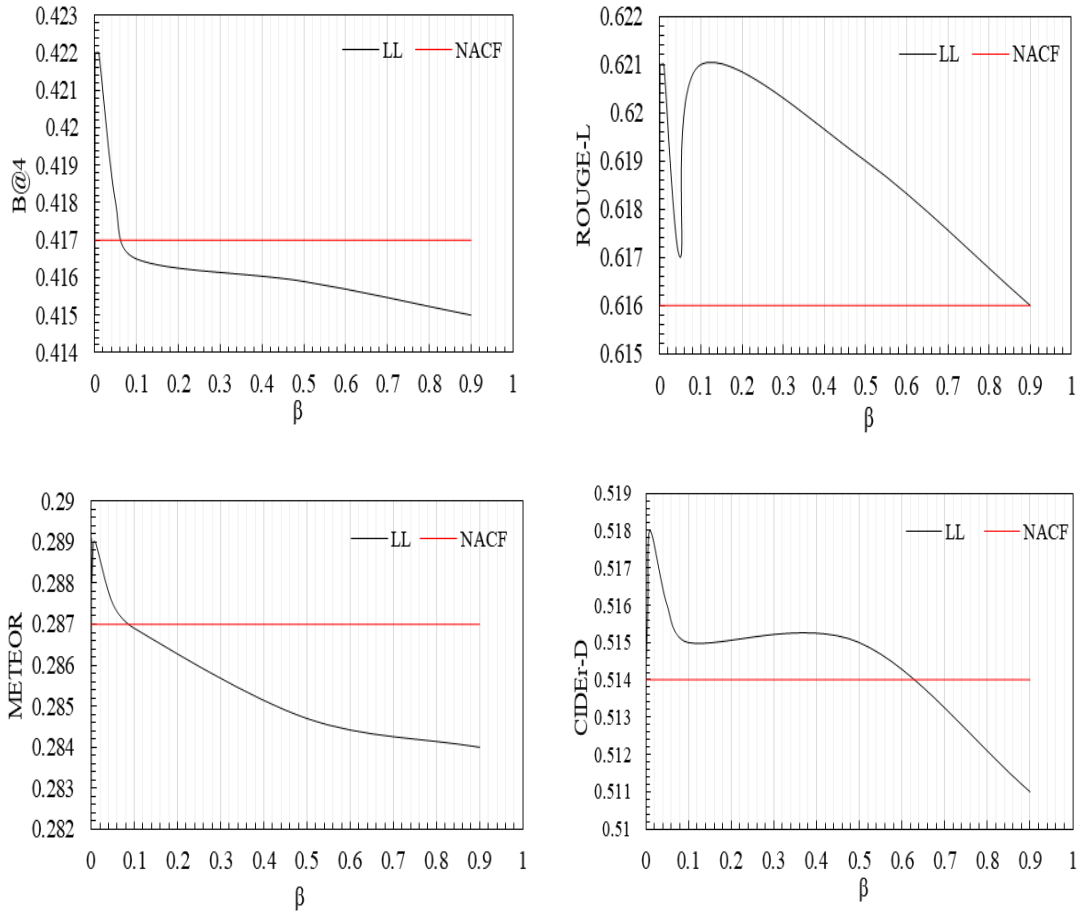

图 2-13 在 MSR-VTT 数据集上不同的采样比例系数 β 对模型性能的影响

3.LL 模型与 NACF 的句长比较

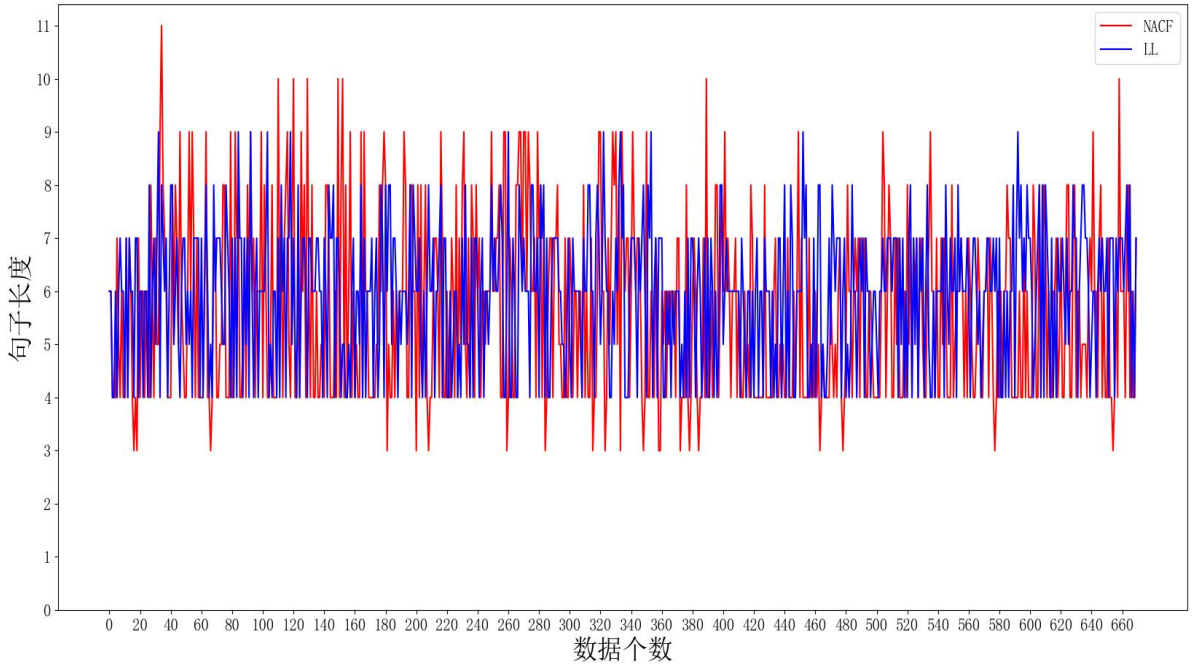

图 2-14 NACF 和 LL 模型的句长对比图

为了更直观、更准确地验证 LL 模型能倾向学习出现概率较高的样本，本节将对 NACF 和 LL 模型在 MSVD 数据集上进行测试获得的生成描述进行对比实验。图 2-14 反映的是 NACF 和 LL 模型的每个测试样本分别对应生成的每一条描述句长对比图。其中如图 2-14 纵轴代表句子长度分布，其范围 1~11 字长代表 LL 模型与 NACF 生成的句子描述的范围，横纵是数据个数，代表 LL 模型与 NACF 生成的总句子描述。由图 2-14 可知引入新的长度损失后，LL 模型比对比 NACF 模型生成的描述更集中在 5~8 句长，进一步验证了 LL 模型能更倾向生成句长出现概率较高的样本。图 2-15 展示了 NACF 和 LL 模型分别使用 MSVD 测试数据集获得的生成描述的句长统计。该图能直观地证明了引入 LL 模型后，生成的句子长度能从 4 和 9~10 句长慢慢移至 5~8 句长，说明 LL 模型更倾向生成句长出现概率较高的样本。

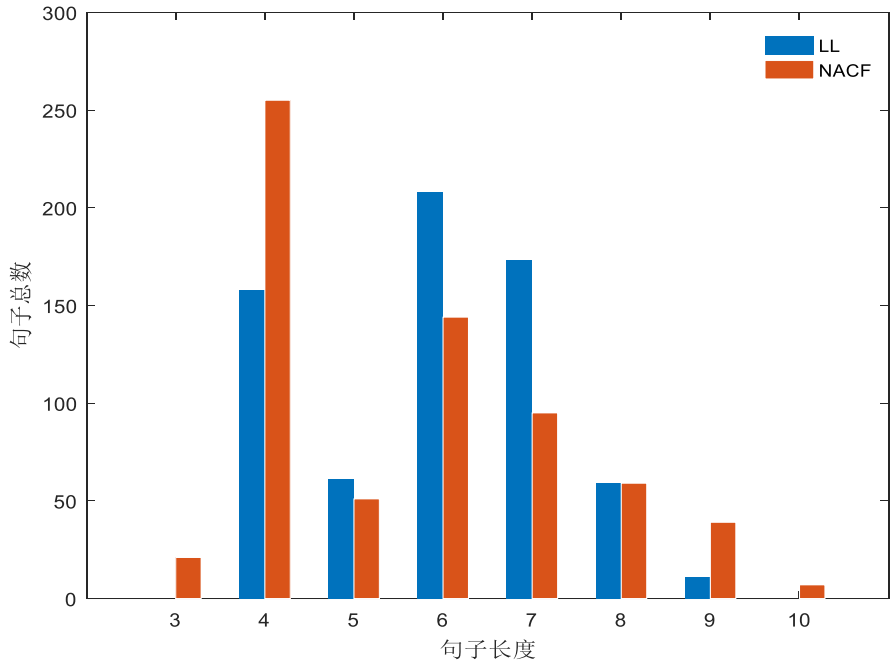

图 2-15 NACF 和 LL 模型的句长分布对比图

§2.5 本章小结

本章设计了一种基于句子语义与长度损失计算的视频描述的方法 SSLL，在 MSVD 和 MSR-VTT 数据集上的实验结果表明了 SSLL 方法的有效性，它的生成描述质量明显优于目前先进的模型 NACF。从消融实验的分析结果来看，本章所提的长度损失函数能提升描述内容语义信息的完整性，所提的句子语义损失能提升模型的准确性。但本章算法仅对视频英文描述进行研究分析，可能会忽略其他语言使用者的需求，导致算法的应用范围受限。因此为了满足不同应用场景的需求以及提高算法的适应性，本团队在之后的工作会进一步研究视频的中文描述方法。

第三章 基于余弦注意力与语义选择的视频中文描述方法

随着互联网和数字设备的普及，视频数据的增长速度呈现爆炸式增长，视频内容的自动描述变得越来越重要。视频描述可以为用户提供更好的搜索体验和更直观的视频内容理解。然而，视频描述任务仍然存在以下问题：（1）现有模型的自注意力机制使用查询和键矩阵的点积计算会导致 Softmax 函数出现饱和。（2）信息间的融合会导致大量冗余信息的产生，影响模型生成描述的语义准确性。另外，视频描述任务中的中文描述方法因缺乏训练样本而受到限制，因此大量的研究工作都是基于英文描述方法进行的，而对于视频中文描述的研究相对较少。一种复杂的视频中文描述方法是先生成英文描述，然后进行机器翻译，即将英文描述翻译成中文描述。该方法主要存在两个方面的缺陷：（1）复杂的视频中文描述方法会消耗大量不必要的时间和成本。（2）由于模型不是直接向着目标进行学习，间接方法会积累视频英文描述和机器翻译任务的损失。针对上述问题，提出一种基于余弦注意力与语义选择的视频中文描述方法。首先，新设计了一个缩放余弦注意力网络，先将查询和键矩阵进行余弦相似度计算，再通过可学习的参数进行放大，使得模型能自适应地关注正确的视觉语义特征，从而生成语义正确的描述。其次，在解码阶段，设计了一种语义选择网络，过滤由视觉语义特征与句子语义特征融合产生的冗余信息，减少干扰，从而提升模型语义的准确性。最后，还将视频英文描述数据集 MSVD 扩展成中文数据集 MSVD-C，并在该数据集上进行实验表明，效果显著优于其他先进模型。

§3.1 相关研究工作

近年，研究者们针对视频描述任务提出了各种模型，主要分为基于模板和基于深度学习两种方法。2017 年，谷歌团队提出的 transformer[34]网络结构，在各个领域都取得了巨大成就，也为该任务提供了新的研究思路。其强大的功能主要归因于该网络的自注意力机制，它是一种将单个序列的不同位置关联起来来计算同一序列表示的机制。2020 年，Iashin 等[53]人利用多头自注意力模块来融合两个序列特征，以产生双模态序列使模型生成更好的描述质量。接着，Li 等[54]人利用自注意力机制将输入文本和某一图像区域联系起来缩小文本与图像之间的语义鸿沟。生成高质量的描述不仅要巧妙地使用自注意力机制，还要恰当地利用视觉语义特征。Chen 等人[55]提出一种视觉特征的空间信息进行提取和聚合的方法，证明了全局聚合也可以为生成描述提供适当的语义上下文。Ramanishka 等人[56]提出一种多模态视频描述方法，这种方法是将视觉、声音以及表示视频主题特征的信息源进行融合以获得准确的文本描述。2021

年，Yang 等人[45]提出了一个非自回归粗到细（NACF）的模型，不仅引入了多头自注意力机制来筛选准确的视觉语义特征，还将其生成的句子语义信息与高级视觉语义信息融合来防止网络退化。近年，为了满足不同应用场景的需求以及提高算法的适应性，视频中文描述方法开始受到了研究者的关注。2021 年，侯等人[57]提出了首个端到端的视频中文描述方法，利用知识蒸馏方法试图将英文描述的高级语义信息融入中文描述以提高中文描述的质量，但未充分利用视觉语义信息，导致生成描述不够准确。此外，中文训练样本缺少，也增加了视频中文描述研究的难度[58]。

§3.2 基于余弦注意力与语义选择的视频中文描述模型的设计

§3.2.1 视频描述模型的整体框架

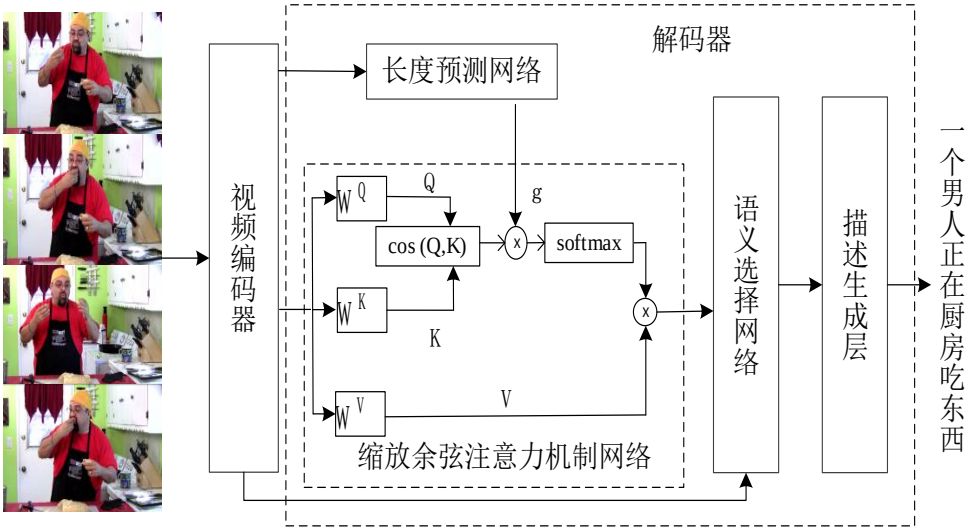

图 3-1 SCA-SS 模型的整体框架

本章提出的基于余弦注意力与语义选择的视频中文描述方法，简称 SCA-SS（Scaling Cosine Attention and Semantic Selection）。SCA-SS 整体框架如图 3-1 所示，该模型由视频编码器和解码器两部分组成。视频编码器通过将由 N 帧组成的视频片段输入到已预训练的 2D/3D-CNN 模型获得视觉特征V_a和运动特征V_m，再将这两个特征进行融合以获得更加丰富的视觉语义信息F。解码器主要由缩放余弦注意力网络、长度预测网络、语义选择网络和文本生成层构成。首先，SCA-SS 模型利用视频编码器对视频进行编码，得到富有高级语义信息的特征表示，长度预测网络根据编码网络的特征预测句子长度分布。然后，解码器的缩放余弦注意力网络先通过余弦相似度计算出高级视觉语义信息之间的联系，再根据句子长度分布值适当地放大该联系来计算出注意力权重值，最后根据注意力权重选择出关键的高级视觉语义特征。接着，语义选择网络通过选择门选择关键的高级视觉语义与句子语义特征，并利用更新门实

现动态地语义融合，从而使模型能准确地过滤冗余信息。最后，利用描述生成层生成输出的预测句子。

为了获取句子长度分布，本文参考文献[45]设计了长度预测网络，根据给定的编码器输出特征 F 来预测句子长度分布 L，计算过程如下所示：

$$L = \text{Softmax}(\text{ReLU}(\text{MP}(F)W_{31})W_{32}) \tag{3-1}$$

其中，MP 为平均池化，ReLU 和 Softmax 为激活函数，$W_{31} \in R^{d_m \times d_m}$ 和 $W_{32} \in R^{d_m \times N_{\max}}$ 为权重参数。

§3.2.2 缩放余弦注意力网络

Henry 等人[59]证明了自注意力机制在查询和键矩阵相乘之前沿着其头部应用归一化，再通过一个可学习的参数放大，能使 Softmax 不容易出现饱和。因此本文受此启发设计了一种缩放余弦自注意力网络（Scaling Cosine Attention,SCA）。首先可使模型对视频编码后的高级视觉语义特征 F 分别进行线性变换获得查询 Q、键 K、值 V 矩阵，线性变换矩阵用 W^Q、W^K、W^V 表示。

$$Q = FW^Q, K = FW^K, V = FW^V \tag{3-2}$$

由于点积计算的值是无限的，容易造成 Softmax 出现饱和的现象。而余弦相似度不仅能将值限定在[-1，1]之间，还比点积更容易发现两个向量的微小差距。因此，缩放余弦自注意力网络使用余弦相似度对查询和键矩阵进行计算，再通过一个可学习的训练参数 g 将其进行放大，从而使模型自适应地调整信息之间的权重，避免 Softmax 出现饱和的问题。

$$f = g\cos(Q, K^T) \tag{3-3}$$

$$g = \text{Sigmoid}(W_{13}\log(L^2 - L)) \tag{3-4}$$

其中，K^T 为键矩阵 K 的转置，W_{13} 为可训练的参数矩阵，L 为描述的预测句子长度，sigmoid 为激活函数。

最后将 f 进行 Softmax 函数计算以获得可自适应调整的权重，再将计算出来的权重值作用于值矩阵 V，最终使模型能准确地关注到高级视觉语义特征。

$$F_{SCA} = \text{Softmax}(f)V \tag{3-5}$$

§3.2.3 语义选择网络

Parisitto 等人[60]的研究证明，在 transformer 的子模块中的关键点添加新的门控机制，能够使模型更快、更准确地过滤冗余信息。受此启发，本文在解码器的缩放余弦注意力网络后添加了一个语义选择网络（Semantic Selection，SS），能有效地过滤由视觉语义特征与句子语义特征融合产生的冗余信息，减少干扰，从而提升模型语义的准确性。整个过程如下：

GRU 中的门控机制可以有选择地记住和遗忘一些特征信息，语义选择网络也通过类似的门控机制来达到动态滤除冗余信息源的效果。首先从视觉编码器获取高级视觉语义特征 F，然后从缩放余弦注意力网络获取句子语义特征 F_{SCA}，最后将获取到的高级视觉语义特征 F 和句子语义特征 F_{SCA} 分别经过激活函数 ReLU 得到 F' 和 F'_{SCA}。

$$F' = \mathrm{ReLU}(F) \tag{3-6}$$

$$F'_{SCA} = \mathrm{ReLU}(F_{SCA}) \tag{3-7}$$

接着，将 F' 和 F'_{SCA} 特征一起输入选择门 r 和更新门 z 以计算出候选隐藏状态 $\hat{h}$，再通过候选隐藏状态 $\hat{h}$ 与更新门 z 实现 F' 和 F'_{SCA} 特征的动态融合。选择门 r 能选择关键的 F' 和 F'_{SCA} 特征，更新门 z 可以根据 F' 特征抽取与 F'_{SCA} 特征相关的特征信息。

$$r = \delta(W_r F' + U_r F'_{SCA}) \tag{3-8}$$

$$z = \sigma(W_z F' + U_z F'_{SCA} - b_g) \tag{3-9}$$

$$\hat{h} = \tanh(W_g F' + U_g (r \odot F'_{SCA})) \tag{3-10}$$

$$F_{SS}(F', F_{SCA}) = (1-z) \odot F' + z \odot \hat{h} \tag{3-11}$$

其中，tanh、ReLU 为激活函数，σ 为激活函数 Sigmoid，W_r、U_r、W_z、U_z、W_g、U_g 为权重矩阵，b_g 为偏置，⊙表示向量点乘。

§3.3 实验结果与分析

§3.3.1 MSVD-C 中文数据集设置

由于视频中文描述方法的模型训练需要使用中文描述数据集，因此将视频英文描述数据集 MSVD 扩展成中文数据集 MSVD-C。整个中文数据集直接使用 MSVD 数据集的 1970 个短视频，其中包含动物、玩耍和烹饪等主题，大部分视频在 9~12 秒之间。

中文数据集 MSVD-C 对应的中文标签是由 20 个受过良好教育且普通话标准的研究生参考 MSVD 英文数据集进行标注。为了减少噪音对模型的干扰，MSVD-C 中文数据集不仅剔除了 MSVD 英文数据集成分不完整以及与视频内容无关的描述，还在整理时对描述进行语言纠错和标准化处理。整理后，中文数据集 MSVD-C 对应的视频大约包含 20~30 条描述，整个数据集大约包含 5 万条中文描述。在实验中，训练、验证和测试集分别为 1200、100 和 670 个视频。

GT1:一个男孩在弹钢琴

GT2:一个男孩坐在房间里弹钢琴

GT3:一个男孩坐在房间里，借着灯光弹钢琴

GT4:演奏音乐的男孩

GT5:一个男孩在演奏音乐

GT6:一个男孩坐着弹钢琴

.....

GT1:两个女人正对着麦克风唱歌

GT2:两个女人在录音棚里唱歌

GT3:两位年轻的女人开心地唱卡拉 ok

GT4:两个女人站在一起，手里拿着麦克风唱歌

GT5:两个女人在卡拉 ok 机前唱歌

GT6:两位年轻的女士在唱歌

.....

图 3-2 中文数据集 MSVD-C 实例

数据集预处理部分。关于视觉特征本章选用在 ImageNet 数据集上进行预训练获得的 2D 卷积 ResNet-101 模型，对每个视频提取 16 帧作为 2D 视觉特征。关于运动特征选用在 ImageNet 数据集上进行预训练获得的 3D 卷积 ResNet-101 模型，对每个视频提取 60 帧作为 3D 运动特征。为了避免以字为单位组合的词表导致部分词语义不完整的问题，在训练前，使用 jieba 中文分词工具对已经整理和清洗的描述进行词性分词，从而构建出一个由字词混合组成新的词表，总词汇表共计 8000 个词。图 3-2 为中文数据集 MSVD-C 标注的部分视频及其参考描述。

§3.3.2 参数与参考指标设置

参数设置方面，训练时将中文数据集 MSVD-C 的最大输入序列长度设置为 20 词长。解码器分别采用 1 个解码器层、512 个模型维度、2048 个隐藏维度和每层 8 个注意力头。Adam 优化器以 0.005 的初始学习速率分批训练 64 个视频，epoch 设置为 50，dropout 设置为 0.5，L2 权重衰减是 0.0005，测试时模型时 batch size 设置为 32。

评价指标方面，为了验证 SCA-SS 方法的有效性，本章使用视频描述方法常用的四个评价指标：BLEU@4、METEOR、ROUGE-L 和 CIDEr-D。

§3.3.3 模型性能比较

表 3-1 SCA-SS 模型和其他模型的对比结果

模型	BLUE@4	METEOR	ROUGE-L	CIDEr-D
NIC-KD	21.3	21.6	47.6	42.1
S2VT-KD	21.1	21.6	48.0	46.8
ConvCapp-KD	19.1	21.5	47.0	42.2
Top-Down-KD	21.1	22.9	49.4	57.4
NACF	37.2	38.0	64.2	73.8
SSLL	39.5	38.3	65.2	75.9
SCA-SS	**44.2**	**39.3**	**66.8**	**88.4**

如表 3-1 所示，本章选择了较为先进的模型在中文数据集 MSVD-C 上进行实验对比来验证 SCA-SS 方法的有效性，这些模型可以分为两类：（1）NIC-KD、S2VT-KD、ConvCapp-KD 和 Top-Down-KD 模型，这四种模型分别由四种经典视频/图像描述方法 NIC[61]、S2VT[62]、ConvCapp[63]和 Top-Down[64]引入了跨语言知识蒸馏的视频中文描述[57]方法获得。跨语言知识蒸馏[57]的视频中文描述方法是目前最先进的视频中文描述方法，也是唯一一种研究视频中文描述的方法，与这四种模型进行对比实验能验

证本文方法的先进性。（2）NACF[45]和 SSLL 模型，其中基于句子语义与长度损失计算的视频描述模型 SSLL 和本章所提方法 SCA-SS 都是基于非自回归的视频描述方法 NACF 的基础上进行研究，与它们进行对比实验可以更直接地反映 SCA-SS 的有效性。

实验结果如表 3-1 所示，其中最佳结果用黑色字突出显示。可以看出基于余弦注意力与语义选择的视频中文描述方法 SCA-SS 相较与目前的先进模型 NACF 拥有大幅度提升，其中视频描述常用的指标 BLUE@4、METEOR、ROUGE-L、CIDEr-D 分别提升了 18.8%、3.1%、4.0%、19.7%，而与 SSLL 模型相比，BLUE@4、METEOR、ROUGE-L、CIDEr-D 指标分别提升了 11.8%、2.6%、2.4%、16.4%，其提升效果同样十分显著，说明所提出的 SCA-SS 方法能够有效避免 Softmax 函数出现饱和以及减少冗余的信息的干扰，从而提升模型描述语义的准确性。

表 3-2 间接翻译模型和端到端生成中文描述模型的对比结果

模型	BLUE@4	METEOR	ROUGE-L	CIDEr-D
NIC-T	11.1	17.7	41.7	16.4
NIC-KD	21.3	21.6	47.6	42.1
S2VT-T	9.3	16.4	39.2	18.9
S2VT-KD	21.1	21.6	48.0	46.8
ConvCapp-T	7.1	16.4	39.0	18.3
ConvCapp-KD	19.1	21.5	47.0	42.2
Top-Down-T	9.7	17.0	39.6	23.0
Top-Down-KD	21.1	22.9	49.4	57.4
NACF-T	37.2	38.0	62.4	73.7
NACF	37.2	38.0	64.2	73.8
SSLL-T	38.6	38.1	65.0	75.9
SSLL	39.5	38.3	65.2	75.9

为了验证端到端的中文生成方法优于从英文描述翻译成中文描述的间接方法，本文将间接方法与端到端的中文描述进行对比实验。在实验间接方法时，利用 MSVD 英文数据集分别训练获得的 NIC、S2VT、ConvCapp、Top-Down、NACF 和 SSLL 模型，再使用对应的测试集测试获得的英文描述，并利用谷歌翻译器将生成的英文描述转化为中文描述，结果如表 3-2 所示，NIC-T、S2VT-T、Top-Down-T、NACF-T、SSLL-T 代表的是利用机器翻译成中文的结果。从表 3-2 可以看出，与直接端到端的中文描述方法相比，间接翻译方法的各个评价指标更低，表明前者生成的描述语义更准确且结构信息更完整。导致这个情况的主要原因可能是由于模型不是直接向着目标进行学习，间接方法会积累了视频英文描述任务和机器翻译任务的损失，从而影响最后描述生成的质量。表 3-2 的实验充分说明了将英文数据集 MSVD 扩展成中文数据集

MSVD-C 不仅为更多研究视频中文描述任务的研究者提供便利，还促进了视频中文描述任务的技术发展。

§3.3.4 消融实验

如表 3-3 所示，本节通过在 MSVD-C 中文数据集上进行消融实验来验证所提组件缩放余弦注意力网络（SCA）和语义选择网络（SS）的有效性。由表 3-3 可知，SCA 和 SS 在 MSVD-C 中文数据集测试的各个指标都优于 NACF 模型，意味着 SS 能够过滤掉冗余信息，SCA 能够关注到正确的视觉语义特征，这两个模型都能通过纠正语义内容使模型生成语义更准确的描述。为了进一步考察上述指标的提升在实际应用中的效果，还将 SCA 和 SS 模型的描述生成结果与 NACF 模型的描述生成结果进行对比分析，后者的 100 条测试样本生成的结果大约有 70 条样本存在语义信息不准确问题，引入本文所提升的 SS 和 SCA 方法后分别有 17 条和 13 条样本获得明显改进。提升幅度大约达 24%和 18%，实际应用的效果得到显著提升。部分实例参见图 3-3。

表 3-3 SCA 与 SS 方法有效性对比

模型	BLUE@4	METEOR	ROUGE-L	CIDEr-D
NACF	37.2	38.0	64.2	73.8
SCA-SS	44.2	39.3	66.8	88.4
-SS	41.5	38.6	64.4	79.2
-SCA	42.7	38.8	66.3	80.5

图 3-3 展示了中文数据集 MSVD-C 的部分视频和分别使用 NACF、SCA、SS 和 SCA-SS 模型测试生成的描述结果，其中 GT 代表参考描述。通过将 NACF 模型作为基线模型进行对比，可以看出 NACF 模型生成的描述与视频内容的语义不匹配。与 NACF 模型生成的描述进行对比可以更直观地反映不同模型的有效性。例如实例分析（a）中参考描述的目标对象是“男人”和“狗”，而基线 NACF 模型生成的描述的生成目标对象是“男人”和“女人”。这可能是因为 NACF 模型的解码器存在大量的冗余信息，以及自注意力机制未能准确地关注到正确的视觉语义特征，导致模型生成描述语义的不准确。对 SS 和 SCA 模型生成的描述结果进行详细分析，可知引入 SS 和 SCA 模型后，描述生成的语义信息得到纠正，证明了 SS 能够过滤掉冗余信息，SCA 能够关注到正确的视觉语义特征。SCA-SS 模型是由 SS 和 SCA 模型融合而成的模型。从图 3-3 的两个实例分析可知，SCA-SS 模型的生成描述的语义信息与参考描述的语义信息是基本一致的，说明了该模型不仅能够避免冗余信息对模型产生干扰，还能够准确地关注到正确的视觉语义特征。

GT:一只黑色狮子狗正在和一个男人击掌

NACF:一个男人在看一个女人

SCA:一个男孩和狗玩

SS:一个男人和狗玩

SCA-SS: 一个男人和狗玩

图 3-3 实例分析（a）

GT:一个女孩在做体操时摔倒了

NACF:一个人正在跑道上跑步

SCA:一个女孩摔倒

SS:一个女孩摔倒在平衡木上

SCA-SS:一个女孩摔倒在平衡木上

图 3-3 实例分析（b）

§3.4 本章小结

本章提出了一种基于余弦注意力与语义选择的视频中文描述方法，提升了生成描述语义信息的准确性，制作了一个中文数据集 MSVD-C 不仅为更多研究视频中文描述任务的研究者提供便利，还促进了视频中文描述任务的技术发展。通过在中文数据集 MSVD-C 上进行实验，实验结果表明基于余弦注意力与语义选择的视频中文描述方法的各项性能指标显著提升，明显优于目前先进的模型。通过消融实验证明了缩放余弦注意力网络能够准确地关注视觉语义特征，语义选择网络能有效滤除冗余信息对模型的干扰。然而，相对于英文视频描述，当前的中文视频描述的技术水平还有提升

空间。在视频描述领域，英文视频描述的研究和应用较早，相关技术和数据集也较为丰富，而中文视频描述的研究还比较新鲜，相关技术和数据集还需要进一步完善和拓展。因此，后续研究将继续深入探索视频中文描述任务，旨在进一步提升中午生成描述语义准确性。

第四章 基于自适应特征选择与融合的视频中文描述方法

在当前人工智能技术的发展下，视频理解和自然语言处理领域受到了广泛的关注。其中，视频描述任务是将视频中的视觉信息转化为相应的自然语言描述，对于视频内容的自动理解和搜索具有重要的实际应用价值。视频描述的编码器需要从高维视频数据中提取低维特征，以便于后续的描述生成。虽然线性变换能够将高维特征压缩成紧凑的低维表示，但静态的降维方法会丢失重要信息，无法处理数据的非线性相关性。因此，如何使用动态的降维方法来避免信息丢失是视频描述任务的难点之一。此外，由于多模态特征能更完整地表达视频信息，越来越多的研究者开始使用多模态特征来提高生成质量。但如何充分融合不同类型的特征，并处理它们之间的相互影响等问题仍需要进一步探索。视频描述任务中的中文描述方法因缺乏训练样本而受到限制，因此大量的研究工作都是基于英文描述方法，而对于视频中文描述的研究相对较少。因此，设计了一种基于自适应特征选择和融合的视频中文描述方法。首先，设计一个特征选择网络，先使用注意力机制关注重要的特征，再使用门控机制选择性地保留或丢弃特征，从而提升模型语义的准确性。其次，设计一种自适应动态融合机制，通过计算运动特征的权重系数将视觉和运动特征向量进行动态融合，减少冗余信息的干扰，从而提升模型语义的准确性。最后本文还在中文数据集 MSVD-C 上进行上测试，其中 BLEU@4 和 CIDEr-D 指标提升尤为显著。

§4.1 相关研究工作

许多研究者证明将高维度的视频数据转换为低维度的特征表示可以有效地减少冗余和不必要的信息，同时提取出视频中的关键特征，使得后续的任务可以更加高效地进行处理。2017 年，Krause 等人[65]提出了一种逐层生成图像段落描述的方法，该方法使用线性变换将图像特征压缩到低维空间，并在每个层次上使用 LSTM 来生成描述性语句。同年，Tu 等人[66]提出了一种基于注意力池化的视频描述方法，通过利用线性变换将视频特征紧凑化，并使用注意力机制来自适应地融合时空特征。近年，Jin 等人[67]又提出一种期望最大化对比学习(EMCL)方法来学习紧凑的视频和语言表示，期望最大化算法来为潜在空间寻找一组紧凑的基，其中的特征可以简明地表示为基的线性组合。上述方法虽然通过线性变换将高维的视频特征压缩成低维的紧凑特征表示，能减少特征的维度和冗余性，但它们是静态的降维方法，无法对数据的非线性相关性进行建模，会丢失重要信息。因此，如何使用动态的降维方法避免重要信息丢

失已经成为视频描述任务的难点内容之一。在视频描述领域，也有不少研究者致力于将 2D 和 3D 特征相结合，以提高视频描述的性能。Yao 等人[68]通过使用 3D 卷积神经网络提取视频的空间信息，空间注意力机制来关注视频中重要的片段，再将获得的时空特征进行融合来获取更优的描述。Yuan 等人[69]通过设计一个分层句子语法编码器来提取例句的句法结构，同样也采用了 2D 视觉特征与 3D 运动特征融合来获得的特征表示。Zhong 等人[70]提出了一种新的频率扩散（RSFD）精细语义增强方法，同样也使用 2D 视觉特征与 3D 运动特征融合的方法来实现一种不断感知不常用词的视频描述模型。

§4.2 基于自适应特征选择与融合的视频中文描述模型的设计

§4.2.1 视频描述模型的整体框架

本章提出的基于自适应特征选择与融合的视频中文描述方法，简称 AFSF（Adaptive Feature Selection and Fusion）。AFSF 整体框架如图 4-1 所示，该模型由编码器和解码器两部分组成。其中编码器分别由特征提取网络、特征选择网络（Feature Selection，FS）和自适应动态融合机制(Adaptive Feature Fusion，AFF)组成。首先 AFSF 模型先对输入的视频进行预处理获得标准的输入帧，然后利用特征提取网络对预处理后的视频帧进行 2D 视觉特征和 3D 运动特征提取，接着特征选择网络分别对已提取的 2D 视觉特征和 3D 运动特征进行强化重要特征的降维处理，再接着自适应特征融合网络将特征选择网络的降维特征进行自适应特征融合，最后将融合后的特征输入解码器以获得准确的句子描述。

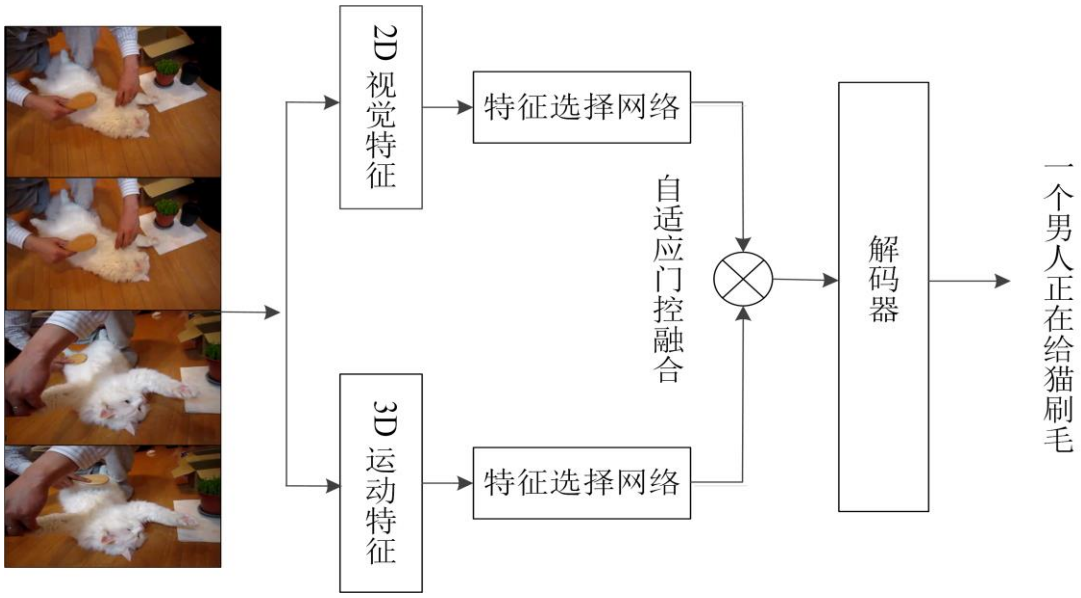

图 4-1 AFSF 模型的整体框架

§4.2.2 特征选择网络

在视频描述任务中，由于视频包含非常丰富的信息内容，往往需要从不同的角度进行特征提取。因此，提取出来的特征通常是高维特征。然而，高维特征向量往往包含大量冗余和不必要的信息，若仅进行简单的线性变换容易丢失重要信息。因此针对该问题，受王等人[71]使用注意力有效地区别每个特征的重要程度的启发，本文提出了一个特征选择网络，该网络通过将注意力机制和门控组合方法，将高维特征降低到低维特征并强化重要的特征信息，以获得更好的特征表达。

为了使网络能正确区分视频特征向量的重要程度，特征选择网络分别计算高维度的 2D 视觉特征向量V_a和 3D 运动特征向量V_m的相似矩阵，再利用相似度矩阵计算出每个特征向量对于整个视频注意力权重系数，并根据注意力权重系数，初步筛选出具备表达性和判别性的特征$\bar{V}$，计算过程如公式所示：

$$A_{ij} = \alpha(V_x) \tag{4-1}$$

$$\bar{V} = \sum_{j=1}^{n} \frac{\exp(A_{ij})}{\sum_{k=1}^{n} \exp(A_{kj})} V_x \tag{4-2}$$

其中，V_x代表特征提取网络输出的 2D 视觉特征或 3D 运动特征，A_{ij}代表由$\alpha(\cdot)$计算获得的相似矩阵。

最后将输入特征V_x和特征$\bar{V}$进行拼接，再使用门控机制控制信息的流动，使网络可以更准确地保留 2D 视觉特征和 3D 运动特征的重要信息，达到更好的降维效果，计算过程如公式所示：

$$z_i = \tanh(W_{e1}[V_x : \bar{V}] + b_1) \tag{4-3}$$

$$r_i = \delta(W_{e2}[V_x : \bar{V}] + b_2) \tag{4-4}$$

$$f_i = \delta(W_{e3}[V_x : \bar{V}] + b_2) \tag{4-5}$$

$$V' = r_i \odot V_x + f_i \odot z_i \tag{4-6}$$

其中，W_{e1}、W_{e2}、$W_{e3} \in R^{2d_v \times d_m}$和$b_1$、$b_2$、$b_2$代表可训练参数，[:]代表行拼接，$\delta$是非线性激活函数 sigmoid，$\odot$代表元素向量点乘。

§4.2.3 基于视觉特征和运动特征的自适应特征融合机制

在视频描述任务中，2D 视觉特征和 3D 运动特征分别提取视频中的图像和运动

序列，这两种特征分别关注视频的空间信息和时间信息，但是它们之间存在着信息的差异性和互补性。特征选择网络虽然能避免了重要信息的丢失，但模型仅对其进行简单地拼接会产生冗余信息，从而降低视频描述的性能。因此，针对上述问题，本文受 Balazs 等人[72]提出的子词和词组表征的融合门启发，设计了一种基于视觉特征和运动特征的动态融合机制，该机制不依赖于外部信息来源，仅将来自特征选择网络输出的 2D 视觉特征V_a'和 3D 运动特征V_m'作为输入，并以 3D 运动特征为基准计算动态选择系数g_i，通过门控动态机制计算 2D 视觉特征V_a'和 3D 运动特征V_m'融合后的特征表示。具体的计算过程如下：

$$g_i = \delta(W_{e4}V_m' + b_4) \tag{4-7}$$

$$V = g_i \odot V_a' + (1 - g_i) \odot V_m' \tag{4-8}$$

其中，$W_{e1} \in R^{d_m \times d_m}$和$b_1$代表可训练参数，$g_i \in (0, 1)$，$\delta$代表非线性激活函数 Sigmoid，$\odot$为元素向量点乘。

§4.3 实验结果与分析

§4.3.1 实验数据与参数设置

数据集和参数设置方面，本章选择了中文描述数据集 MSVD-C 进行实验，训练时将中文数据集 MSVD-C 的最大输入序列长度设置为 20 词长。解码器采用了 1 个解码器层、512 个模型维度、2048 个隐藏维度和每层 8 个注意力头。Adam 优化器以 0.005 的初始学习速率分批训练 64 个视频，epoch 设置为 50，dropout 设置为 0.5，L2 权重衰减是 0.0005，测试时模型时 batch size 设置为 32。

评价指标方面，为了验证本文方法的有效性，本章使用视频描述方法常用的四个评价指标分别为 BLEU@4、METEOR、ROUGE-L 和 CIDEr-D。

§4.3.2 模型性能比较

为了验证 SCA-SS 方法的有效性，本文选择了较为先进的模型在中文数据集 MSVD-C 上进行实验对比，这些模型可分为两类：（1）NIC-KD[61]、S2VT-KD[62]、ConvCapp-KD[63]和 Top-Down-KD[64]模型，这四种模型分别由四种经典视频/图像描述方法 NIC、S2VT、ConvCapp 和 Top-Down 引入了跨语言知识蒸馏的视频中文描述方法获得。跨语言知识蒸馏的视频中文描述方法是目前最先进的视频中文描述方法，也是唯一一种研究视频中文描述的方法，与它们进行对比实验能验证本文方法的先进

性。（2）NACF、SSLL 以及 SCA-SS 模型，其中基于句子语义与长度损失计算的视频描述方法 SSLL、基于余弦注意力与语义选择的视频中文描述方法 SCA-SS 模型、本章所提的 AFSF 模型和 NACF 模型都是使用非自回归的视频描述方法，与它们进行对比实验可以更直接的反映本章所提的基于自适应特征选择与融合的视频中文描述方法 AFSF 的有效性。所述 SSLL 和 SCA-SS 模型分别为第二章和第三章所设计模型。实验结果如表 4-1 所示，其中最佳结果用黑色字突出显示。可以看出本文方法 AFSF 相较与目前的先进非自回归生成视频描述生成模型 NACF 拥有大幅度提升，其中视频描述常用的指标 BLUE@4、METEOR、ROUGE-L 和 CIDEr-D 分别提升了 20.6%、3.6%、6.2%、20.5%，与 SCA-SS 模型进行比较 BLUE@4、METEOR、ROUGE-L 和 CIDEr-D 指标分别提升了 1.5%、0.2%、2.0%、0.6%，说明提出的方法 AFSF 能够避免重要信息的丢失和滤除冗余信息，进而提升模型描述语义的准确性。

表 4-1 AFSF 模型和其他模型的对比结果

模型	BLUE@4	METEOR	ROUGE-L	CIDEr-D
NIC-KD	21.3	21.6	47.6	42.1
S2VT-KD	21.1	21.6	48.0	46.8
ConvCapp-KD	19.1	21.5	47.0	42.2
Top-Down-KD	21.1	22.9	49.4	57.4
NACF	37.2	38.0	64.2	73.8
SSLL	39.5	38.3	65.2	75.9
SCA-SS	44.2	39.3	66.8	88.4
AFSF	**44.9**	**39.4**	**68.2**	**89.0**

§4.3.3 消融实验

表 4-2 FS 与 AFF 方法有效性对比

模型	BLUE@4	METEOR	ROUGE-L	CIDErD
NACF	37.2	38.0	64.2	73.8
-FS	43.1	38.9	66.9	81.0
-AFF	41.1	38.8	66.2	85.5
AFSF	44.9	39.4	68.2	89.0

如表 4-2 所示，本节通过在 MSVD-C 中文数据集上进行消融实验来验证所提组件特征选择网络（FS）和自适应特征融合机制（AFF）的有效性。从表 4-2 可知 FS 和 AFF 在 MSVD-C 中文数据集测试的各个指标都优于 NACF 模型，说明特征选择网络（FS）能在特征降维时避免重要信息丢失，自适应特征融合机制（AFF）能自适应

融合视觉特征和运动特征避免产生冗余信息。为了进一步考察上述指标的提升在实际应用中的效果，FS、AFF 与 NACF 模型的描述生成结果进行对比分析，后者的 100 条测试样本生成的结果大约有 70 条样本存在语义信息不准确问题，引入本文所提升的 FS 和 AFF 方法后分别有 18 和 15 条样本获得明显改进。提升幅度大约达 25%和 21%，实际应用的效果得到显著提升。部分样例如图 4-2 所示。

GT:一个男人坐在餐厅外面弹吉他

NACF:一个人在弹吉他

AFSF:一个男人坐在屋外弹吉他

FS:一个男人坐着弹吉他

AFF:一个男人在屋外弹吉他

图 4-2（a）：实例分析

GT:一只赖皮狗正在游泳池边玩水

NACF:一个狗正在池边玩

AFSF:一个狗正在池边玩水

FS:一个狗正在池边玩水

AFF:一个狗正在池边玩水

图 4-2（b）：实例分析

图 4-2 展示了中文数据集 MSVD-C 的部分视频和分别使用 NACF、FS、AFF 和 AFSF 模型测试生成的描述结果，其中 GT 代表参考描述，NACF 模型作为参考模型进行比照。从图 4-2 的两个实例分析可以看出 NACF 模型生成的描述与视频内容的语义存在差异。例如实例分析（a）中参考描述的目标对象有“男人”、“餐厅”和“吉

他”，而参考模型 NACF 生成的描述的目标对象仅有“人”和“吉他”，缺少了“餐厅”和具体的性别，说明 NACF 模型的编码器在特征降维时会丢失重要信息，以及在视觉特征和运动特征融合时产生冗余信息，导致模型生成描述语义的不准确。图中 FS 和 AFF 模型生成的描述结果表明，引入 FS 和 AFF 模型后，描述视频的语义信息得到纠正，这从侧面证明了引入 FS 方法后能在特征降维时避免重要信息丢失，AFF 方法能自适应融合视觉特征和运动特征避免产生冗余信息。AFSF 模型是由 FS 和 AFF 融合而成的模型，通过将图 4-2 中 AFSF 模型的生成描述的语义信息与参考描述的语义信息对比可知，它们的语义信息是基本一致的，证明了该模型不仅在特征降维时避免丢失重要信息，还能自适应融合视觉特征和运动特征避免产生冗余信息。

§4.4 本章小结

本文提出了一种基于自适应特征选择与融合的视频中文描述方法，通过设计一种特征选择网络，使得模型得能在特征降维时避免重要信息丢失，提升了生成描述语义信息的准确性。其次，通过设计一种自适应动态融合机制，能使视觉特征和运动特征自适应进行融合，避免产生冗余信息，提升了模型的准确性。通过在中文数据集 MSVD-C 上进行实验，各项性能指标都表现出显著的提升，远优于当前先进的模型说明本文方法对于提升描述内容语义信息准确性是十分有效的。在以后的研究中，团队还会继续对视频中文描述任务进行研究，旨在进一步解决视觉语义与描述内容语义不一致的问题。

第五章 总结与展望

§5.1 本文总结

近年来，随着直播行业和短视频在全世界范围内兴起，人们观看和制作短视频的需求急剧上升，视频内容描述技术也应用越来越广泛。视频内容描述的核心是进行视频语义解析，而视频理解领域的大量工作都是针对英文描述方法，较少研究视频中文描述方法。这是因为英文的视频描述训练样本更丰富，而中文训练样本相对较少。因此，本文先基于英文描述进行了视频语义内容解析的研究，解决了生成描述语义信息不完整和不准确的问题。然后在英文视频描述方法的基础上，提出了一种基于余弦注意力和语义选择的视频中文描述方法，解决了中文描述的语义冗余和语义不准确的问题。接着，在视频中文描述研究的基础上，又提出了一种基于自适应特征选择和融合的视频中文描述方法，进一步提升了生成的中文描述语义的质量。最后为了解决中文样本缺少的问题，将视频英文描述数据集 MSVD 扩展成中文数据集 MSVD-C。本文研究工作总结如下：

1.提出了一种基于句子语义与长度损失计算的视频描述的方法，通过设计一个新的长度损失函数，提升了生成描述语义内容的完整性。其次，通过设计一个基于句子语义的描述生成损失函数，提升了生成描述语义信息的准确性。在 MSVD 和 MSR-VTT 数据集上测试，各项性能指标显著提升，均优于目前先进的模型。其中 BLEU@4 和 METEOR 指标提升尤为显著，说明该方法能提升描述内容语义信息的完整性和准确性。

2.提出了一种基于余弦注意力与语义选择的视频中文描述方法，通过重新设计一个缩放余弦注意力网络，使得模型能准确地关注到视觉语义特征，提升了生成描述语义信息的准确性。其次，通过设计一个语义选择网络能滤除冗余信息对模型的干扰，提升了模型的准确性。为了验证该方法的有效性，制作了一个中文数据集 MSVD-C，并在该数据集上进行实验。实验结果表明该方法在各项性能指标上均显著优于目前先进的模型，说明该方法能够有效提高描述内容语义信息的准确性。

3.提出了一种基于自适应特征选择与融合的视频中文描述方法。通过设计了一种特征选择网络，避免了冗余信息产生，提升了生成描述语义信息的准确性。其次，通过设计一种自适应动态融合机制，将视觉特征和运动特征进行了自适应动态融合，提升了模型的准确性。在中文数据集 MSVD-C 上进行实验，各项性能指标显著提升，均优于目前先进的模型，说明该方法能够有效提高描述内容语义信息的准确性。

综上所述，本文针对视频语义解析问题的研究已经达到了预期课题所设定的研究目

标，实验验证了算法的有效性，根据实际生成的内容对比观察，本文方法生成的描述比现有最优模型的语义描述更准确和完整，同时也为进一步对视频内容进行准确和优美描述打下基础。

§5.2 后续与展望

虽然本文设计的基于句子语义与长度损失计算的视频描述模型、基于余弦注意力与语义选择的视频中文描述模型以及基于自适应特征选择与融合的视频中文描述模型都取得不错的效果，但是仍然存在不足需要进一步研究，具体如下：

1.研究的基于深度学习的视频语义解析都是针对视觉和文本特征进行联合建模，但是视频还包含音频和其他重要信息，因此未来的研究可以探索多模态信息的利用，如联合建模动画和真实视频等以拓展应用场景。此外，还可以研究如何有效地融合多模态信息，平衡不同模态信息的贡献，提高视频语义解析的准确性和鲁棒性。多模态视频语义解析是未来的重要研究方向，有望实现更全面、准确的视频理解。

2.研究的基于深度学习的视频语义解析都集中在短视频和网络视频的场景下，但是随着网络带宽和存储技术的不断提升，长视频和复杂场景下的视频数据变得越来越普遍，因此在这些场景下进行视频语义解析的需求也越来越迫切。未来的研究可以考虑在设计模型时引入更加复杂的结构，如分层注意力机制、多层次特征融合等方法，以提高视频语义解析的精度和效率。此外，可以考虑结合强化学习等技术，进一步提升视频语义解析的效果和稳定性，使其具有更广泛的应用前景。

3.所提出方法都是基于视频的语义进行解析，未深入研究视频特征提取和表示方法。然而，视频特征提取和表示方法是视频语义解析的关键环节，当前方法仍有局限性。未来的研究可以探索更有效的特征提取和表示方法，例如基于图像和视频联合学习的方法。

4.所提出的算法使用的数据集都是基于大规模的短视频数据集，其存在标注质量和数据集样本的丰富性都存在局限性。未来的研究可以探索更准确和丰富的标注方法，以提高视频语义解析的性能。同时，为了更好地模拟真实世界的视频场景，研究人员可以探索使用更广泛和多样化的数据集，包括真实场景下的视频数据集，将有助于提高算法的泛化能力和适应性，从而更好地应用。

参考文献

[1] 牛满堂. 基于语义信息的视频描述算法研究[D]. 西安: 西安电子科技大学, 2022.

[2] 李栋. 基于时空关联性的视频动作识别与检测方法研究[D]. 合肥: 中国科学技术大学, 2021.

[3] 王永. 基于 Transformer 网络和双向解码的视频描述研究方法[D]. 南昌:江西师范大学, 2022.

[4] 闫雨寒. 基于多注意力和语义检测的视频描述方法研究[D]. 徐州: 中国矿业大学,2021.

[5] 常志. 基于深度学习的多特征多模态视频描述方法研究[D].天津: 天津理工大学, 2022.

[6] 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: 171-184.

[7] 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. Piscataway, NJ: IEEE, 2013: 433-440.

[8] Thomason J, Venugopalan S, Guadarrama S, et al. Integrating language and vision to generate natural language descriptions of videos in the wild[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Ireland: ACL, 2014: 1218-1227.

[9] Lin K, Li L, Lin C C, et al. Swinbert: End-to-end transformers with sparse attention for video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 17949-17958.

[10] 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. Piscataway, NJ: IEEE, 2022: 17939-17948.

[11] Seo P H, Nagrani A, Arnab A, et al. End-to-end generative pretraining for multimodal video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 17959-17968.

[12] Zeng P, Zhang H, Gao L, et al. Visual Commonsense-aware Representation Network for Video Captioning[J]. arXiv preprint arXiv:2211.09469, 2022.

[13] Venugopalan S, Xu H, Donahue J, et al. Translating videos to natural language using deep recurrent neural networks[J]. arXiv preprint arXiv:1412.4729, 2014.

[14] Wu X, Li G, Cao Q, et al. Interpretable video captioning via trajectory structured localization[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6829-6837.

[15] Rothenpieler D L, Amiriparian S. METEOR Guided Divergence for Video Captioning[J]. arXiv preprint arXiv:2212.10690, 2022.

[16] Xu J, Yao T, Zhang Y, et al. Learning multimodal attention LSTM networks for video captioning[C]//Proceedings of the 25th ACM international conference on Multimedia. New York: ACM,2017: 537-545.

[17] Xu H, Ye Q, Yan M, et al. mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video[J]. arXiv preprint arXiv:2302.00402, 2023.

[18] Ghaderi Z, Salewski L, Lensch H P A. Diverse Video Captioning by Adaptive Spatio-temporal Attention[C]//Pattern Recognition: 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27–30, 2022, Proceedings. Cham: Springer International Publishing, Konstanz: DAGM, 2022: 409-425.

[19] Monfort M, Jin S Y, Liu A, et al. Spoken moments: Learning joint audio-visual representations from video descriptions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14871-14881.

[20] Tanaka T, Simo-Serra E. Lol-v2t: Large-scale esports video description dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4557-4566.

[21] Zhang J, Peng Y. Object-aware aggregation with bidirectional temporal graph for video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8327-8336.

[22] Wang B, Ma L, Zhang W, et al. Reconstruction network for video captioning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2018: 7622-7631.

[23] Zala A, Cho J, Kottur S, et al. Hierarchical Video-Moment Retrieval and Step-Captioning[J]. arXiv preprint arXiv:2303.16406, 2023.

[24] Yang B, Liu F, Zou Y, et al. ZeroNLG: Aligning and Autoencoding Domains for Zero-Shot Multimodal and Multilingual Natural Language Generation[J]. arXiv preprint arXiv:2303.06458, 2023.

[25] Huang Z, Chen Z, Li Q, et al. 1st Place Solutions of Waymo Open Dataset Challenge 2020--2D Object Detection Track[J]. arXiv preprint arXiv:2008.01365, 2020.

[26] Zhang Y, Song X, Bai B, et al. 2nd Place Solution for Waymo Open Dataset Challenge--Real-time 2D Object Detection[J]. arXiv preprint arXiv:2106.08713, 2021.

[27] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2015: 3431-3440.

[28] Chen Y. Convolutional neural network for sentence classification[D]. University of Waterloo, 2015.

[29] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[30] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[31] 席超. 基于计算智能的图像变化检测方法研究[D]. 无锡市: 江南大学, 2022.

[32] 闫严. 基于深度学习的多模态数据处理算法研究[D].天津: 天津大学, 2019.

[33] 闻婷. 基于深度学习的长视频描述技术研究与实现[D].东南大学,2021.

[34] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[35] Pan Y, Mei T, Yao T, et al. Jointly modeling embedding and translation to bridge video and language[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: CVPR, 2016: 4594-4602.

[36] Pei W, Zhang J, Wang X, et al. Memory-attended recurrent network for video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: CVPR, 2019: 8347-8356.

[37] Ryu H, Kang S, Kang H, et al. Semantic grouping network for video captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2021, 35(3): 2514-2522.

[38] Wang Y, Huang G, Yuming L,et al. MIVCN: Multimodal interaction video capti Jointly modeling embedding and translation to bridge video and language oning network based on semantic association graph[J]. Applied Intelligence, 2022, 52(5): 5241-5260.

[39] Wang B, Ma L, Zhang W, et al. Controllable video captioning with pos sequence guidance based on gated fusion network[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul: ICCV, 2019: 2641-2650.

[40] Hou J, Wu X, Zhao W, et al. Joint syntax representation learning and visual cue translation for video captioning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: ICCV, 2019: 8918-8927.

[41] Zheng Q, Wang C, Tao D. Syntax-aware action targeting for video captioning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: CVPR, 2020: 13096-13105.

[42] Aafaq N, Akhtar N, Liu W, et al. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12487-12496.

[43] Tan G, Liu D, Wang M, et al. Learning to discretely compose reasoning module networks for video captioning[J]. arXiv preprint arXiv:2007.09049, 2020.

[44] Vaidya J, Subramaniam A, Mittal A. MITTAL A. Co-Segmentation Aided Two-Stream Architecture for Video Captioning[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Long Beach: WACV, 2022: 2774-2784.

[45] Yang B, Zou Y, Liu F, et al. Non-autoregressive coarse-to-fine video captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2021, 35(4): 3119-3127.

[46] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: CVPR,2016: 770-778.

[47] Kang D, Hashimoto T. Improved natural language generation via loss truncation[J]. arXiv preprint arXiv:2004.14589, 2020.

[48] Li X, Meng Y, Yuan A, et al. Lava nat: A non-autoregressive translation model with look-around decoding and vocabulary attention[J]. arXiv preprint arXiv:2002.03084, 2020.

[49] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.

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

[51] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language[C]//Proceedings of the ninth workshop on statistical machine translation. Pennsylvania,: ACL, 2014: 376-380.

[52] 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. Piscataway, NJ: IEEE, 2015: 4566-4575.

[53] Iashin V, Rahtu E. A better use of audio-visual cues: Dense video captioning with bi-modal transformer[J]. arXiv preprint arXiv:2005.08271, 2020.

[54] Li L H, Yatskar M, Yin D, et al. Visualbert: A simple and performant baseline for vision and language[J]. arXiv preprint arXiv:1908.03557, 2019.

[55] 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: WACV, 2021: 1543-1552

[56] Ramanishka V, Das A, Park D H, et al. Multimodal video description[C]//Proceedings of the 24th ACM international conference on Multimedia. Xiamen: ICIAI, 2016: 1092-1096.

[57] 侯静怡, 齐雅昀, 吴心筱, 等. 跨语言知识蒸馏的视频中文字幕生成[J]. 计算机学报, 2021, 44(9):

1907-1921.

[58] 常志, 赵德新. 基于深度学习的视频描述方法研究综述[J]. 天津理工大学学报, 2020.36(6):17-23

[59] Henry A, Dachapally P R, Pawar S, et al. Query-key normalization for transformers[J]. Association for Computational Linguistics, 2020:4246–4253

[60] Parisotto E, Song F, Rae J, et al. Stabilizing transformers for reinforcement learning[C]//International conference on machine learning. PMLR, online: ICML,

[61] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]//Proceedings of the IEEE conference on computer vision and pattern recognition.USA:IEEE，2015:3156-3164

[62] Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to sequence-video to text[C]//Proceedings of the IEEE international conference on computer vision. Santiago:iEEE, 2015: 4534-4542

[63] Aneja J, Deshpande A, Schwing A G,et al.Convolutional image captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake City, USA:IEEE，2018:5561-5570

[64] 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:IEEE，2018:6077-6086.

[65] Krause J, Johnson J, Krishna R, et al. A hierarchical approach for generating descriptive image paragraphs[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2017: 317-325.

[66] Tu Y, Zhang X, Liu B, et al. Video description with spatial-temporal attention[C]//Proceedings of the 25th ACM international conference on Multimedia. New York: ACM, 2017: 1014-1022.

[67] Jin P, Huang J, Liu F, et al. Expectation-Maximization Contrastive Learning for Compact Video-and-Language Representations[J]. Advances in Neural Information Processing Systems, 2022, 35: 30291-30306.

[68] Yao L, Torabi A, Cho K, et al. Describing videos by exploiting temporal structure[C]//Proceedings of the IEEE international conference on computer vision. Piscataway, NJ: IEEE, 2015: 4507-4515.

[69] Yuan Y, Ma L, Zhu W. Syntax Customized Video Captioning by Imitating Exemplar Sentences[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway, NJ: IEEE, 2021, 44(12): 10209-10221.

[70] Zhong X, Li Z, Chen S, et al. Refined Semantic Enhancement towards Frequency Diffusion for Video Captioning[J]. arXiv preprint arXiv:2211.15076, 2022.

[71] 王鑫岚. 社交网络中节点影响力与身份对齐方法研究[D].桂林: 桂林电子科技大学, 2022.

[72] Balazs J A, Matsuo Y. Gating mechanisms for combining character and word-level word representations: an empirical study[J]. arXiv preprint arXiv:1904.05584, 2019.