文章编号　1004-924X(2022)24-3198-12

视频描述中链式语义生成网络

毛　琳，高　航*，杨大伟，张汝波
(大连民族大学 机电工程学院，辽宁 大连 116600)

摘要：针对视频描述中语义特征表达能力不足导致文本描述不准确问题，本文提出一种视频描述中链式语义生成网络(Chained Semantic generation Network，ChainS-Net)。构建了多阶段双路交叉的链式特征提取结构，该结构以全局域和局部域模块为基本单元，分别从视觉特征的全局和局部捕获视频语义；在网络的各阶段，将语义信息在全局域和局部域之间变换解析，实现视觉和语义信息的交互参考，提升语义特征表达能力；在此基础上，网络通过多阶段迭代的处理方式获取更为有效的语义表示，提升视频描述模型性能。在MSR-VTT和MSVD数据集上的实验结果表明，本文提出的链式语义生成网络ChainS-Net优于现有同类方法，相比于语义辅助视频描述网络(Semantics-Assisted Video Captioning network，SAVC)，视频描述的四个评价指标平均提升了2.5%。

关　键　词：视频描述；语义特征；全局；局部；域变换；多阶段
中图分类号：TP394.1；　　**文献标识码**：A　　**doi**：10.37188/OPE.20223024.3198

Chained semantic generation network for video captioning

MAO Lin，GAO Hang*，YANG Dawei，ZHANG Rubo

(*School of Electromechanical Engineering*，*Dalian Minzu University*，*Dalian* 116600，*China*)
* *Corresponding author*，*E-mail*：*gao_hang*2021@163.*com*

Abstract：Aiming to address the unsatisfactory expression ability of semantics，which results in inaccurate text descriptions in video captioning，a chained semantic generation network (ChainS-Net) for video captioning is proposed. A multistage two-branch crossing chained feature extraction structure is constructed that uses global and local domain modules as basic units and captures the video semantics from global and local visual features，respectively. At each stage of the network，semantic information is transformed and parsed between the global and local domains. This method allows visual and semantic information to be cross referenced and improves the semantic expression ability. Furthermore，it allows a more effective semantic representation to be obtained through multistage iterative processing，thereby improving video captioning. Experimental results on MSR-VTT and MSVD datasets show that the proposed ChainS-Net outperforms other similar algorithms. Compared with the semantics-assisted video captioning network，SAVC，ChainS-Net shows average improvements of 2.5% in four metrics of video captioning.

Key words：video captioning；semantic feature；global；local；domain transformation；multi-stage

收稿日期：2022-05-16；**修订日期**：2022-08-21.
基金项目：国家自然科学基金资助项目(No. 61673084)；辽宁省自然科学基金资助项目(No. 20180550866)；民族创新联合基金资助项目(No. 2020-MZLH-24)

1　引　言

视频描述任务是自动生成视频的概括性文本描述[1]，可以应用于短视频、安防监控和视频-文本双向检索等领域。由于语义特征具备较强的可解释性，在视频描述中常常被用来提升模型性能。当前大多语义检测网络的特征提取方式单一导致表达能力不足，影响文本描述效果且不利于实际应用，因此如何通过提升语义特征有效性来获取准确的文本描述是当前研究热点。

视频描述模型大多可以分为编码器和解码器两部分，编码器采用卷积神经网络（Convolutional Neural Network，CNN）[2]和循环神经网络（Recurrent Neural Network，RNN）[3]等提取视频特征，解码器则利用RNN对视频特征进行分析，生成文本描述结果。在图像描述任务的基础上，Subhashini等人首次提出针对视频序列的描述模型[4]，该网络利用CNN作为编码器对视频帧提取特征，由于长短时记忆网络（Long Short Term Memory Network，LSTM）[5,6]具有输入输出长度任意性和记忆能力等特点，被用于在解码阶段生成文本序列，这种编码-解码框架也成为此后主流方法。以该框架为基础，目标检测、语义检测和图卷积等编码方式[7-9]；门控循环单元（Gated Recurrent Unit，GRU）、双层LSTM和双向LSTM等解码方式不断提升视频描述性能[10]。其中语义特征是多标签分类形式的特征表示[11]，相比于其他特征，其具备较强的可解释性，可以作为一种编码特征提升视频描述效果。

为了利用这种语义分类信息，Rakshith等人采用支持向量机（Support Vector Machine，SVM）提取语义特征[12]，有选择地将语义特征、密集轨迹视频特征和关键帧特征作为输入，输出文本描述。Tu等人[13]采用Fast RCNN对视频帧提取表示目标的语义信息，通过Attention机制输入到LSTM中解码。目标语义信息可以相对准确地表示局部目标，但是忽略了视频中场景和行为等语义表示，存在表征能力不足的问题。Gan等人[14]由视觉信息捕获一种可以表示名词、形容词和动词等词汇的语义特征，更进一步概括了视频语义，在图像和视频描述任务中均取得了性能提升。Chen等人[8]提出语义辅助视频描述网络—SAVC，网络采用多层感知机（Multi-Layer Perceptron，MLP）提取词汇语义特征，以语义信息辅助视觉特征提升模型性能。但当前大多数生成语义特征的网络结构单一，且不能兼顾全局概括信息和局部细节信息，不能从多角度全面分析视频内容，导致语义特征表达能力有限，无法获取较准确的文本描述结果。

针对视频描述中语义特征表达能力不足而导致的文本描述不准确问题，本文从全局-局部多角度分析出发，提出视频描述中链式语义生成网络—ChainS-Net。设计了以MLP为基本单元的全局域和局部域双路结构，从全局和局部两个角度捕获视频语义。进一步结合域变换的思想，对语义特征交叉提取，分别将两种语义特征转换到另一个域中重新解析。最后对双路交叉的语义提取方式进行多阶段迭代，获取更准确的语义特征进而提升视频描述模型性能。

2　ChainS-Net

在深度学习任务中，对于特征的提取与增强，多角度分析是较为有效的手段之一，由于仅从单一角度对视觉内容进行解读（即单一处理方式或单一支路等），无法获取比较丰富的特征表示，或带来对内容的不全面和错误认知。为消除这种单一角度认知偏见，并进一步实现对特征的充分利用和增强，本文构建了一种用于深度学习特征提取与增强的链式网络结构，图1所示为网络结构的大体示意图，数学表达如式(1)和式(2)：

$$\begin{cases} \boldsymbol{x}_i^{\varphi}=\Phi\left(\boldsymbol{x}\oplus\boldsymbol{x}_{i-1}^{\psi}\right) \\ \boldsymbol{x}_i^{\psi}=\Psi\left(\boldsymbol{x}\oplus\boldsymbol{x}_{i-1}^{\varphi}\right) \end{cases},\ i\in\{2,3,\cdots,k\}, \tag{1}$$

$$\boldsymbol{x}'=\boldsymbol{x}_k^{\varphi}+\boldsymbol{x}_k^{\psi}, \tag{2}$$

其中：$\boldsymbol{x}$是输入的视觉信息（图像或相关特征等），$\boldsymbol{x}'$是增强后的目标特征，采用提出的链式网络获取较为有效的特征$\boldsymbol{x}'$。Φ和Ψ是网络中两种特

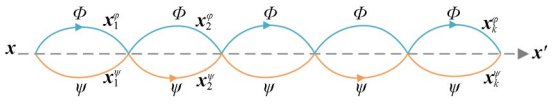

图1　链式网络结构示意图

Fig. 1　Structure diagram of chained network

征处理方式，网络迭代 k 次，在第 1 阶段（$i=1$），Φ 和 Ψ 的输入仅为 $\boldsymbol{x}$。在此后各阶段，将前一阶段输出的 $\boldsymbol{x}_{i-1}^{\psi}$ 和 $\boldsymbol{x}_{i-1}^{\varphi}$ 与视觉信息 x 融合，交叉输入至 Φ 和 Ψ 中进一步处理。通过这种多阶段双路交叉的方式对特征连续提取和增强，其具备的三个结构要素，可显著增强特征提取能力：

（1）双路：即两种特征处理方式（Φ 和 Ψ）。站在不同角度对视觉特征进行分析，一定程度上消除单一角度分析内容的不足；

（2）交叉：将两种处理方式的输出结果，分别采用另一种处理方式进一步解析（$\boldsymbol{x}_i^{\varphi}$ 由 Ψ 处理，$\boldsymbol{x}_i^{\psi}$ 由 Φ 处理）；

（3）多阶段：对上述双路交叉操作进行多阶段迭代，进一步加强特征的表达效果。

本文利用上述结构捕获视频语义特征，提出链式语义生成网络—ChainS-Net，提升视频描述性能。如图 2 所示，为实现从不同角度对视频内容进行分析，本文对视频卷积获取的视觉特征均匀切分为多份。全局视觉特征表达完整视频内容，切分后的局部视觉特征是视觉特征片段的集合，表示视频局部细节，由此设计了处理两种视觉特征的全局域模块 G 和局部域模块 L，从全局和局部两方面综合分析视频语义。下文将详述全局域和局部域模块，以及以两模块为基础构建的 ChainS-Net，最后介绍整体视频描述模型。

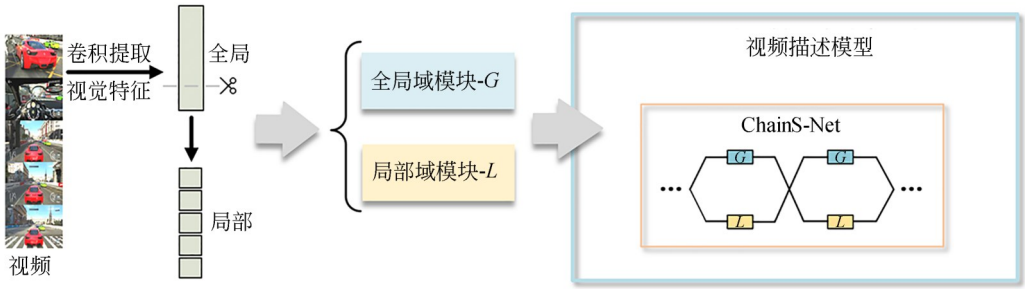

图 2　视觉特征切分及整体网络结构示意图

Fig. 2　Diagram of visual feature segmentation and overall network structure

2.1　全局域和局部域模块

2.1.1　特征提取基本结构—MLP

链式语义生成网络以全局域和局部域两个基本模块构成，由于 MLP 在分类任务中有较好表现，本文在两个模块中采用 MLP 获取相应语义特征，如图 3 所示为 MLP 网络结构，相应数学表达如式（3）～式（4）：

$$\boldsymbol{y}=F(\boldsymbol{x})=f_m(\cdots f_2(f_1(\boldsymbol{x}))), \tag{3}$$

$$f_m(\boldsymbol{u})=\sigma_m(\boldsymbol{w}_m\boldsymbol{u}+\boldsymbol{b}_m), \tag{4}$$

其中：F 为 MLP 整体函数表示，$\boldsymbol{x}$ 和 $\boldsymbol{y}$ 分别为输入和输出，f 是 MLP 的每一层处理，其具体表达如式（4）所示，以第 m 层为例，$\boldsymbol{u}$ 为该层输入特征，$\boldsymbol{w}_m$ 和 $\boldsymbol{b}_m$ 是第 m 层权重和偏置，且 $\boldsymbol{w}_m$ 和 $\boldsymbol{u}$ 进行全连接计算，σ_m 表示非线性激活函数。

2.1.2　全局域模块

为了从全局角度提取视频语义，设计了以 MLP 为基本单元的全局域模块，其网络结构如图 4，相应数学表达如式（5）所示。其中 $\boldsymbol{v}$ 表示视频经 2D 和 3D 卷积处理得到的视觉特征，$\boldsymbol{s}^g$ 为模

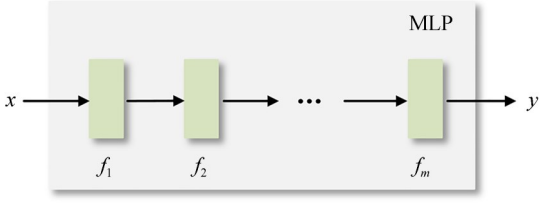

图 3　MLP 网络结构

Fig. 3　Network structure of MLP

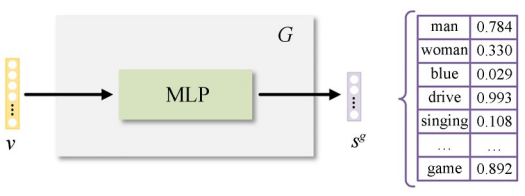

图 4　全局域模块

Fig. 4　Global domain module

块输出的全局语义特征，v 和 s^g 分别是维度 p 和 q 的特征向量。用函数 G 表示全局域模块处理，且模块简单地由 MLP 进行特征提取。该模块对视觉特征（全局视觉特征）进行分析，得到概括性的全局语义特征。

$$s^g = G(v) = F(v). \tag{5}$$

2.1.3　局部域模块

为了对视觉特征的局部深入分析进而捕获相应语义表示，设计了 ChainS-Net 中局部域模块，其网络结构如图 5 所示。

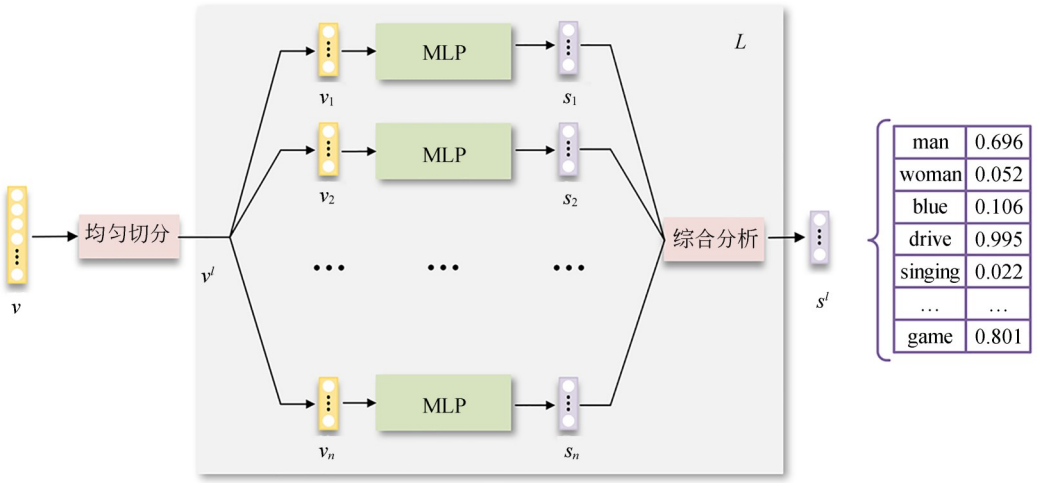

图 5　局部域模块

Fig. 5　Local domain module

首先采用均匀切分方式对视觉特征进行分割，获取可以表达视频局部细节的局部视觉特征，如式（6），D 表示均匀切分函数，具体操作为采用深度学习框架中 split 函数对向量进行切分（如 tensorflow. split），v 是视频卷积提取的原始视觉特征，维度为 p，v^l 是切分得到的局部视觉特征，包括 $v_1, v_2, \cdots, v_n$，维度均为 p/n。

$$v^l = D(v), v^l = \{v_1, v_2, \cdots, v_n\}. \tag{6}$$

将获取的局部视觉特征通过局部域模块处理，该模块的整体数学表达如式（7）所示，v^l 和 s^l 是局部视觉和局部语义特征。

$$s^l = L(v^l). \tag{7}$$

由该模块获取局部视频语义可分为如下提取和融合两个步骤。

局部语义特征提取：利用 MLP 结构分别对 n 个局部视觉特征 $v_1, v_2, \cdots, v_n$ 提取相应语义，如式（8）所示，F 是 MLP 处理函数，具体操作参考式（3）～式（4），由此得到 n 个局部语义特征 $s_1, s_2, \cdots, s_n$，其向量维度均为 q。

$$s_i = F(v_i), v_i \in v^l. \tag{8}$$

局部语义特征融合：如式（9）所示，将获取的局部语义特征融合，即求取上述特征的平均表达，得到局部语义特征的综合表示 s^l，向量维度仍为 q：

$$s^l = \frac{\sum_{i=1}^{n} s_i}{n}. \tag{9}$$

该模块中首先通过均匀切分获得局部视觉特征，采用 MLP 提取相应局部语义特征后将其融合，综合考虑这些局部分析结果提升局部语义特征的有效性。

2.1.4　全局域-局部域结合

全局域和局部域模块分别对完整和切分后的视觉特征进行处理，生成表达概括性和细节性的语义信息，具备较强的互补性，由此将二者结合，输出全面的视频语义表示，如图 6 将两个模块结合，其数学表达如式（10）～式（11）：

$$\begin{cases} s_1^g = G(v) \\ s_1^l = L(v^l) \end{cases}, \tag{10}$$

$$s = s_1^g \oplus s_1^l. \tag{11}$$

由于上述特征提取仅为本文提出链式网络的第 1 阶段，故将两模块输出的 s^g 和 s^l 记为 s_1^g 和 s_1^l，其中 $\oplus$ 表示将两种特征拼接（concat），本文采用拼接的融合方式可以最大程度保留两种特征的属性信息。

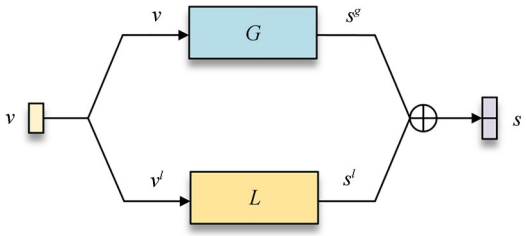

图 6　全局域-局部域结合

Fig. 6　Combination of global and local domains

2.2　链式语义生成网络

上述模块可以从全局和局部角度综合分析视频语义，为进一步对特征有效利用和增强，结合域变换的思想，本文提出链式语义生成网络，其网络结构如图 7，以第 1 阶段和第 2 阶段为例，给出了语义提取的具体操作。

2.2.1　双路语义预提取（第 1 阶段）

上文详述了全局域 G 和局部域 L 的特征提取过程，利用双路这一结构要素实现对视觉特征整体和细节的语义感知。如图 7 所示，网络采用多阶段（k 阶段）迭代推理，进一步求取较准确的语义信息。如式（10），在第 1 阶段，分别将两种视觉特征 $\boldsymbol{v}$ 和 $\boldsymbol{v}^l$ 经 G 和 L 模块进行语义预提取，生成第 1 阶段语义特征 $\boldsymbol{s}_1^g$ 和 $\boldsymbol{s}_1^l$。

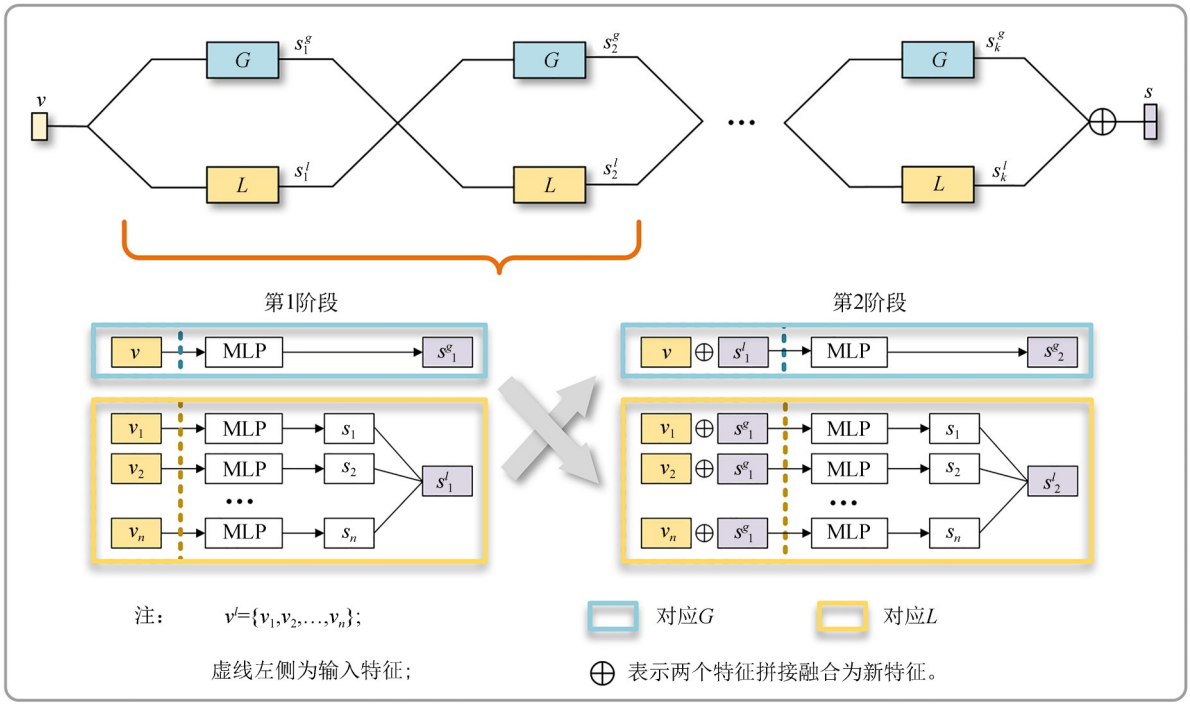

图 7　链式语义生成网络

Fig. 7　Chained semantic generation network

2.2.2　交叉提取（第 2 阶段）

在后续阶段中，链式网络通过交叉提取得到更有效的语义表示。以第 2 阶段为例，首先对输入特征进行重构，数学表达如式（12）：

$$\begin{cases}\tilde{\boldsymbol{v}}=\boldsymbol{v}\oplus\boldsymbol{s}_1^l\\ \tilde{\boldsymbol{v}}^l=\boldsymbol{v}^l\oplus\boldsymbol{s}_1^g\end{cases},\tag{12}$$

其中：$\oplus$ 表示特征拼接，将全局视觉 $\boldsymbol{v}$ 和局部语义 $\boldsymbol{s}_1^l$ 拼接，局部视觉 $\boldsymbol{v}^l$ 和全局语义 $\boldsymbol{s}_1^g$ 拼接，得到广泛意义上的全局和局部视觉特征 $\tilde{\boldsymbol{v}}$ 和 $\tilde{\boldsymbol{v}}^l$。

由图 7 可知，与第 1 阶段不同，第 2 阶段并非仅以视觉特征为输入，为了更充分地利用视觉特征和第 1 阶段预提取的语义特征，本文将视觉和语义特征结合，重构第 2 阶段的输入。由于第 1 阶段两个模块生成的语义特征 $(\boldsymbol{s}_1^g,\boldsymbol{s}_1^l)$ 有一定差异性，若仍以相对应的视觉和语义信息直接组合作为输入 $(\boldsymbol{v}\oplus\boldsymbol{s}_1^g,\boldsymbol{v}^l\oplus\boldsymbol{s}_1^l)$，会导致输出的语义信息 $(\boldsymbol{s}_2^g,\boldsymbol{s}_2^l)$ 差异性进一步增大且偏离最佳表示。由此将两种视觉和语义特征交叉结合得到 $\tilde{\boldsymbol{v}}$ 和 $\tilde{\boldsymbol{v}}^l$，

使得二者存在适当差异性的基础上,具备一定的相关性。该操作也使得不同属性的视觉和语义信息可以相互参考(v 和 s_1^l,v^l 和 s_1^g),通过这种不同特征间的交互作用,增强特征的鲁棒性和有效性。

如式(13),将重构后的特征 $\tilde{v}$ 和 $\tilde{v}^l$ 采用相应模块进行更精确的语义提取:

$$\begin{cases} s_2^g = G(\tilde{v}) \\ s_2^l = L(\tilde{v}^l) \end{cases}, \tag{13}$$

其中,s_2^g 和 s_2^l 为第 2 阶段全局和局部语义特征。

以重构后的丰富的特征表示作为输入,经 G 和 L 捕获第 2 阶段语义信息。该操作将两种语义特征转换到相对的另一个域视角下进一步解析(s_1^g 由 L 处理,s_1^l 由 G 处理),有利于网络对视频语义正确理解。

2.2.3　多阶段提取

网络进一步采用多阶段策略获取表达能力更强的语义表示,对第 2 阶段所示的交叉提取过程进行迭代操作,在网络的最后阶段(第 k 阶段)输出相应语义特征,数学表达如式(14)所示,其中 $\tilde{v}$ 和 $\tilde{v}^l$ 均包含两种特征属性($\tilde{v}$ 包含 v 和 s_{k-1}^l,$\tilde{v}^l$ 包含 v^l 和 s_{k-1}^g),s_k^g 和 s_k^l 为第 k 阶段语义特征:

$$\begin{cases} s_k^g = G(\tilde{v}) \\ s_k^l = L(\tilde{v}^l) \end{cases}. \tag{14}$$

最后将两种特征拼接结合,得到语义特征 s,可描述为:

$$s = s_k^g \oplus s_k^l,\ k, \in Z, k \geqslant 1. \tag{15}$$

本文提出的链式语义生成网络,采用全局域和局部域模块获取多角度语义表示,利用交叉提取操作进一步提升语义特征的鲁棒性和有效性,最后通过多阶段迭代来增强语义特征表达能力。

2.3　视频描述网络模型

本文将链式语义生成网络 ChainS-Net 用于视频描述生成,构建视频描述网络模型,利用 ChainS-Net 获取表达能力更强的语义特征,进而生成准确的文本描述结果,视频描述网络模型如图 8 所示。

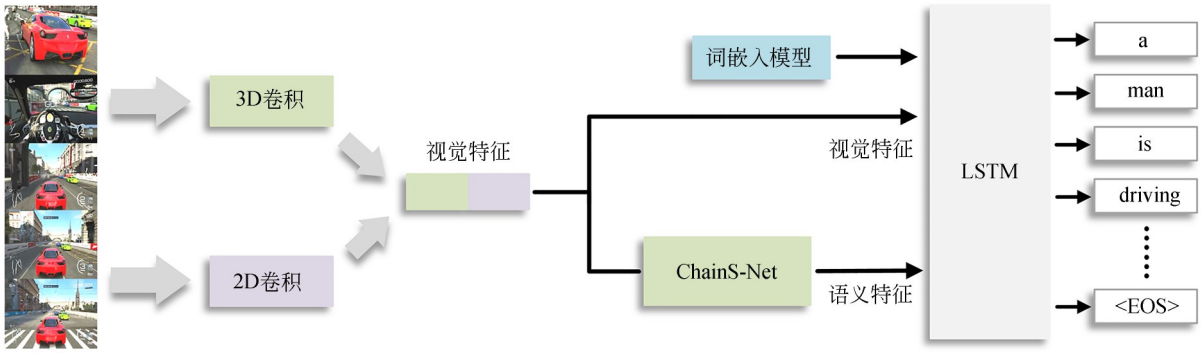

图 8　视频描述网络模型

Fig. 8　Video captioning network

算法流程如下:

第 1 步:对视频进行预处理。从视频中提取固定数量的帧图像,并将图像剪裁为固定尺寸;

第 2 步:提取视觉特征。分别采用 3D 和 2D 卷积对视频提取特征,3D 和 2D 特征分别倾向于表达视频的动态和静态信息[15],将两种特征向量拼接得到视觉特征,更全面地表达视频内容;

第 3 步:提取语义特征。以视觉特征为输入,采用本文提出的 ChainS-Net 提取视频语义特征;

第 4 步:生成文本描述。将视觉特征和语义特征作为 LSTM 网络的输入,输出文本描述结果。

3　仿真分析

3.1　数据集

本文选择在视频描述通用数据集 MSR-VTT[16] 和 MSVD[17] 上进行实验和分析,MSR-VTT 数据集中有 10 000 个短视频,且视频时长约为 10 s,其内容包括生活中诸多情景,每个视频配有 20 个描述句子(Ground Truth, GT)。遵循

数据集官方设置及前人经验，分别将 6 513，497 和 2 990 个视频作为训练集、验证集和测试集。MSVD 数据集与 MSR-VTT 类似，共 1 970 个视频，本文分别将 1 200，100 和 670 个视频用于网络训练、验证和测试。

3.2 实验设置

本文算法实现的系统环境为 Ubuntu16.04，采用 Python 语言及 TensorFlow 深度学习框架编程，用于对网络训练和测试的显卡为单张 NVIDIA 1080Ti。实验中单独训练和测试 ChainS-Net，将输出的语义特征用于 LSTM 输入，并重新对 LSTM 网络进行训练和测试后输出描述结果，网络训练细节如下。

3.2.1 视频描述网络

在整体视频描述网络中，对视频预处理后采用预训练的 ECO[18] 和 ResNeXt[19] 网络提取 3D 和 2D 特征，向量维度分别为 1 536 和 2 048，将两种特征拼接为 3 584 维视觉特征 $\boldsymbol{v}$。采用 ChainS-Net 生成语义特征 $\boldsymbol{s}$，将 $\boldsymbol{v}$ 和 $\boldsymbol{s}$ 输入至 LSTM 生成描述结果。(LSTM) 训练时学习率设置为 0.000 4，批尺寸设置为 64，迭代次数为 50，采用 Adam 算法对网络进行优化。

本文采用视频描述标准的评价指标对网络进行评估，分别为 BLEU-4(B-4)，CIDEr(C)，METEOR(M) 和 ROUGE-L(R)，可以从准确率、召回率、流畅性和同义词等诸多因素判断句子质量，4 个评价指标的值越大说明描述得越准确。

3.2.2 ChainS-Net

ChainS-Net 以维度为 3 584 的视觉特征 $\boldsymbol{v}$ 为输入，在第 1 阶段，将 $\boldsymbol{v}$ 经全局域模块输出 300 维全局语义特征 $\boldsymbol{s}_1^g$，将 $\boldsymbol{v}$ 均匀切分为 7 个 512 维的局部视觉特征 $\boldsymbol{v}^l(\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_7)$，其中考虑到整除，并防止在某一个局部发生 3D 和 2D 特征混杂，切分数量设为 7，经局部域模块输出 300 维局部视觉特征 s_1^l。在后续阶段中，将 $\boldsymbol{s}_{i-1}^l$ 和 $\boldsymbol{v}$ 拼接(3 884 维)输入到全局域模块 G 生成第 i 阶段全局语义特征 $\boldsymbol{s}_i^g$，$\boldsymbol{s}_{i-1}^g$ 和 $\boldsymbol{v}^l$ 拼接(812 维)输入到局部域模块 L 生成第 i 阶段局部语义特征 $\boldsymbol{s}_i^l$，在最后阶段将两种特征拼接为 600 维的语义特征 $\boldsymbol{s}$。

网络中用于提取语义特征的基本单元均是层数为 3 的 MLP，每一层神经元个数分别为 512，392 和 300。语义特征的 GT 分别是两个数据集中出现频次最高的实词，第 i 个视频的语义 GT 表示为 $\hat{s}_i=\{\hat{s}_{i,0}, \hat{s}_{i,1}, \cdots, \hat{s}_{i,299}\}$，$\hat{s}_{i,j}\in\{0,1\}$，特征中某元素为 1 表示相应单词在视频中有所涉及。训练时学习率设置为 0.000 2，批尺寸设置为 128，迭代次数为 1 000。

3.3 消融实验

为验证 ChainS-Net 结构的有效性，本文进行如下消融实验，采用以下结构生成视频语义特征，用于生成视频描述，性能对比如表 1 所示。

3.3.1 全局域-局部域双路结构有效性分析

如图 9 所示，本文分别或同时采用 G 和 L 模块获取视频语义，G 和 L 分别表示全局域和局部域模块，其输出维度均为 300，由此(c)结构输出 $\boldsymbol{s}$ 的维度为 600(该结构为简单的 $k=1$)，性能对比如表 1 数据部分第 1～3 行。由实验数据可知，在 MSR-VTT 数据集上同时采用两个模块可以达到较为优异的视频描述性能，在 MSVD 数据集四个评价指标上也均优于两个模块单独使用。两个模块从不同角度对视觉特征进行分析，具备较强互补性，将二者结合可以更全面地表达视频内容，验证了该方法的有效性。

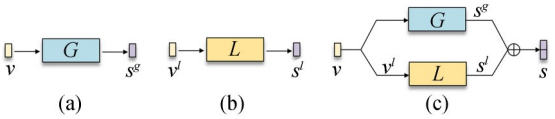

图 9 两模块消融实验结构图

Fig. 9 Structures of ablation experiment for two modules

3.3.2 交叉提取方式的有效性分析

本文进行如下实验探究链式网络中交叉提取方式的作用，网络结构对比如图 10 所示，(a)为本文提出方法，(b)结构中(非交叉方式)在每一阶段将视觉特征 $\boldsymbol{v}$ 和全局语义 $\boldsymbol{s}^g$ 拼接，局部视觉特征 $\boldsymbol{v}^l$ 和局部语义 $\boldsymbol{s}^l$ 拼接，分别输入到下一阶段全局域 G 和局部域 L 模块中。两种结构性能对比如表 1 第 4～5 行，且实验数据均为 $k=2$ 情况下(2 阶段链式网络)。在两个数据集上，当阶段数相同时本文方法明显更优，根据上文分析，

ChainS-Net在每阶段将语义特征输入到另一个域中重新分析,同时实现了两种视觉和语义特征的交互参考,可以有效增强语义特征的表达能力。

表1　消融实验性能对比

Tab. 1　Performance comparison of ablation experiments

	MSR-VTT				MSVD			
Mode	B−4	C	M	R	B-4	C	M	R
G	45.83	**53.16**	29.28	63.64	62.35	109.71	39.04	77.04
L	45.21	52.63	29.38	63.59	61.07	113.41	39.71	77.34
G−*L*	**46.70**	52.93	**29.55**	**64.03**	**65.20**	**119.58**	**41.51**	**79.02**
交叉	**47.44**	**52.98**	**29.75**	**64.11**	**66.29**	119.93	**42.25**	**79.30**
非交叉	46.72	52.67	29.60	63.99	65.47	**120.09**	41.49	79.24
k=1	46.70	52.93	29.55	64.03	65.20	119.58	41.51	79.02
k=2	47.44	52.98	29.75	64.11	66.29	119.93	42.25	79.30
k=3	47.70	53.64	29.78	64.31	65.62	**122.44**	**42.25**	79.93
k=4	48.20	53.75	**29.98**	**64.48**	**66.51**	122.26	41.98	**80.05**
k=5	**48.35**	**54.22**	29.90	64.47	64.45	121.09	42.13	79.85

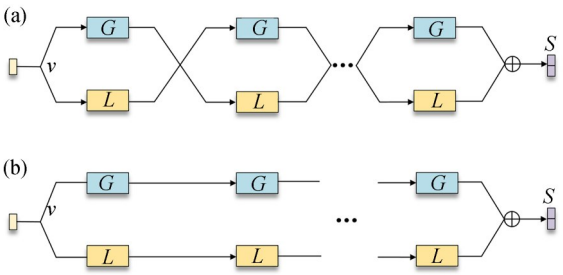

图10　交叉提取方式消融实验结构图

Fig. 10　Structures of ablation experiment for cross-extraction

3.3.3　多阶段结构有效性分析

为进一步验证ChainS-Net中多阶段结构的有效性,同时确定最佳阶段数*k*,进行了*k*逐渐递增的实验(*k*=1,2,…),性能对比如表1第6~10行。由表可知,若均衡考虑4个评价指标,在MSR-VTT数据集上,视频描述性能随着*k*增大逐渐增高,*k*=4时达到最佳,*k*=5时趋于稳定。MSVD数据集上类似,且均衡的最佳性能出现在*k*=3。实验验证了上述双路交叉方式可以进行多次迭代,将性能最优的语义特征应用于视频描述,生成更准确的文本描述句子。

3.4　结果分析

本文在两个数据集上进行仿真实验,将评价指标与现有模型进行对比,如表2和表3所示,在MSR-VTT数据集上,相比于仅采用单阶段全局

表2　MSR-VTT数据集模型性能对比

Tab. 2　Performance comparison on MSR-VTT dataset

Method	B-4	C	M	R
MTVC[20]	40.8	47.1	28.8	60.2
CIDEnt-RL[21]	40.5	51.7	28.4	61.4
SibNet[22]	40.9	47.5	27.5	60.2
HACA[23]	43.4	49.7	29.5	61.8
TAMoE[24]	42.2	48.9	29.4	62.0
POS[25]	41.3	53.4	28.7	62.1
MARN[10]	40.4	47.1	28.1	60.7
JSRL-VCT[26]	42.3	49.1	29.7	62.8
GRU-EVE[27]	38.3	48.1	28.4	60.7
STG-KD[28]	40.5	47.1	28.3	60.9
SAAT[29]	39.9	51.0	27.7	61.2
ORG-TRL[9]	43.6	50.9	28.8	62.1
SAVC[8]	45.8	53.2	29.3	63.6
ChainS-Net	**48.2**	**53.8**	**30.0**	**64.5**

进行语义提取的SAVC网络，ChainS-Net($k=4$)在4个评价指标上均有不同程度提升，且优于现有同类方法。MSVD数据集上，ChainS-Net($k=3$)的4个评价指标也取得较好效果，优于其他15个视频描述算法。

将ChainS-Net与SAVC网络进行可视化结果对比，在两个数据集上选择6个视频样本，将ChainS-Net，SAVC和GT的文本描述对比分析，如图11所示。相比于仅单阶段全局处理的SAVC网络，ChainS-Net可以生成更准确、全面和具体的文本描述，如(a)视频中准确描述出SAVC忽略的“男人”，(d)更具体地描述了开车场景—“在电子游戏中”，而(e)视频则更全面地表达了“一群士兵在战场上战斗”。以上例子均表明链式网络对语义特征及视频描述模型的增强效果，验证了本文方法的优越性。

为衡量算法在应用层面的实时性，如表4所示，对算法运行时间进行对比，表中数据为测试阶段在两个数据集上单个视频生成文本描述所用的平均时间，可知ChainS-Net所用时间稍高于SAVC，但由此带来了较大的性能提升，且基本满足实时性要求。

表3 MSVD数据集模型性能对比

Tab. 3 Performance comparison on MSVD dataset

Method	B-4	C	M	R
LSTM-E[30]	45.3		31.0	
h-RNN[31]	49.9	65.8	32.6	
aLSTMs[32]	50.8	74.8	33.3	
SCN[14]	51.1	77.7	33.5	
MTVC[20]	54.5	92.4	36.0	72.8
ECO[18]	53.5	85.8	35.0	
SibNet[22]	54.2	88.2	34.8	71.7
POS[25]	53.9	91.0	34.9	72.1
MARN[10]	48.6	92.2	35.1	71.9
JSRL-VCT[26]	52.8	87.8	36.1	71.8
GRU-EVE[27]	47.9	78.1	35.0	71.5
STG-KD[28]	52.2	93.0	36.9	73.9
SAAT[29]	46.5	81.0	33.5	69.4
ORG-TRL[9]	54.3	95.2	36.4	73.9
SAVC[8]	62.4	109.7	39.0	77.0
ChainS-Net	**65.6**	**122.4**	**42.3**	**79.9**

图11 视频描述可视化结果对比

Fig. 11 Comparison of video captioning results

表 4　运行时间对比

Tab. 4　Runtime comparison

算法	运行时间/ms
SAVC	517
ChainS-Net	774

4　结　论

为了捕获表达能力更强的视频语义特征,进而生成准确的文本描述,本文提出视频描述中链式语义生成网络。设计了以MLP为基本单元的全局域和局部域模块,分别对视觉特征的整体和局部片段进行处理,从全局和局部两方面捕获视频语义,为特征的多角度分析提供一种便捷有效的新思路。以此为基础构建了链式网络结构,其交叉提取操作,使得特征可以在两个域之间变换解析,同时实现多种属性信息的交互融合,增强了神经网络对视频语义的解析能力,网络的多阶段迭代方式可以实现语义特征的进一步增强,提高视频描述准确性。在后续工作中,将继续优化网络各阶段的特征融合机制,进一步提高链式网络性能。

参考文献:

[1] 汤鹏杰, 王瀚漓. 从视频到语言:视频标题生成与描述研究综述[J]. 自动化学报, 2022, 48(2): 375-397.
TANG P J, WANG H L. From video to language: survey of video captioning and description[J]. *Acta Automatica Sinica*, 2022, 48(2): 375-397. (in Chinese)

[2] GU J X, WANG Z H, KUEN J, *et al*. Recent advances in convolutional neural networks[J]. *Pattern Recognition*, 2018, 77: 354-377.

[3] ELMAN J L. Finding structure in time[J]. *Cognitive Science*, 1990, 14(2): 179-211.

[4] VENUGOPALAN S, ROHRBACH M, DONAHUE J, *et al*. Sequence to sequence: video to text [C]. 2015 *IEEE International Conference on Computer Vision. Santiago, Chile*. IEEE, 2015: 4534-4542.

[5] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory [J]. *Neural Computation*, 1997, 9(8): 1735-1780.

[6] 陈科峻,张叶. 循环神经网络多标签航空图像分类[J]. 光学 精密工程,2020,28(6):1404-1413.
CHEN K J, ZHANG Y. Recurrent neural network multi-label aerial images classification[J]. *Opt. Precision Eng.*, 2020, 28(6): 1404-1413. (in Chinese)

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

[8] YAN C G, TU Y B, WANG X Z, *et al*. STAT: spatial-temporal attention mechanism for video captioning [J]. *IEEE Transactions on Multimedia*, 2020, 22(1): 229-241.

[9] CHEN H R, LIN K, MAYE A, *et al*. A semantics-assisted video captioning model trained with scheduled sampling[J]. *Frontiers in Robotics and AI*, 2020, 7: 475767.

[10] ZHANG Z Q, SHI Y Y, YUAN C F, *et al*. Object relational graph with teacher-recommended learning for video captioning [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA*. IEEE, 2020: 13275-13285.

[11] PEI W J, ZHANG J Y, WANG X R, *et al*. Memory-attended recurrent network for video captioning[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA*. IEEE, 2019: 8339-8348.

[12] 赵海英, 周伟, 侯小刚, 等. 多标签分类的传统民族服饰纹样图像语义理解[J]. 光学 精密工程, 2020, 28(3): 695-703.
ZHAO H Y, ZHOU W, HOU X G, *et al*. Multi-label classification of traditional national costume pattern image semantic understanding [J]. *Opt. Precision Eng.*, 2020, 28(3): 695-703. (in Chinese)

[13] SHETTY R, LAAKSONEN J T. Video captioning with recurrent networks based on frame- and video-level features and visual content classification [J]. *arXiv preprint, arXiv:1512.02949.*, 2015

[14] TU Y B, ZHANG X S, LIU B T, *et al*. Video description with spatial-temporal attention [C]. *MM'17: Proceedings of the 25th ACM international conference on Multimedia*. 2017: 1014-1022.

[15] GAN Z, GAN C, HE X D, *et al*. Semantic compositional networks for visual captioning[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA*. IEEE, 2017 : 1141-1150.

[16] 杨其利, 周炳红, 郑伟, 等. 注意力卷积长短时记忆网络的弱小目标轨迹检测[J]. 光学 精密工程, 2020, 28(11): 2535-2548.
YANG Q L, ZHOU B H, ZHENG W, *et al*. Trajectory detection of small targets based on convolutional long short-term memory with attention mechanisms[J]. *Opt. Precision Eng.*, 2020, 28(11): 2535-2548. (in Chinese)

[17] XU J, MEI T, YAO T, *et al*. MSR-VTT: a large video description dataset for bridging video and language[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA*. IEEE, 2016 : 5288-5296.

[18] GUADARRAMA S, KRISHNAMOORTHY N, MALKARNENKAR G, *et al*. YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition[C]. 2013 *IEEE International Conference on Computer Vision. Sydney, NSW, Australia*. IEEE, 2013: 2712-2719.

[19] ZOLFAGHARI M, SINGH K, BROX T. ECO: efficient convolutional network for online video understanding [C]. *Computer Vision-ECCV* 2018, 2018: 713-730.

[20] XIE S N, GIRSHICK R, DOLLÁR P, *et al*. Aggregated residual transformations for deep neural networks[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA*. IEEE, 2017: 5987-5995.

[21] PASUNURU R, BANSAL M. Multi-task video captioning with video and entailment generation [C]. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*). *Vancouver, Canada. Stroudsburg, PA, USA*: *Association for Computational Linguistics*, 2017: 1273-1283.

[22] PASUNURU R, BANSAL M. Reinforced video captioning with entailment rewards[C]. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark. Stroudsburg, PA, USA*: *Association for Computational Linguistics*, 2017: 979 - 985.

[23] LIU S, REN Z, YUAN J S. SibNet: sibling convolutional encoder for video captioning[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(9): 3259-3272.

[24] WANG X, WANG Y F, WANG W Y. Watch, listen, and describe: globally and locally aligned cross-modal attentions for video captioning [C]. *Proceedings of the* 2018 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *Volume* 2 (*Short Papers*). *New Orleans, Louisiana. Stroudsburg, PA, USA*: *Association for Computational Linguistics*, 2018: 795 - 801.

[25] WANG X, WU J W, ZHANG D, *et al*. Learning to compose topic-aware mixture of experts for zero-shot video captioning [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 8965-8972.

[26] WANG B R, MA L, ZHANG W, *et al*. Controllable video captioning with POS sequence guidance based on gated fusion network[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *Seoul, Korea* (*South*). IEEE, 2019: 2641-2650.

[27] HOU J Y, WU X X, ZHAO W T, *et al*. Joint syntax representation learning and visual cue translation for video captioning[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *Seoul, Korea* (*South*). IEEE, 2019 : 8917-8926.

[28] AAFAQ N, AKHTAR N, LIU W, *et al*. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Long Beach, CA, USA*. IEEE, 2019: 12479-12488.

[29] PAN B X, CAI H Y, HUANG D A, *et al*. Spatio-temporal graph for video captioning with knowledge distillation[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Seattle, WA, USA*. IEEE, 2020: 10867-10876.

[30] ZHENG Q, WANG C Y, TAO D C. Syntax-aware action targeting for video captioning [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Seattle, WA, USA*. IEEE, 2020: 13093-13102.

[31] PAN Y W, MEI T, YAO T, *et al*. Jointly modeling embedding and translation to bridge video and language[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas*, *NV*, *USA*. IEEE, 2016: 4594-4602.

[32] YU H N, WANG J, HUANG Z H, *et al*. Video paragraph captioning using hierarchical recurrent neural networks[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas*, *NV*, *USA*. IEEE, 2016: 4584-4593.

[33] GAO L L, GUO Z, ZHANG H W, *et al*. Video captioning with attention-based LSTM and semantic consistency[J]. *IEEE Transactions on Multimedia*, 2017, 19(9): 2045-2055.

作者简介:

毛　琳(1977—),女,山东省荣成人,副教授,2000年于哈尔滨工程大学获得学士学位,2005年于黑龙江大学获得硕士学位,2011年于哈尔滨工程大学获得博士学位,主要从事目标跟踪与多传感器信息融合方面的研究。E-mail:maolin@dlnu.edu.cn

通讯作者:

高　航(1997—),男,内蒙古赤峰人,硕士研究生,2020年于大连民族大学获得学士学位,主要从事计算机视觉视频描述方面的研究。E-mail:gao_hang2021@163.com