摘　要

视频描述是一个与计算机视觉、自然语言处理和机器学习等领域相关的多模态信息处理任务，它可以将输入的视频自动生成与视频内容对应的文本描述。

观看视频时，人们能够获取视频中的整体视觉信息、运动信息、目标信息以及目标之间的关系信息。由于视频中目标数量众多，而视频的事件通常发生在目标密集的区域，为了方便理解视频内容，人们会重点关注密集的目标特征，因此，这些目标的特征会被增强。此外，人脑会将获取的视频信息分别融合在一起，以减少不同特征之间的干扰。基于这种人脑对视频的处理方式，本文提出了两种视频描述方法：一种是增强视频目标特征的方法，另一种是分步融合视频特征的方法。

（1）增强视频目标特征的视频描述方法。这种方法采用了双向 Transformer 模型，由于目标特征具备图结构数据的特点，使用图卷积神经网络来加强目标特征。同时，它还利用目标置信度来决策多个被检测目标的结果，以减少由于目标特征不准确或虚假而导致的负面影响。并且该方法还使用了多头注意力机制，通过学习目标特征的高层语义信息，从而帮助模型生成更准确的描述。

（2）分步融合视频特征的视频描述方法。这种方法基于 Transformer 模型，并且考虑了视频特征的不同特性，使用多个神经网络编码视频特征，从而减少特征融合时造成的信息丢失。该方法在多头注意力机制的基础上，提出了分步融合特征的方法。除此以外，它还利用视频目标的关系特征作为先验知识，指导模型生成视频描述。

本文提出的方法充分考虑了目标特征和目标关系特征，从而得到了更准确、更完整的视频表示。实验结果表明，在 MSVD 和 MSR-VTT 这两个常用的数据集上，这两种方法能够生成更为精细、内容更丰富的文本描述，使得生成的描述文本更符合人类的语法规则。

关键词：视频描述；特征增强；特征融合；先验知识

Abstract

Video captioning is a multimodal information processing task related to computer vision, natural language processing, and machine learning. It can automatically generate textual descriptions corresponding to the content of input videos.

When people watch a video, the content they see includes the overall visual information of the video, the motion information of the video, the object information of the video and the relationship information between the video objects. Usually, the number of objects in the video is very large, and the video content occurs in the object dense area, and the object features in this area will be focused by the human brain. And according to the object features around each object, the object feature is enhanced. At the same time, after obtaining the video information, the human brain will fuse the semantic information of the video respectively through the connection and difference between them to help the human brain understand the video content. Inspired by the above idea of "object feature enhancement + step-by-step feature fusion" in human brain processing video, this thesis constructs a video caption method for enhancing video object features and a video caption method for step-by-step fusion of video features, respectively.

(1)A video caption method to enhance the features of video objects. In this method, two-way Transformer model is adopted. Since object features have the characteristics of graph structure data, graph convolutional neural network is used to strengthen object features, then uses the video object confidence to decide multiple detection objects, and learns the high-level semantic information of the object features based on the multi-head attention mechanism to reduce the negative impact of the inaccurate and false object features.

(2)Video caption method by fusing video features step by step. The method is based on the Transformer model. According to the characteristics of video features, this thesis uses different neural networks to encode video features, and based on the multi-head attention mechanism, this thesis proposes a step-by-step feature fusion method, which uses the relationship features of video objects as prior knowledge to guide the model to generate video captions, and reduces the information loss of video features in feature fusion.

The proposed method in this thesis fully takes into account the object features and object relationship features, resulting in a more accurate and complete representation of videos. Experimental results show that on the commonly used datasets, MSVD and MSR-VTT, these two methods can generate more fine-grained and informative textual descriptions, making the generated descriptions more compliant with human grammatical rules.

Key Word: Video Captioning; Feature Enhancement; Feature Fusion; Prior Knowledge

目 录

1 绪论

1.1 研究背景与意义

随着科技的不断发展，人们越来越喜欢使用短视频分享生活。作为连接人际关系的平台，短视频受到越来越多人的青睐。根据中国互联网络信息中心发布的第 50 次《中国互联网络发展状况统计报告》显示，截至 2022 年 6 月，我国短视频的用户规模已经达到 9.62 亿。然而，随着短视频的发展，一些弊端也逐渐浮出水面，例如血腥暴力、内容低俗、侵权频发等问题。由于短视频用户数量庞大，产生的视频数据也十分惊人，采用人工审核和监督视频的效率过低，并且不能保证能够识别出不符合规定的视频。此外，在监控视频领域中，监控人员需要花费大量的人力物力对视频进行标记，但这种标记视频的方法效率过低，并且不便于后期检索。因此，人们急需一种能够让机器理解视频内容的方法，以帮助人们更快速、高效地监督和管理视频。

视频描述任务是一个复杂的多模态学习任务，旨在根据视频内容自动生成对应的文本描述。由于它涉及计算机视觉和自然语言处理两个领域的知识，因此需要采用复杂的模型结构。视频描述任务包括两个子任务：第一个子任务是通过预训练的神经网络模型从视频中提取多种特征；第二个子任务是使用类似于机器翻译的"编码器+解码器"结构，将提取的特征序列转换成关于视频内容的文字描述。这个过程类似于将一种模态的信息转换成另一种模态的信息，需要综合使用计算机视觉和自然语言处理的知识。

除了上述提到的领域，视频描述还有广泛的应用前景。例如，在人机交互方面，通过将演示视频中的动作转化为简单指令，视频描述任务可以生成书面程序，以供人类或机器人使用。在视觉障碍人士的辅助设备方面，视频描述可以帮助他们获取外界的视觉信息，提供生活便利。此外，视频描述生成的描述还可以作为视频的标签，从而使视频检索更加快速、全面和高效。

1.2 国内外研究现状

视频描述任务的研究可以分为两个阶段。第一阶段采用预定义的模板来描述视频，研究人员从视频中提取关键物体信息，并将这些信息填充到符合语义规则的固定模板中。随着深度学习的快速发展和视频描述数据集的扩充，第二阶段采用了基于序列的视频描述方法，该方法采用编码器-解码器(encoder-decoder)结构，

该结构已在机器翻译任务中广泛应用并表现出良好的性能。由于视频描述任务与机器翻译任务有一定的相似性，因此该结构被移植到视频描述任务中，并且当前大部分视频描述的研究都是基于该结构[1]。

1.2.1 基于模板的视频描述方法

Kojima 等人[2]的研究工作是视频描述研究领域的先驱之一。他们采用人的姿势测量技术，获取视频中头部方向、头部位置和手部状态等信息，并结合动词的语义特征和从视频中获取的语义特征，学习语义规则并选取最合适的名词和宾语。最后，将选取的语言成分填入预定义的模板句子中。这类早期的方法都是基于预定义模板的，需要手动设计模板，因此存在一定的局限性。

Rohrbach 等人[3]的研究将视频描述任务视为机器翻译任务，通过获得对视觉内容的丰富描述来预测语义表示。他们使用条件随机场模拟各种输入内容之间的关系，并将描述文本的生成视为机器翻译问题，把视觉语义特征翻译成文本。这种方法的主要缺点是依赖预定义的句子模板，生成的句子具有预定的语言格式，因此准确率较低。此外，这种方法只能对视频进行简单的描述，并且生成的句子句型固定，不够美观。

1.2.2 基于序列的视频描述方法

基于序列的视频描述方法采用了一种较先进的机器学习方法，它可以通过学习自然语言和视频特征在同一空间中的概率分布，更加灵活地生成具有不同语法结构的句子。该方法将视频特征向量被提取出来后，经过循环神经网络对视频特征向量的转换，得到对应的视频描述语句，而这个过程中，卷积神经网络则起到了提取视频特征的关键作用。相比于传统的基于预定义模板的方法，这种方法可以更好地处理视频中的时间信息，提高描述的准确性和连贯性，同时能够生成更加灵活多样的语言描述，使得描述更加自然流畅。

Donahue 等人[4]提出的视频描述方法采用了长期循环卷积网络和条件随机场来预测视频中的物体、运动和关系的词语，并利用 LSTM（Long and Short Term Memory）解码器生成描述文本。然而，该方案缺乏对预测词语在空间和时间上的信息考虑，且使用条件随机场得到的描述容易遗漏信息。为了解决这些问题，Venugopalan 等人[5]提出了一种基于 CNN（Convolutional Neural Network）-RNN（Recurrent Neural Network）的视频描述模型 MP-LSTM，该方法在提取视频特征时，通过预训练的卷积神经网络逐帧处理，随后对特征进行平均池化，使其成

为固定维度的向量，接下来，将向量输入到 LSTM 解码器中，以逐字生成视频描述。然而，该方法仍然未考虑到视频的时间顺序关系。为了解决这个问题，Venugopalan 等人[6]提出了一种端到端的视频到文本(S2VT)模型，结合了图像特征和光流图像特征，采用 LSTM 编码和解码，并融合两种特征得到结果。这个方法虽然考虑了时间相关性，但是存在长期依赖信息损失的问题。Yao 等人[7]提出使用 3D 卷积神经网络提取视频特征作为局部动作特征，使用注意力机制表示时间特征，他们的方法能够动态地关注每个词的相关视频特征。这些方法都取得了一定的进展，但仍然存在一些缺陷和挑战，需要进一步研究和改进。

在视频描述任务中，使用 3D 卷积神经网络[8]提取的特征可以有效捕捉视频的时序特征，从而提高模型的性能。在随后的研究中，这种方法得到了广泛应用，通常是通过同时使用 2D 和 3D 卷积神经网络来表示视频特征。这种方法的优点在于可以同时考虑视频的空间和时序特征，从而提高描述文本的准确性和连贯性。

为了提高视频描述任务的性能，Chen 等人[9]考虑到提取视频特征时存在大量冗余视频帧的问题，因此提出使用强化学习方法来选择关键帧，并使用这些关键帧来表示整个视频输入，从而减少计算量，同时保持模型的性能。Wang 等人[10]提出了一种对偶学习方法，即使用重构网络（RecNet）来重新生成视频特征向量，并与原始视频特征向量进行相似度比较，从而提高生成描述的质量。Wang 等人[11]还提出了基于视频多重表示的门控融合网络，利用预测描述词的词性信息缩小预测词的选择范围，进一步引导生成更准确的词。丁恩杰等人[12]提出了一种用于视频描述任务的多模态视觉特征提取与融合方法，采用多层长短期记忆网络来融合多维、多模态信息，以生成描述文本。Ryu 等人[13]提出了语义组网络，它通过使用部分解码的短语来对视频的不同帧进行分组，并在预测下一个单词时解码这些语义对齐的组，以提高模型的描述性能。

微软团队提出了 2D-TAN 网络[14]来处理视频的时序信息，该网络将视频转换为二维时间图，并在此基础上进行时间定位。而 Zhou 等人[15]则在密集视频描述任务中引入 Transformer 模型，将其称为 vanilla Transformer。

1.3 本文的主要研究内容和章节安排

1.3.1 本文的主要研究内容

视频描述任务是一种将视频转化为文本的生成任务，类似于自然语言生成文本的过程。因此，使用将卷积神经网络和循环神经网络结合的 CNN-RNN 网络，从视频中提取特征并生成描述文本是一种常用的方法。目前，许多研究都是在这

个框架的基础上进行改进的。本文的主要研究流程图如图 1-1 所示。

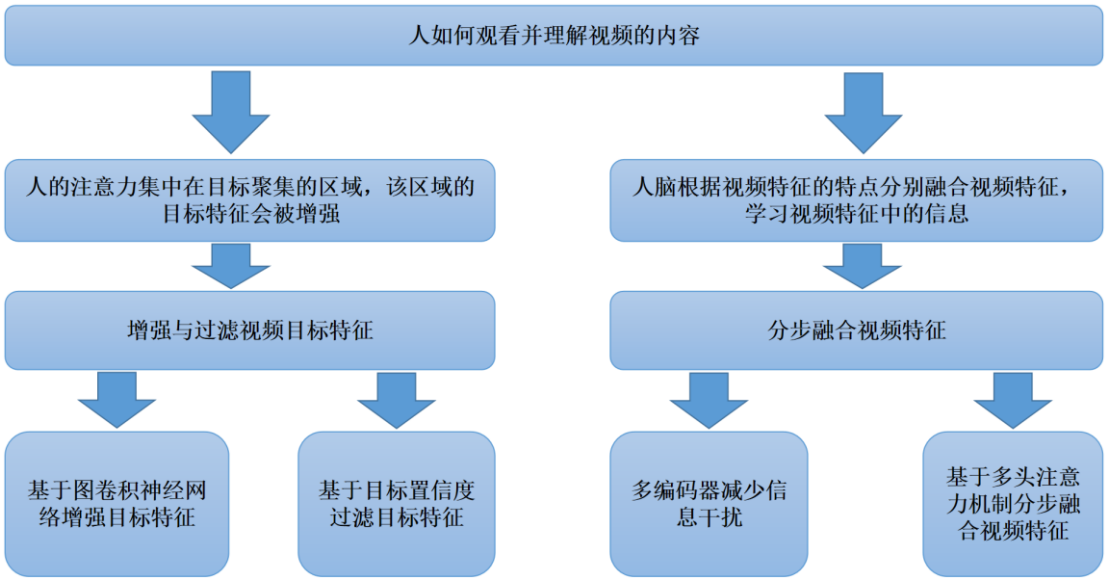

图 1-1 本文研究流程图

本文基于图卷积神经网络与多头注意力机制，针对视频描述任务进行了如下研究工作:

（1）引入 Transformer 模型用于生成描述文本：传统使用 RNN 生成视频描述文本存在一些挑战，例如内存限制阻碍了长序列训练样本的批处理，而且训练样本难以并行化。为了解决这些问题，谷歌团队提出了 Transformer 模型[19]，依赖于注意力机制，该模型可以有效地建模输入和输出之间的全局依赖关系。使用 Transformer 模型可以克服 RNN 模型无法并行计算的限制，并且自注意力机制可以使结果更具可解释性。Transformer 模型利用多头注意力机制使模型关注不同的位置，在机器翻译等自然语言处理领域取得了不错的效果。因此，为了提取输入的不同特征的关键信息，本文使用 Transformer 模型生成描述文本，并根据不同特征分布应用注意力机制。

（2）基于图卷积神经网络增强目标特征：人们对于结构化数据的处理能力不断提升，但是在现实生活中，很多数据是非结构化的。传统的 CNN 无法对非结构化数据进行局部卷积特征提取，因为非结构化数据中每个顶点的邻居结点个数可能不同，不能使用传统的离散卷积提取特征，而使用图神经网络处理图结构数据更为合适。对于视频的目标特征，可以将它们视为图结构数据中的节点，并搭配目标置信度来进一步处理目标特征。因此，本文第一个模型的核心思想是首先将目标特征构建成图结构数据，然后使用设计的模块增强其特征，并根据目标置信度对其进行过滤。

（3）基于多头注意力机制分步融合视频特征：传统视频特征融合方法过于

简单，不利于提取视频语义信息。同时，这种方法也会造成信息干扰和模型开销大。为了解决这些问题，本文第二个模型采用多头注意力机制来合理地处理多个输入特征，并根据输入特征选择编码器，以减少模型开销。通过分步融合从视频中提取的语义信息，该模型能够生成更加合理且具有可解释性的描述文本。

1.3.2 本文的主要创新

本文提出的基于特征强化与特征融合的视频描述方法在 MSVD 和 MSR-VTT 两个公共数据集上取得了较好的效果，本文的创新点主要体现在以下几个方面：

（1）基于图卷积神经网络增强目标特征。本文针对视频中存在目标特征不完整和不准确的问题，提出了两个解决方案。首先，本文基于图卷积神经网络，在视频中相邻区域之间搭建通信网络，以完善目标的低层视觉信息，增强目标特征。其次，针对不准确的目标特征，本文采用目标置信度标记目标特征，并结合多头注意力机制，动态学习每个被捕获区域中的目标特征，从而降低不准确目标特征对模型理解视频内容的影响。

（2）基于多头注意力机制分步融合视频特征。本文解决视频特征融合问题的方法是使用多个编码器对视频特征进行分别编码，然后通过多头注意力机制分步融合编码后的特征，以减少视频特征在被编码时信息干扰。为了减少模型结构复杂度，本文使用了 SRU（Simplified Recurrent Unit）代替原本的 transformer 编码器，SRU 可以编码视频目标之间的关系特征，并缩短模型的训练时间，提高模型性能。这样做的好处是能够有效地处理视频特征融合问题，同时减少了不必要的计算开销，提高了模型的效率和性能。

1.3.3 本文的章节安排

本文章节安排如下：

第 1 章主要讲述了本文研究的背景、意义和现状，以及介绍了视频描述任务的相关信息。在第 1 节中，介绍了短视频在当今社交网络中的流行趋势，以及视频描述任务在不同领域中的应用方向。在第 2 节中，对国内外视频描述任务的研究现状进行了归纳和总结。最后，在第 3 节中，讲解了本文的研究内容和主要创新点，并对文章结构进行了概述。

第 2 章介绍了本文使用的相关方法和技术的理论基础。在第 1 节中，对视频描述任务的特征提取方法进行了列举，包括对本文使用的两个预训练卷积神经网

络结构的分析，以及对用于提取目标特征的 Mask R-CNN 网络结构和提取目标关系特征的 OpenKE 框架整体结构的讲解。第 2 节对本文的模型主体 Transformer 的理论基础进行了阐述，解释了 Transformer 的编码器和解码器结构以及其在序列到序列学习中的应用。第 3 节探讨了图卷积神经网络的网络结构，详细说明了其在处理具有复杂结构的数据时的优势。第 4 节对本文使用的简易循环单元的网络结构进行了说明，包括其与传统循环神经网络的区别以及其在减少模型开销和提高训练速度方面的优点。

第 3 章主要介绍一种基于图卷积神经网络的视频描述方法，通过增强视频中目标的特征信息来提高描述的质量。第 1 节介绍了本文提出的 EFbiT-VC 模型的总体思路和结构图。第 2 节详细阐述了模型的具体结构，包括目标特征增强模块、目标信息过滤模块、编码器和解码器四部分。第 3 节列举了实验的细节设置，包括使用的数据集、评价指标和模型参数配置等。第 4 节分析了 EFbiT-VC 模型的实验结果，证明了本文方法的有效性。

第 4 章介绍了一种基于多头注意力机制进行分步融合视频特征的视频描述方法——SFF-MA 模型，同时也介绍了其实验细节和实验结果分析。第 1 节中，概述了 SFF-MA 模型的总体思路和总体结构；第 2 节详细阐释了 SFF-MA 模型的具体结构；第 3 节首先列举了 SFF-MA 模型的参数配置，随后展示和分析了 SFF-MA 模型的实验结果，并通过与现有方法的对比实验以及消融实验验证了分步特征融合模块的有效性。

第五章是对本文工作的总结与展望。

2 理论基础

2.1 视频特征提取技术

2.1.1 Inception-ResNet-V2 网络

在 ResNet[20]网络被提出之前，增加神经网络的深度会导致网络性能下降的问题。这是因为反向传播调整网络参数时，可能会出现梯度爆炸或梯度消失的问题，因此神经网络的训练过程可能难以收敛。为了解决这个问题，ResNet 提出了一种解决方法，即在残差模块内部加入跳跃连接，将前面层的信息直接传递到后面层，这样即使在更远的距离，网络也不会忘记先前的信息，从而成功地克服了神经网络训练中梯度消失和梯度爆炸的挑战。

由于 ResNet 解决了网络层数过深的问题，因此现在许多卷积神经网络通过增加网络深度来提高性能。然而，Inception 网络[22][23]提出了一种不同的方法，即通过扩展网络宽度来提高性能。如图 2-1 所示，Inception 块是一个基本的结构，它能够同时进行多个卷积或池化操作，然后将它们的输出整合成一个大的特征图，从而加速神经网络的计算。这种设计的原因是不同尺度的卷积和池化操作在输入图像上关注的区域不同，将这些不同尺度的图像信息组合起来相当于融合了多种表达，从而得到更加合理、丰富的图像表达。

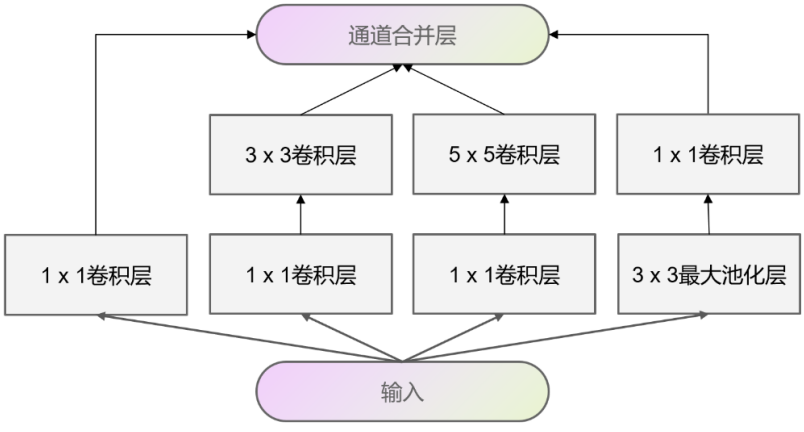

图 2-1 Inception 块结构

Inception 模块通过在同一层次上设置多个尺寸的过滤器来实现更宽的网络，从而提高性能。然而，随着残差网络的兴起，Google 团队提出了一种名为混合 Inception 的改进版本，即 Inception-ResNet-V2[24]网络（图 2-2）。混合 Inception 模块采用了与传统 Inception 模块相似的并行卷积的思想，但它还在这些并行卷积中加入了残差连接，从而使网络更加深层次，此方法同时还解决了梯度消失和

梯度爆炸等问题。此外，混合 Inception 还采用了 1x1 的卷积层来降低计算量和模型参数数量，提高了网络的效率和准确性。

Inception-ResNet-V2 网络有以下几点改进：

1. 使用批归一化（Batch Normalization）层进行归一化处理；
2. 通过借鉴 VGG 网络的设计思路，可以用两个 3×3 的卷积层串联代替 Inception 模块中的 5×5 卷积模块，这样既可以降低计算复杂度，又可以增加网络的非线性拟合能力；
3. 使用非对称卷积将 3×3 的卷积操作进一步分解为两个操作：一个 3×1 的卷积和一个 1×3 的卷积。这种方式可以进一步提高卷积层的效率和准确性，同时还可以减少卷积操作的参数数量。

其整体流程如图 2-2 所示：

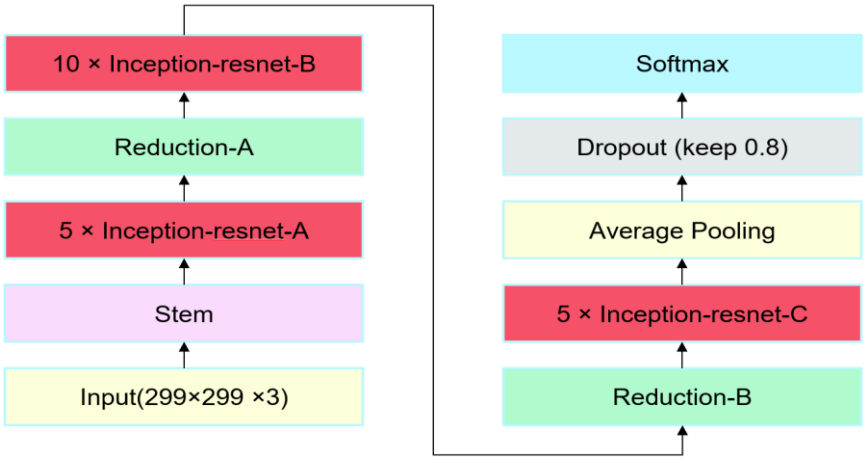

图 2-2 Inception-Resnet-V2 整体流程

其中 Inception-resnet-A、Inception-resnet-B、Inception-resnet-C 的结构如图 2-3 所示。

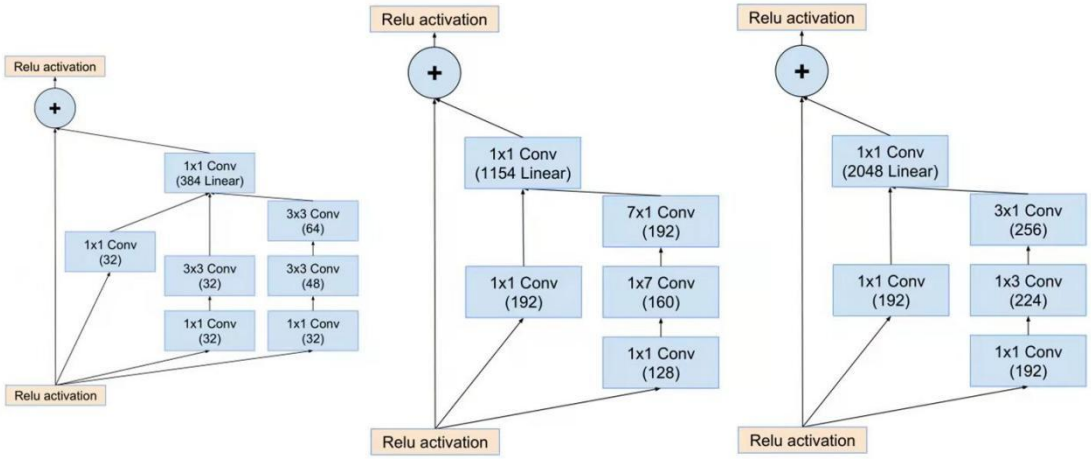

图 2-3 Inception-resnet-A、B、C 的结构

本文选用 Inception-ResNet-V2 网络作为视频空间特征提取的方法，是因为

该网络在图像特征提取方面具有出色的性能，可以很好地适应视频特征提取任务。

2.1.2 I3D 网络

I3D 网络[25]是 DeepMind 团队于 2017 年提出的一种方法，旨在通过将在 2D ImageNet 数据集上使用的滤波器和池化内核替换为 3D 结构，从而更好地提取视频的时序特征。I3D 网络使用 3D 卷积操作来处理视频数据的时空特征，可以看作是对 2D 卷积的扩展。通过将 I3D 网络与现有的 2D 卷积网络进行比较，研究表明，在视频分类和动作识别等任务中，I3D 网络具有更好的性能。I3D 网络结构如图 2-4(e)。

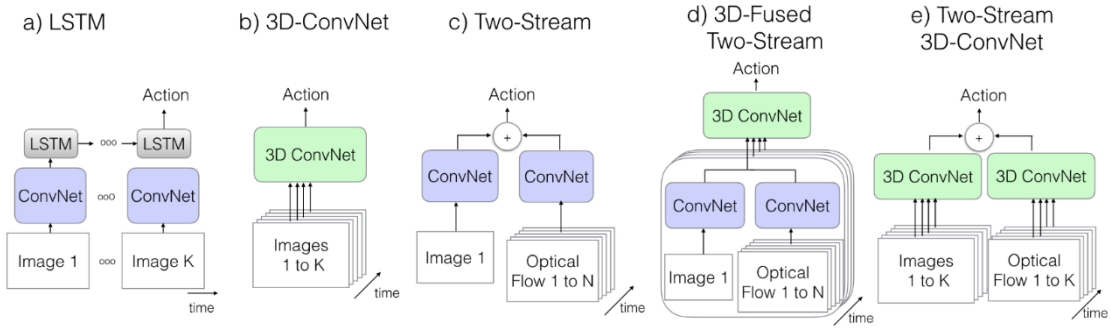

图 2-4 I3D 网络与先前图片特征提取网络的结构对比

图 2-4 比较了不同视频特征提取方法的结构。LSTM 方法是将图像分类方法直接应用于视频，将每一帧看作独立的部分，没有利用视频的时间序列信息；3D-ConvNet 方法使用 3D 卷积进行视频处理，将图像转化为视频，但是 3D 网络的训练成本更高，参数更多，且无法与在图片数据集上预训练的 2D 网络共享参数；Two-Stream 方法使用 RGB 帧和光流特征作为输入，以更好地提取时间和空间信息；最后两种方法是将在图像分类任务中表现良好的模型扩展为 3D 模型，以便可以直接使用预训练的模型参数进行初始化。

总的来说，由于在 Kinetics[27]数据集（人类行为识别数据集）上进行了预训练，I3D 方法在大多数据集上都取得了很好的性能，因此，本文决定采用 I3D 方法来提取视频的时间特征。

2.1.3 Mask R-CNN

Mask R-CNN 是一种基于 Faster R-CNN 的改进模型，主要用于解决实例分割任务。与 Faster R-CNN 不同的是，Mask R-CNN 在 RPN（Region Proposal Network）之后增加了一个分支，用于生成每个候选区域的二进制掩码。He 等人[32]在 Mask R-CNN 的基础上提出了更快的 R-CNN 和 FPN[37]（Feature Pyramid Network）主

干，用于进行更快速和准确的特征提取和实例分割。

如图 2-5 所示，Mask R-CNN 主要由以下 4 个模块构成：

1. 主干卷积网络（convolutional backbone）：可以使用各种图片特征提取网络作为特征提取网络，用于生成特征图（feature map），例如特征金字塔网络（Feature Pyramid Network，FPN）等；
2. 区域候选网络（Region Proposal Network, RPN）：其功能是根据特征图生成潜在的目标物体位置和大小的候选框（bounding boxes），也被称为目标提议（object proposals）或候选框（bounding box proposals）；
3. 目标框分支（Box head）：预测目标框的分类和回归；
4. 语义分割分支（Mask head）：通过对感兴趣区域进行精确的空间对齐（ROIAlign），Mask R-CNN 能够实现在特定区域内进行语义分割任务，从而更加准确地确定物体的位置。

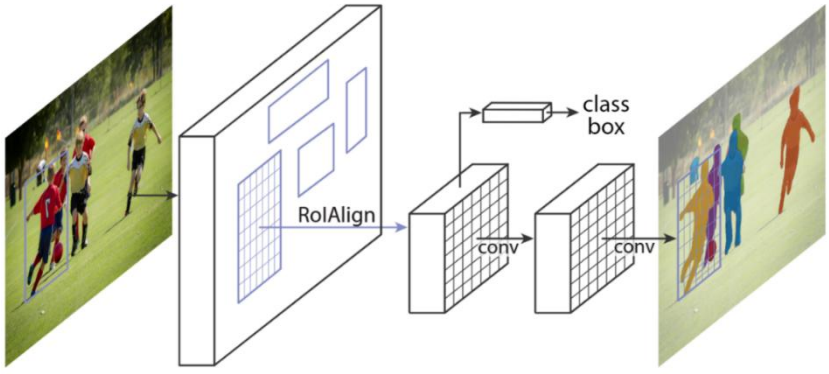

图 2-5 Mask R-CNN 框架整体结构

Mask R-CNN 是一种将目标检测和语义分割相结合的模型，它集成了多种先前模型的优点。它使用了 R-CNN 中的边界框回归技术，同时利用了 Fast R-CNN 中提出的感兴趣区域对齐（ROI Align）和优化感兴趣区域池化（ROI Pooling）方法。此外，它还继承了 Faster R-CNN 中的区域生成网络（RPN）。通过增加语义分割分支，Mask R-CNN 实现了对类别预测和语义分割的解耦，使得它在性能上有了显著的提升。

由于 Mask R-CNN 的成熟性和易用性，本文选择它作为目标检测框架，以提取目标特征。

2.1.4 OpenKE

OpenKE[46]是一个开源框架，由 THUNLP 团队基于 TensorFlow 和 PyTorch 开发，用于将知识图谱表示为低维连续向量空间中的嵌入。该框架提供了稳定且快速的各种接口，并实现了多个经典的知识表示学习模型。OpenKE 还具有易于

扩展的特点，因此使用该框架设计新的知识表示模型非常方便。具体来说，OpenKE 具有如下特点：

（1）接口设计简单，可以轻松在各种不同的训练环境下部署模型。

（2）底层的数据处理进行了优化，模型训练速度较快。

（3）提供了轻量级的 C++模型实现，在 CPU 多线程环境下也能快速运行。

（4）提供了大规模知识图谱的预训练向量，可以直接在下游任务中使用。

（5）长期的工程维护来解决问题和满足新的需求。

以下是改写后的内容：OpenKE 框架如图 2-6 所示，其整体设计分为三层：底层数据处理、中层模型构建和上层训练/评估策略。每个功能块都提供了足够的封装，以确保调用的便捷性。用户可以通过简单的代码调用不同层次的模块，最终达到训练和部署知识图谱表示学习模型的目的。因此，在本文中，我们选择使用 OpenKE 作为目标关系特征提取器。。

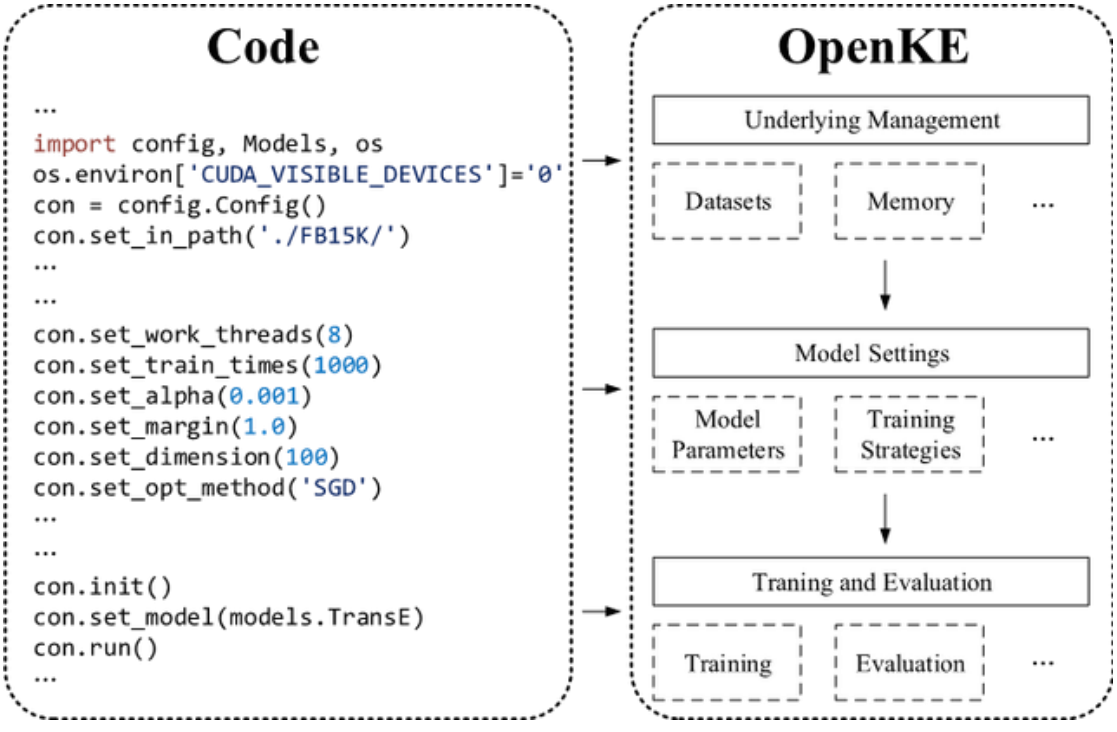

图 2-6 OpenKE 框架的整体结构

2.2 Transformer 网络

CNN-RNN 结构是视频描述领域最为常见的模型，但其存在一些问题。RNN 模型在处理句子时是逐词进行的，根据当前节点的隐藏状态和输入词更新节点的隐藏状态。因此，该模型只能按顺序一个一个地处理单词，无法并行处理，训练和推理的时间较长。此外，由于循环神经网络固有的长期依赖性，该模型也很难

解决长序列的问题。

为了解决长期依赖性问题，Transformer[19]模型采用了多头注意力机制。该模型有以下优点：

（1）每层网络的计算时间复杂度优于 CNN 和 RNN;

（2）可直接计算点乘结果;

（3）通过一步矩阵计算来解决长时依赖问题。与 CNN 需要增加卷积层数来扩大视野，RNN 需要逐个计算从第 1 步到第 n 步的不同状态不同，自注意力机制只需要通过一步矩阵计算即可处理所有状态。因此，相较于 RNN，自注意力机制能够更有效地处理长期依赖问题。在序列长度 n 大于序列维度 d 导致计算量过大时，可以通过限制窗口大小来控制自注意力机制的计算量。

（4）模型可解释性强。注意力计算结果的分布表明了该模型学习到了一些语法和语义信息，增强了模型的可解释性。

如图 2-7 所示，Transformer 模型的核心是三种不同的多头注意力操作：编码器中的多头注意力、解码器中的遮蔽多头注意力、解码器中的多头注意力。这些操作的过程可以简单地理解为一种输入特征的转换方式，通过特征之间的关系（注意力）来增强或减弱特征中不同维度的权重。这种转换能够有效地捕捉到输入序列之间的依赖关系，使得 Transformer 模型在处理长序列时更加高效和准确。

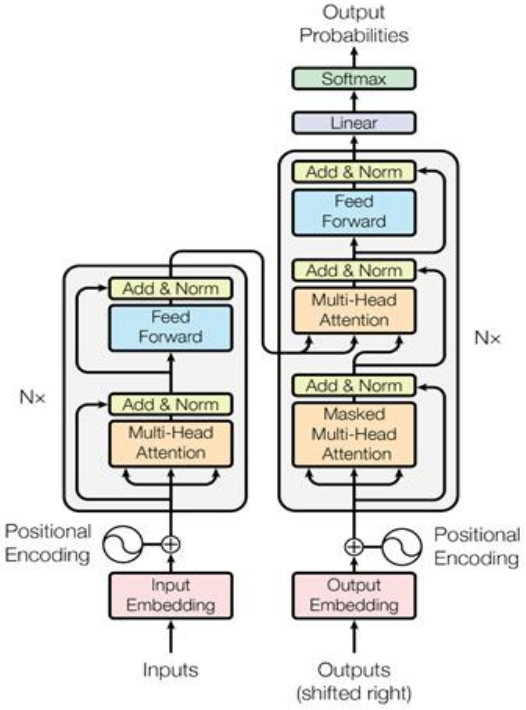

图 2-7 Transformer 网络整体结构

除了主要的多头注意力操作，Transformer 模型还通过位置编码、残差连接、层归一化、dropout 等操作将输入、注意力和多层感知器连接起来，以解决

长距离依赖问题，并取得了出色的成绩。因此，在本文中，Transformer 被选为基准模型。

2.3 图卷积神经网络

图数据是一种数据结构，用于表示实体之间的关系。在图中，实体表示为节点(nodes)，它们之间的关系表示为边(edges)。然而，许多现实世界中的数据是非欧几里得数据，如图 2-8 所示，这些数据的顶点邻居数量可能不同，传统的卷积神经网络无法在这些数据上执行局部卷积以提取特征。因此，可以处理这些非欧几里得数据的图神经网络应运而生。除了图形卷积以外，图神经网络还具有一些其他操作，例如节点嵌入、图注意力和图池化等。

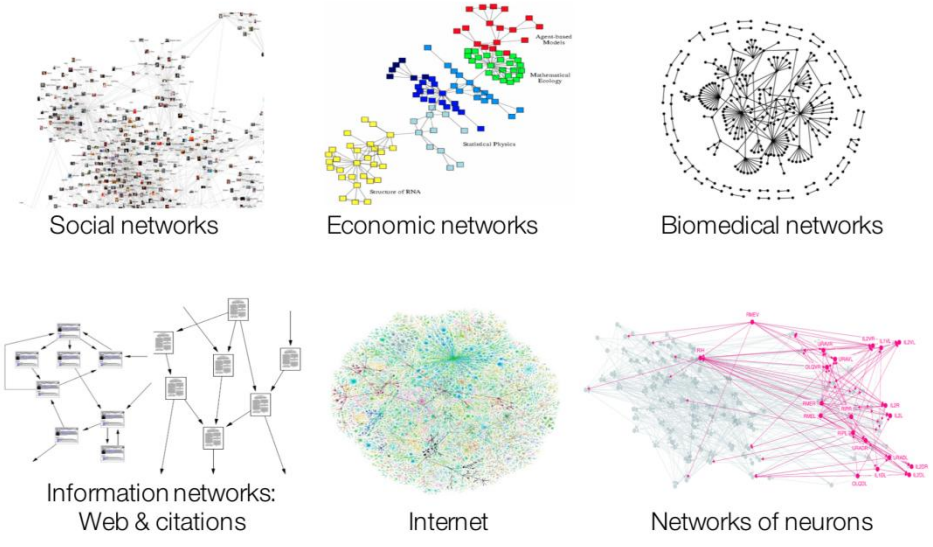

图 2-8 现实中的图数据

在以图像为代表的欧式空间中，节点的邻居数量是固定的，例如绿色节点始终有 8 个邻居节点（边缘上的节点可以通过填充进行补齐）。但在非欧式空间中，如图 2-9 所示，节点的邻居数量并不固定。例如，绿色节点当前有 2 个邻居节点，但其他节点可能有 5 个邻居节点。在欧式空间中，卷积操作是通过可学习的固定大小卷积核提取像素的特征，例如在图像中，绿色节点及其相邻节点的特征会被提取。然而，在非欧式空间中，由于节点的邻居数量不固定，传统的卷积核无法直接用于提取节点的特征。

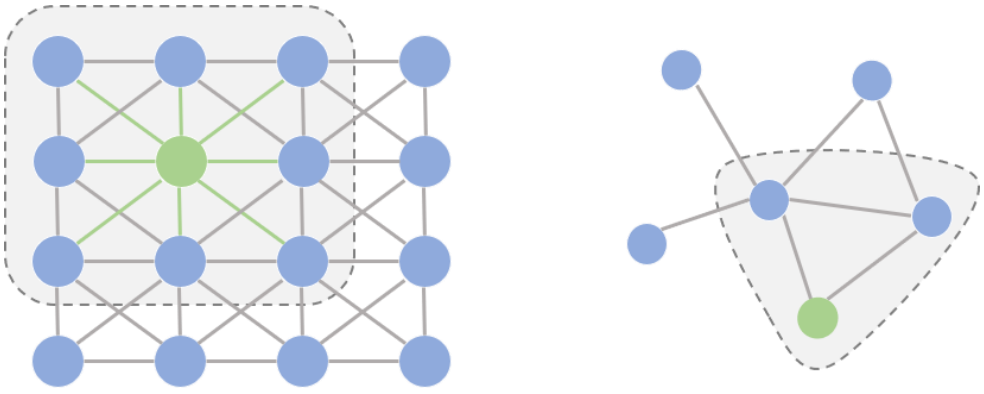

图 2-9 非欧空间中的节点

图卷积的关键思想在于通过考虑节点之间的边关系，从而生成新的节点表示。GCN（Graph Convolutional Neural Networks）[31]的主要目的是提取拓扑图的空间特征。

图卷积神经网络可以分为基于空间域或顶点域(vertex domain)和基于频域或谱域(spectral domain)的两类。例如，在图 2-10 中，我们可以将左边的图表示转换为右边的形式。以节点 A 为例，A 依赖于 B、C、D；而 B 依赖于 A、C，C 依赖于 A、B、E、F，D 依赖于 A，A 是由 B、C、D 通过神经网络生成的，即将它们的特征取平均值。因此，可以直观地理解图神经网络是图拓扑结构和神经网络的结合。GNN 的本质是聚合邻居节点的信息，对于图中的任何节点，每次更新节点特征时，就会聚合更高阶邻居节点的信息。

综上所述，本文使用图卷积神经网络增强目标特征。

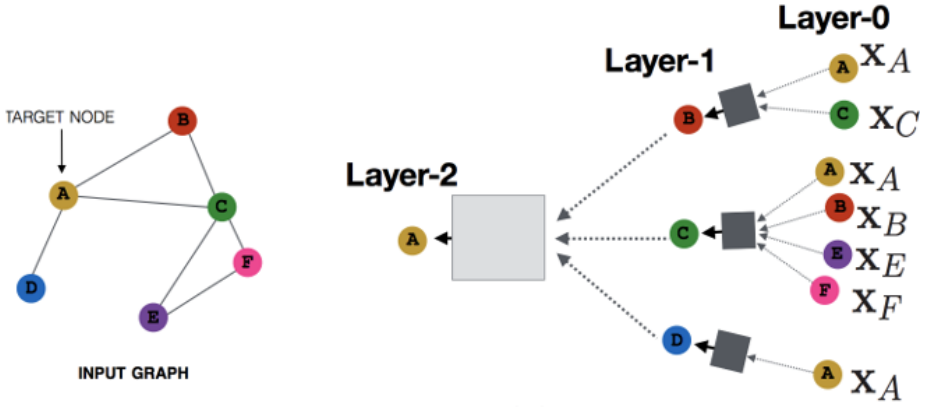

图 2-10 空间域卷积

2.4 简易循环单元

为了解决训练深度学习模型时计算速度慢的问题，使用 GPU 进行加速训练的并行化方法已经被广泛采用，在使用 GPU 进行加速的卷积神经网络中，训练

速度得到了显著提升。然而，像 RNN 和 LSTM 这样的模型无法实现并行化方法，这是因为在 RNN 和 LSTM 的计算过程中，需要等待前一时刻的计算结果才能计算当前时刻的结果，这限制了其实现并行化处理。Lei 等人[33]提出的 SRU（Simple Recurrent Unit）解除了这种限制，如图 2-11 所示。SRU 的结构不再依赖于前一时刻的计算，因此可以实现并行化处理,它的训练速度比 LSTM 快，能够达到与 CNN 相同的训练速度。

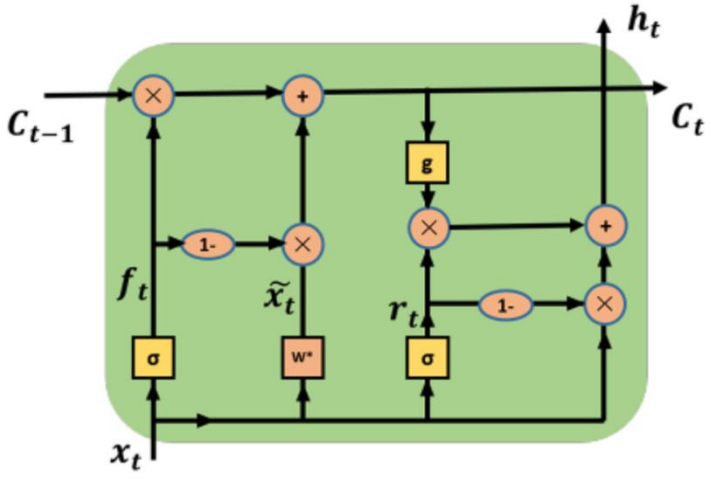

图 2-11 SRU 结构

递归神经结构通常由于其难以并行计算的特点而难以扩展，而 SRU 的设计目标是一个提供表现力强大的递归模型，它支持高度并行的实现，并且具有一定范围的初始化，以方便深度模型的训练。Lei 等人[33]验证了 SRU 在多个 NLP 任务上的有效性，与 cudnn 优化的 LSTM 相比，SRU 在分类和问题回答数据集上实现了 5 到 9 倍的加速，并提供了比 LSTM 和卷积模型更强的性能。

SRU 有以下几点优势：

（1）SRU 的并行性与卷积和前馈网络相同，这是通过平衡序列依赖和独立性来实现的，虽然 SRU 的状态计算是时间依赖的，但每个状态维度是相互独立的。这种简化使得 SRU 的计算可以跨越隐藏维度和时间步长进行 cuda 优化，并有效地利用现代 GPU 的全部容量。

（2）SRU 的使用取代了卷积操作。和 QRNN、KNN 一样，SRU 具有更多的循环连接，这种设计保持了模型的建模能力，同时使用了更少的计算资源和超参数。

（3）SRU 采用了高速公路连接和特殊的参数初始化方案，以改进深度递归模型的训练，并减少梯度传播的问题。

综上所述，本文使用 SRU 提取目标关系特征的语义信息。

2.5 本章小结

本章主要对视频描述任务需要使用的一些基本技术和已有开源工具进行阐述。第 1 节探讨了视频描述任务中提取视频特征的四个预训练神经网络；第 2 节阐述了本文所使用的编码器和解码器框架——Transformer 网络；第 3 节介绍了本文增强目标特征的神经网络——图卷积神经网络的基本结构；第 4 节讲述了本文提取目标关系特征语义信息的神经网络——简易循环单元的基本结构。

3　基于增强与过滤目标特征的视频描述方法

由于视频中存在多个目标，观众会将注意力集中在目标数量较多的区域，从而导致该区域内的目标特征被额外增强，但这些特征不一定准确反映视频的真实内容。为了解决这一问题，本文提出了一种基于图卷积网络和目标置信度的视频描述方法，旨在增强和过滤目标特征。该方法根据目标特征密度程度的不同，在图卷积神经网络中增强部分目标特征，并使用目标置信度标记增强后的特征，以过滤目标特征，这些举措有望提高视频描述的准确性和可读性。该方法简称EFbiT-VC（Enhanced and Filtered bi-Transformer for Video Caption）。

3.1　概述

3.1.1 总体思路

生成视频描述是一个序列到序列的过程，需要设计一个端到端的模型，通过给定视频的帧序列，生成对应视频内容的描述。这个过程可以分为两个阶段：视频特征提取和描述语句生成。在视频特征提取阶段，需要提取视频的 2D 特征、3D 特征、目标特征和目标之间的关系特征。在描述语句生成阶段，根据提取的视频特征和模型在训练阶段学习到的视频高层语义规律，生成关于视频内容的描述。模型生成视频描述的流程如图 3-1 所示。

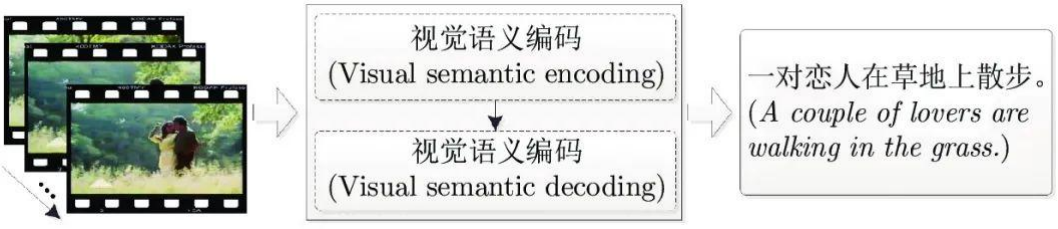

图 3-1 生成视频描述的流程

为了增强视频目标特征，本文对传统视频中目标特征的使用进行了分析。

首先，现有大部分研究在学习视频目标特征以理解视频内容时，往往忽略了这些目标特征的不完整性。在检测视频中的目标时，通常分为两个阶段：第一阶段扫描视频的关键帧以捕获可能存在目标的区域，第二阶段对被捕获到的区域进行分类，生成边界框和掩码。虽然从分割成多个区域的关键帧中提取目标特征可以降低受视频背景噪声影响的程度，但关键帧被分割后，各个区域中的目标特征就失去了一些关于视频的真实信息，导致目标特征无法发挥出应有的作用，使得模型无法通过学习目标特征充分理解视频内容，如图 3-2 所示。

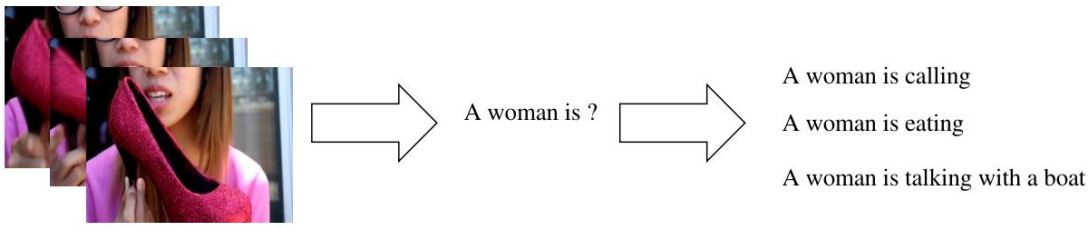

图 3-2 不准确的目标特征误导模型生成错误的描述

其次，视频中检测到的目标特征可能包含虚假信息，导致模型学习错误的目标特征。在视频目标检测的第二阶段，预训练的目标检测模型被用来对被捕获的区域进行分类，并提取目标特征及对应置信度。虽然目前预训练的目标检测模型在目标识别任务上已达到人类水平，但同一个区域仍可能被检测模型识别成多个目标类别。这种情况下，多数的目标特征可能是不准确的，使得模型难以从中学习到有关该区域的真实信息，严重影响了模型对视频内容的理解。

本章研究的重点在于改善视频目标特征的不完整性和不准确性。为此，本文采用图卷积神经网络，在相邻区域之间建立通信网络，从而完善目标的低层视觉信息，强化目标特征。同时，本文利用目标置信度标记目标特征，并采用多头注意力机制，动态地学习每个被捕获区域中的目标特征，以降低不准确的目标特征对模型理解视频内容的影响。。

3.1.2 模型总体结构

针对现有基于目标检测的视频描述模型中出现的目标特征不完整和目标特征中混杂错误的信息的问题，本文在 BiTransformer 的研究工作基础上，提出基于图卷积神经网络的目标特征增强与过滤双向 Transformer 视频描述模型（Enhanced and Filtered bi-Transformer Video Caption，简称为 EFbiT-VC)，模型的结构如图 3-3 所示。EFbiT-VC 由四部分组成：

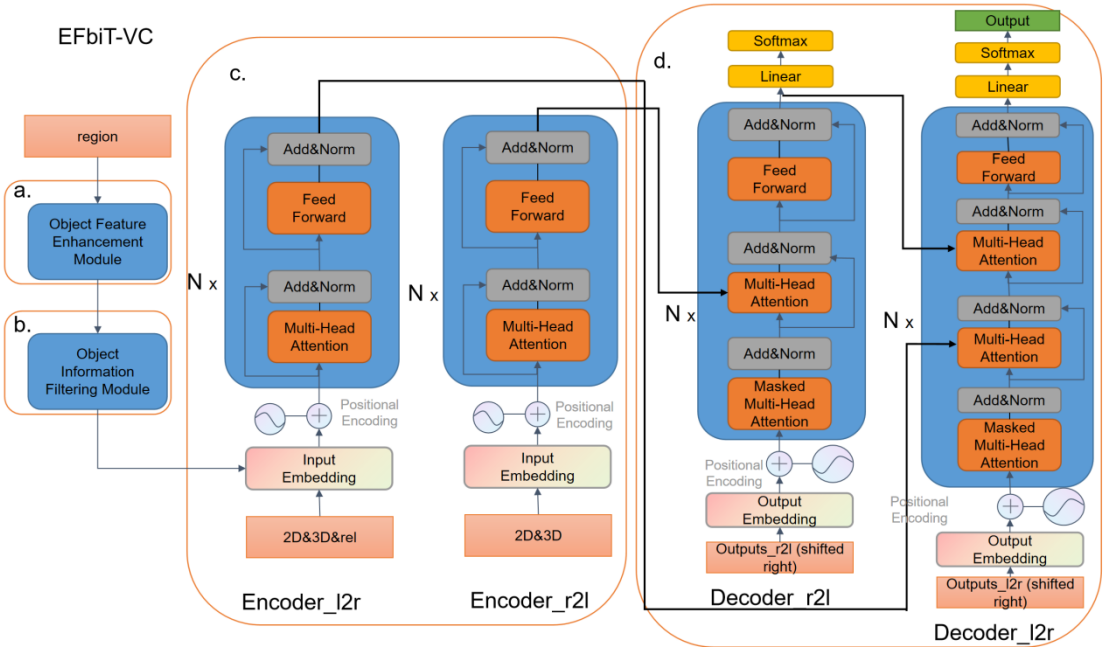

图 3-3 基于增强视频目标特征的视频描述方法的整体结构

（a）目标特征增强模块（Object Feature Enhancement Module，简称为 OFEM）：基于图卷积神经网络增强目标特征，改善了目标特征不完整的问题；（b）目标信息过滤模块（Object Information Filtering Module，简称为 OIFM）：基于目标置信度过滤目标特征，降低不准确的目标特征造成的影响；（c）双向编码器（Bidirectional Encoder，简称为 BE）：减少视频时序特征和非时序特征之间的信息干扰；（d）双向解码器：学习视频的正向信息和反向信息。

3.2 模块设计

3.2.1 目标特征增强模块

本文使用 Mask-RCNN 提取视频关键帧中的目标特征，但这些特征可能不完整且不可靠，因此需要补全和增强目标特征。由于某些被捕获区域之间具有空间上的依赖关系，本文将从这些区域提取的目标特征作为图结构数据的节点，并建立区域之间的联系，构建图结构数据。目标增强模块主要包括三个步骤：第一步，将每个节点的特征信息转换并发送给与之有联系的邻居节点；第二步，聚集节点接收到的特征信息，以融合节点的局部结构信息；第三步，对聚集的信息进行非线性变换，以增加模型的表达能力。通过应用两个图卷积层和一个 ReLU 函数激活层，本章模型增强了节点的特征，从而增强了目标特征。目标增强模块的工作流程如图 3-4 所示。

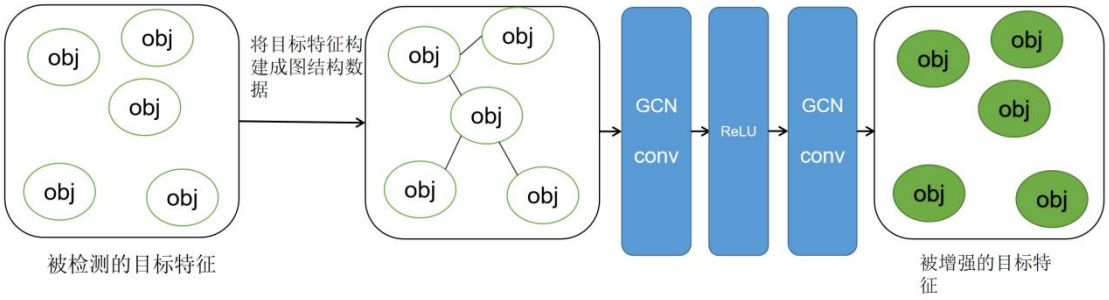

图 3-4 目标特征增强模块工作流程

目标特征增强模块的计算如公式(3-1)、(3-2)所示：

$$OFEM(v^o,r) = GCN(ReLU(GCN(v^o,r)),r) \tag{3-1}$$

$$GCN(X^{(L)},A) = \sigma(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}X^{(L)}W^{(L)} + b^{(L)}) \tag{3-2}$$

其中F_{obj}是目标特征,r 是节点之间的依赖关系，σ和 ReLU 是激活函数，$X^{(L)}$是节点在 GCN 第 L 层的特征，D 是度矩阵，$W^{(L)}$是第 L 层的权重，$b^{(L)}$是第 L 层的偏置。

本文在使用图卷积神经网络增强目标特征之前需要解决两个问题。首先，需要确定节点之间的联系是什么。视频通常描述视频中多个区域内目标的互动事件。当多个区域进行互动时，它们的空间距离通常较近。因此，本文将区域在空间层面上的依赖关系作为节点之间的联系，并将图结构数据的节点用于存储目标特征。其次，需要确定图结构数据中具有联系的区域的数量。目标检测模型可以从一个视频关键帧捕获到多个包含目标的区域。如果所有被捕获区域都被认为有联系，那么过多的区域交流信息可能会出现信息干扰的情况。因此，本文进行了实验，验证了图结构数据中具有联系的区域个数对模型性能的影响。结果表明，当区域个数为 3 时，模型性能最佳。详细的分析过程，请参见 3.4.2 中的消融实验部分。

3.2.2 目标信息过滤模块

Shen 等人[34]在使用目标特征时，忽略了目标特征的可信程度这一条件，直接将目标特征作为模型输入之一，造成模型学习到错误的目标特征。如果使用单个被置信度标记的目标特征，会导致最终的实验效果并不好，主要原因是单个被标记目标特征无法表现同一区域被检测目标的相同低层视觉信息。因此，该文在使用目标置信度标记目标特征的基础上，结合多头注意力机制提取局部的高层语义信息，有选择地学习每个区域的目标特征，充分发挥目标特征的作用

为了解决目标特征不准确的问题，本文使用基于目标置信度和多头注意力机

制的目标信息过滤模块来过滤被增强的目标特征。在对被捕获区域进行分类时，目标检测模型根据被捕获区域信息在不同语义层次上的表达，将其分类为多个目标，并生成相应的置信度。置信度数值越高，说明目标检测模型认为该目标越有可能是被捕获区域的真实目标特征。虽然置信度可以作为模型是否学习到目标特征的参考，但置信度比较低的目标特征也有可能是被捕获区域的真实目标特征。因此，本文使用置信度对目标特征进行标记，并结合多头注意力机制来综合学习目标特征。这样可以让模型有针对性地学习置信度比较高的目标特征，而目标置信度较低的目标特征则作为参考信息，辅助模型理解被捕获区域的内容。目标信息过滤模块的结构如图 3-5 所示。

目标信息过滤模块的计算如公式(3-3)～(3-5)所示：

其中$v^{``o}$表示经过 OFEM 处理并且被目标置信度标记的目标特征，Concat表

$$\mathrm{MultiHead}(v^{``o},v^{``o},v^{``o}) = \mathrm{Concat}_{i=1,\ldots,h}(\mathrm{head}_i)W_1 \tag{3-3}$$

$$\mathrm{head}_i = \mathrm{Attention}(v^{``o}W_i^Q,\ v^{``o}W_i^K,\ v^{``o}W_i^V) \tag{3-4}$$

$$\mathrm{Attention}(Q_{obj},K_{obj},V_{obj}) = \mathrm{softmax}(\frac{Q_{obj}K_{obj}^T}{\sqrt{d}})V_{obj} \tag{3-5}$$

示拼接操作，W_1、W_i^Q、W_i^K、W_i^V表示可训练的权重参数，softmax 是激活函数，d 表示 Q、K、V 的维数。

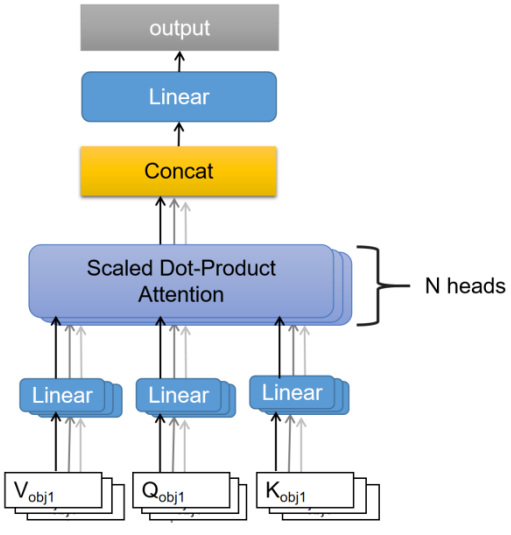

图 3-5 目标信息过滤模块的结构

3.2.3 编码器

假设一段输入视频$F = \{f_i\}_{i=1}^N$,通过预训练的 2D-CNN$\emptyset^a$和 3D-CNN$\emptyset^m$提取

其中的视频特征，提取特征的过程如公式(3-6)、(3-7)所示:

$$v_i^a = \emptyset^a(f_i) \tag{3-6}$$

$$v_i^m = \emptyset^m(c_i) \tag{3-7}$$

其中f_i表示平均采样的采样帧，c_i表示个连续帧$\{c_i\}_{i=1}^N$，这个连续帧由f_i周围的连续帧组成，$\emptyset^a$采用了 Inception-ResNetV2 网络，$\emptyset^m$采用了 I3D 网络。

获得视频的 2D 特征v_i^a与 3D 特征v_i^m后，本文使用 Mask R-CNN 捕捉视频的目标特征 $R_v = [v_1, v_2, \cdots, v_k]$、目标框的位置信息 $R_p = [X, Y, W, H] = [p_1, p_2, \cdots, p_k]$、目标类别 $R_t = [t_1, t_2, \cdots, t_k]$ 以及目标特征的置信度 $R_c = [c_1, c_2, \cdots, c_k]$。本文使用 OpenKE 根据$R_t$预测目标之间的关系$v^r$,并通过公式(3-8)可以得到拥有基本信息的目标特征v^o:

$$v^o = Concat_{i=1\cdots k}\left(R_v R_p\right) \tag{3-8}$$

其中 Concat 表示拼接操作。

获得视频的特征后，本文根据目标边框的位置信息 R_p 构建关于目标特征的邻接矩阵 r，使用目标特征增强模块对目标特征进行增强，计算过程如公式(3-9)所示：

$$v`^o = OFEM(v^o, r) \tag{3-8}$$

经过目标特征增强模块处理的目标特征拼接上目标置信度，作为目标信息过滤模块的输入，进行信息过滤，计算过程如公式(3-9)、(3-10)所示：

$$\overline{v^o} = \text{MultiHead}(v``^o, v``^o, v``^o) \tag{3-9}$$

$$v``^o = Concat_{i=1\cdots k}(v`^o R_c) \tag{3-10}$$

得到处理好的目标特征$\overline{v^o}$后，接下来将 2D 特征v_i^a、3D 特征v_i^a、目标特征$\overline{v^o}$的维度映射到d_{model}，这里使用线性变化，计算过程如公式(3-11)所示：

$$v_i' = w_v v_i + b_v \tag{3-11}$$

其中$w_v \in \mathbb{R}^{d_{\text{model}} \times d_I}, b_v \in \mathbb{R}^{d_{\text{model}}}$，最终得到映射向量$V_a' = \{v_1', \cdots, v_i'\}, v_i' \in \mathbb{R}^{d_{\text{model}}}$。

因为 2D 和 3D 特征仍然具有时间关系，本文使用位置编码来标记输入的视频特征的位置信息。如公式(3-12)~(3-14)所示：

$$PE(pos, 2i) = sin(pos/10000^{\frac{2i}{d_{model}}}) \tag{3-12}$$