基于多维度和多模态信息的视频描述方法

丁恩杰[1]，刘忠育[1]，刘亚峰[1]，郁万里[2]
（1. 中国矿业大学物联网（感知矿山）研究中心，江苏 徐州 221008；2. 不来德大学电动学与微电子研究所，不来德 28359）

摘 要：针对视频自动描述任务中的复杂信息表征问题，提出一种多维度和多模态视觉特征的提取和融合方法。首先通过迁移学习提取视频序列的静态和动态等多维度特征，并采用图像描述算法提取视频关键帧的语义信息，完成视频信息的特征表征；然后采用多层长短期记忆网络融合多维度和多模态信息，最终生成视频内容的语言描述。实验仿真表明，所提方法与目前已有方法相比，在视频自动描述任务中取得了较好的效果。

关键词：视频描述；多模态；迁移学习；长短期记忆网络；循环神经网络
中图分类号：TP391.4
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2020037

Video description method based on multidimensional and multimodal information

DING Enjie[1], LIU Zhongyu[1], LIU Yafeng[1], YU Wanli[2]
1. IoT/Perception Mine Research Center, China University of Mining & Technology, Xuzhou 221008, China
2. Institute of Electrodynamics and Microelectronics, University of Bremen, Bremen 28359, Germany

Abstract: In order to solve the problem of complex information representation in automatic video description tasks, a multi-dimensional and multi-modal visual feature extraction and fusion method was proposed. Firstly, multi-dimensional features such as static and dynamic attributes of the video sequence were extracted by transfer learning, and the image description algorithm was also used to extract the semantic information of the key frames in the video. By doing this, the video features extraction was carried out. Then, multi-layer long and short memory networks were used to fuse multi-dimensional and multi-modal information, and finally generated a language description of the video content. Compared with the existing methods, experimental simulations results show that the proposed method achieves better results in the video automatic description task.

Key words: video description, multimodal, transfer learning, long and short term memory network, recurrent neural network

1 引言

随着大数据、计算机算力、机器学习模型不断发展，视频描述技术再度掀起研究热潮。视频描述有着十分广泛的应用，如视频检索、视频标注、行为识别、人机交互、视频内容讲解等场景[1-2]。然而该任务相对复杂，涉及计算机视觉理解和自然语言处理 2 个领域，本质上属于跨模态的映射问题，现有的方法还有较大的提升空间[3]。

视频描述主要分 2 类。一类是抽象概括一段视频的主要内容，该类任务的输入通常是一个视频片段，而输出则是一句或若干句自然语言[4]。另一类则是视频内容的密集描述，通常需要将视频片段中的人、物、场景状态及其相互关系和变化过程描述

收稿日期：2019–10–21；**修回日期**：2020–01–14
通信作者：刘忠育，zhongyuliu6@163.com
基金项目：国家重点研发计划基金资助项目（No.2017YFC0804400, No.2017YFC0804401）
Foundation Item: The National Key Research and Development Program of China (No.2017YFC0804400, No.2017YFC0804401)

清楚[5-6]。本文所提方法主要解决第一类问题。

传统的视频描述方法是基于模板的方法[7-10]，如主语–动词–宾语三元组（SVO, subject-verb-object）[9]和主语–动词–宾语–地点（SOVP, subject-verb- object-place）[10]等。这类方法通常预先设定产生句子的词法和语法规则，并且预先定义主语、谓语和宾语等要素的视觉类别，当检测到相应的视觉目标时，将视觉语义映射到模板中。显然，该方法总能够根据视觉要素在预定义的模板中直接生成语法正确的描述，不足之处在于该类方法高度依赖预定义的语言模板，生成语句受到预定义的视觉类别和语法结构的限制，句子描述的形式和内容缺乏灵活性和多样性。另一类方法是基于深度学习的方法。鉴于循环神经网络（RNN, recurrent neural network）在自然语言翻译中的惊人表现[11]，相关学者逐渐开始使用此类方法生成视频的语言描述。文献[12]首先用卷积神经网络（CNN, convolutional neural network）提取视频中的图像特征，然后用 RNN 类的方法对图像特征进行编码，最后解码生成视频内容的自然语言描述。然而，该方法提取的视觉特征较单一，对视频内容的语言描述不够丰富[13]。文献[14]提出递归编码器并结合注意力模型，使用在 Imagenet 上预训练的深度卷积神经网络提取视频关键帧的视觉特征，然后按照时序输入长短期记忆网络（LSTM, long and short term memory network）[15]进行编码。文献[16]提出基于多模态融合的视频描述方法，提取视频中动态特征和静态视觉特征，并融合音频特征产生语言描述。然而对于视频中单帧图像而言，没有充分考虑场景中的背景和语义信息。文献[17]提出一种提取视频关键帧的方法来提升描述语言的准确度，然而该方法同样没有考虑视频的物体、背景和时空等多维度信息。此类方法使用预训练的 CNN 提取目标视频的特征，本质上是采用迁移学习对目标视频的视觉特征进行提取。然而，迁移学习要求选取的源域和目标域特征分布越接近越好[18]，但由于描述视频内容的多样性和随机性，目标域的特征很难和某个图像数据集特征分布完全相同，因此该问题是此类方法的主要瓶颈之一。

为解决上述瓶颈问题，本文采用迁移学习，从视频包含的物体、背景和时空动态关系等多个维度提取视频中静态、动态和语义信息，并采用 LSTM 处理多维度和多模态信息完成编解码，最终生成视频的语言描述。根据以上分析，本文的主要贡献如下。

1) 提出采用多个源域预训练的模型提取视频中的静态、动态和语义信息，以提升视频语言描述的准确度。

2) 提出一种多模态信息融合方法，将视频的关键帧进行语言描述，并将视觉模态信息和语言模态信息融合，从而进一步提升模型生成语言的多样性。

3) 在微软视频描述（MSVD, microsoft video description）[19]和微软视频到文本研究（MSR-VTT, microsoft research-video to text）[20]公共数据集上进行模型验证，并采用多种评价指标对产生的语言进行评价，结果显示所提方法取得了良好的效果。

2　改进的视频描述方法

本文提出的视频描述方法分为编码和解码 2 个阶段，如图 1 所示。在编码阶段，首先对输入视频预处理，获取包含静态特征的图像和动态特征的视频片段 2 种模态。然后分别用二维卷积神经网络（2DCNN, two dimensional convolutional neural network）和三维卷积神经网络（3DCNN, three dimensional convolutional neural network）对图像模态和视频模态进行特征提取。近几年，人们对单幅图像的语言描述取得了较大进展，且视频关键帧的语义信息对整个视频的描述有较大帮助。本文采用文献[21]中的方法提取关键帧中的语义信息，然后用多层 LSTM 对获取的语义信息和视觉信息进行编码，并将各层 LSTM 的隐藏层状态作为解码阶段的输入，最终生成视频的语言描述。

2.1　视觉信息提取

视频中存在大量的物体、场景及时空关系等多维度信息，提取并融合这些信息对视频的准确描述至关重要，本文采用迁移学习的方法提取目标视频的视觉特征。为解决源域和目标域特征分布不一致的问题，本文采用在多个源域预训练好的模型提取目标域的特征。具体而言，对于所要描述视频中的物体和场景特征提取，采用在数据集 Imagenet[22]和 Place365[23]上预训练的 2DCNN 模型。Imagenet 数据集用于图像分类任务的模型训练和检测，包含约 1 400 万张图像，共 1 000 个物体类别。Palce365 数据集用于场景识别，包含约 1 000 万张场景图片，共 365 个场景类别。对于视频中时空动态特征的提

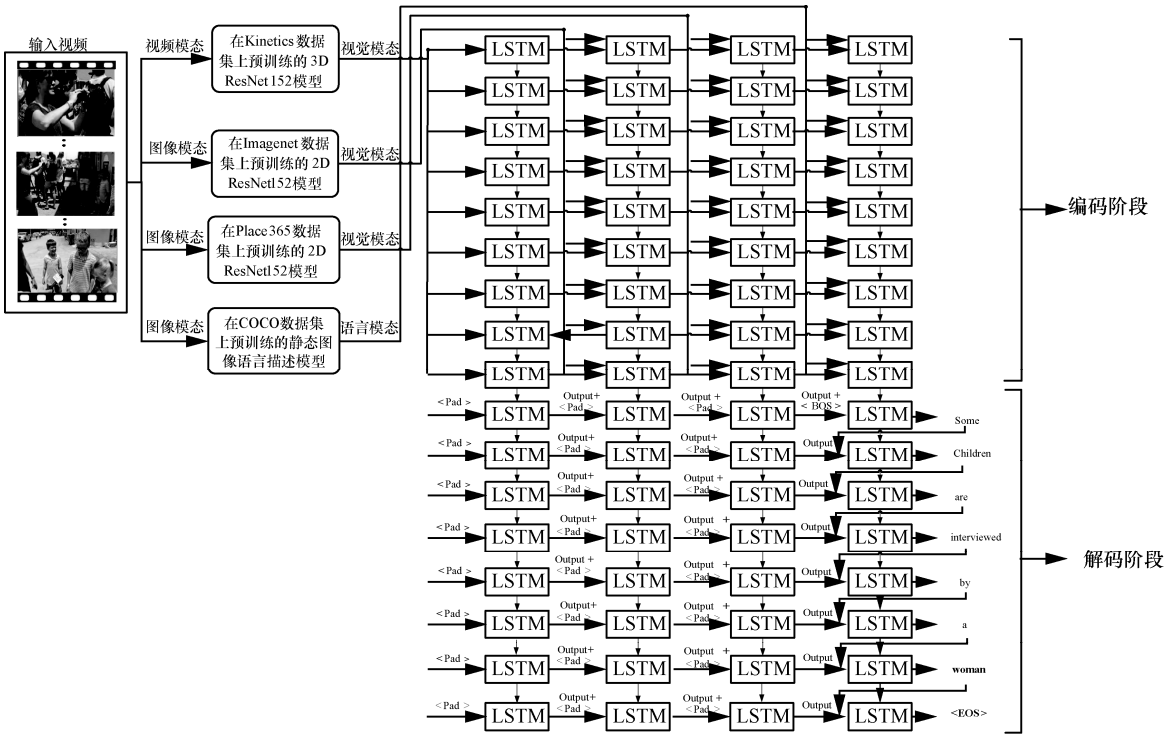


图 1　模型原理

取，采用在行为识别 Kinetics[24]数据集上预训练的 3DCNN 模型。本文采用的 Kinetics 数据集包含约 50 万个视频片段，平均视频长度约为 10 s。

对于 2DCNN 和 3DCNN 模型，均使用 152 层的残差网络（ResNet, residual network）[25]。ResNet 因其网络深度、残差块以及在图像分类问题上的优异表现而成为经典的 CNN。该网络引入残差块解决了梯度消失问题。152 层的二维残差网络（2DResNet152, two dimensional ResNet152）和 152 层的三维残差网络（3DResNet152, three dimensional ResNet152）分别采用二维和三维的卷积核，并且以残差块为其基本单元。残差块如图 2 所示。其中图 2(a)表示 2DResNet152 的卷积块，其卷积核大小分别为 1×1 和 3×3，F 表示卷积核的数量，BN 表示批量标准化（BN, batch normalization）[26]，ReLU 表示激活函数。图 2(b)表示 3DResNet152 的残差块，其卷积核大小分别为 1×1×1 和 3×3×3。

2.2　语义信息提取

视频是由多帧静态图像按时序构成的，因此提取视频中的关键帧并对其内容进行描述有助于理解整个视频的内容。本文采用在微软通用目标检测（COCO, microsoft common object in context）[27]数据集上预训练的静态图像描述模型，对视频的关键帧进行描述。视频中往往包含大量的冗余信息，因而只需抽取视频中若干关键帧进行描述。由于视频内容大多是缓慢变化的，因此采用固定时间间隔的方法来提取 5 个关键帧作为图像描述方法的输入，而输出为图像的语言描述。然后将描述语句中的单词表示为 512 维度的向量，将 2 个单词拼接成 1 024 维度，最后将拼接后的词向量与视觉特征向量进行融合。视觉语义特征提取单元原理如图 3 所示。

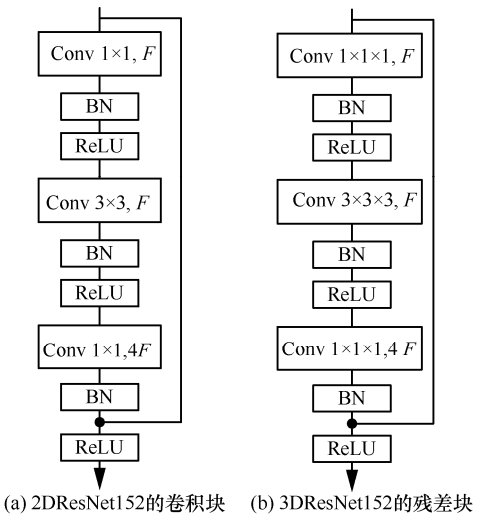


图 2　二维和三维 ResNet152 网络的残差块

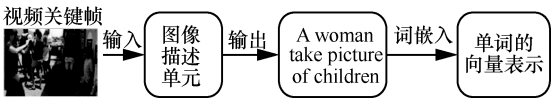


图 3　视觉语义特征提取单元原理

2.3　特征融合模型

如图 1 所示，第一层 LSTM 主要对动态信息建模，将 3DCNN 输出的特征向量按照时序作为 LSTM 的输入。将第一层 LSTM 的输出与场景特征向量拼接后作为第二层 LSTM 单元的输入。同理，由第三层和第四层 LSTM 完成物体信息和语义信息的特征融合建模。将视频按照时序每 16 帧一组作为 3DCNN 的输入，取平均池化后的输出作为动态信息的特征向量。由于视频的长度不一，3DCNN 产生的视频动态特征向量个数不同。本文取 50 个该特征向量，对于超过 50 个的则将多余的部分舍去；对于不足 50 个特征向量的，不足部分用维度相同的零向量代替。对于场景特征向量和物体特征向量均取 50 个。而对于语义信息，将视频等间隔切分，取其中 5 帧生成静态图的语言描述。

本文的视频描述模型以 LSTM 为基础网络进行编码，LSTM 可以有效地提取时序特征，LSTM 原理如图 4 所示。假设在时刻 t，输入的特征向量为 $\boldsymbol{x}_t$，对应输入的隐藏层特征为 $\boldsymbol{h}_{t-1}$，记忆单元的特征为 $\boldsymbol{c}_t$，则任意时刻 LSTM 单元的计算式为

$$\boldsymbol{i}_t=\sigma(W_{xi}\boldsymbol{x}_t+W_{hi}\boldsymbol{h}_{t-1}+b_i) \tag{1}$$

$$\boldsymbol{f}_t=\sigma(W_{xf}\boldsymbol{x}_t+W_{hf}\boldsymbol{h}_{t-1}+b_f) \tag{2}$$

$$\boldsymbol{o}_t=\sigma(W_{xo}\boldsymbol{x}_t+W_{ho}\boldsymbol{h}_{t-1}+b_o) \tag{3}$$

$$\boldsymbol{g}_t=\phi(W_{xg}\boldsymbol{x}_t+W_{hg}\boldsymbol{h}_{t-1}+b_g) \tag{4}$$

其中，i_t、f_t、o_t、g_t 分别为 LSTM 的输入门、遗忘门、输出门、输入调制门，W 和 b 为需要优化的参数。

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\boldsymbol{g}_t \tag{5}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\phi(\boldsymbol{c}_t) \tag{6}$$

其中，σ 为 sigmoid 函数，ϕ 为 tanh 函数，$\odot$ 为哈达玛积运算，即向量的对应元素相乘。σ 和 ϕ 的计算式分别为

$$\sigma(x)=\frac{1}{1+\mathrm{e}^{-x}} \tag{7}$$

$$\phi(x)=\frac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}} \tag{8}$$

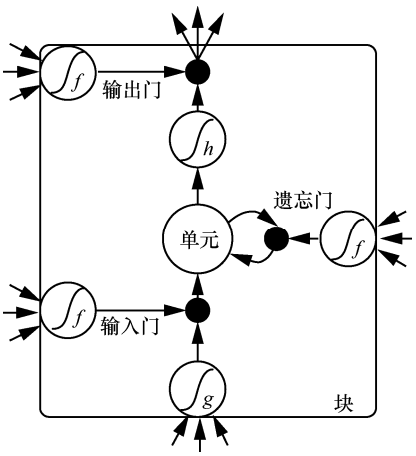


图 4　LSTM 单元原理

在编码阶段，连续选取视频的 16 帧作为 3DResNet152 的输入，经过网络计算得到平均池化输出的 2 048 维向量 $\boldsymbol{X}^d(x_1^d,x_2^d,x_3^d,\cdots,x_t^d,\cdots,x_n^d)$。等间隔地选取视频的 50 帧分别作为 2DResNet152 在 Imagenet 和 Place365 预训练模型的输入，选取 2DResNet152 网络的平均池化输出的 2 048 维向量作为输出，获取场景描述向量 $\boldsymbol{X}^s(x_1^s,x_2^s,x_3^s,\cdots,x_t^s,\cdots,x_n^s)$ 和物体描述向量 $\boldsymbol{X}^g(x_1^g,x_2^g,x_3^g,\cdots,x_t^g,\cdots,x_n^g)$。等时间间隔地选取视频中的 5 帧作为在 COCO 数据集上的静态图像描述模型的输入，输出为 5 句语言描述，采用词嵌入方法将每一个单词编码成 512 维的向量，2 个向量拼接到一起组成一个 1 024 维的向量 $\boldsymbol{s}_t\in\boldsymbol{S}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3,\cdots,\boldsymbol{s}_t,\cdots,\boldsymbol{s}_n)$。其中 LSTM 单元的隐藏层状态表示为 $\boldsymbol{H}(\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3,\cdots,\boldsymbol{h}_t,\cdots,\boldsymbol{h}_n)$，$\boldsymbol{h}_{n+t-1}$ 表示 LSTM 某一时间步的隐藏层状态。模型产生的句子描述为 $Y(y_1,y_2,y_3,\cdots,y_t,\cdots,y_m)$。$Y$ 关于 $\boldsymbol{X}^d$、$\boldsymbol{X}^s$、$\boldsymbol{X}^g$ 和 $\boldsymbol{S}$ 的条件概率分布可表示为

$$P\left(\frac{Y}{\boldsymbol{X}^d},\boldsymbol{X}^s,\boldsymbol{X}^g,\boldsymbol{S}\right)=$$

$$P\left(y_1,\cdots,y_m\mid x_1^d,\cdots,x_n^d;x_1^s,\cdots,x_n^s;x_1^g,\cdots,x_n^g;s_1,\cdots,s_n\right)=$$

$$\prod_{t=1}^{m}P\left(y_t\mid\boldsymbol{h}_{n+t-1},y_{t-1}\right) \tag{9}$$

其中，条件概率分布 $P(y_t\mid\boldsymbol{h}_{n+t})$ 是整个词汇集在 softmax 层对应的输出概率。对于给定的帧序列 F，获取的 $\boldsymbol{X}^d$、$\boldsymbol{X}^s$、$\boldsymbol{X}^g$ 和 $\boldsymbol{S}$ 特征描述向量使 $P\left(\frac{Y}{\boldsymbol{X}^d},\boldsymbol{X}^s,\boldsymbol{X}^g,\boldsymbol{S}\right)$ 最大的词序列 Y 即为预测的句子。本文模型采用 < BOS > 标记表示视频帧序列输

入结束并开始产生语言序列，提示 LSTM 从编码阶段切换到解码阶段。训练阶段采用 < BOS > 标记表示描述语句的次序列输入完成，而测试阶段 < BOS > 表示模型完成描述语句的输出。在完成一个“视频–句子”对的输入并计算信息损失后，通过式(10)来调整模型的参数。

$$\theta^* = \arg\max \sum_{t=1}^{m} \log P\left(y_t \mid \boldsymbol{h}_{n+t-1}, y_{t-1}; \theta\right) \tag{10}$$

3 实验结果与分析

3.1 视频描述数据集

模型验证采用常用的 MSVD 和 MSR-VTT 这 2 个公共数据集。MSVD 是由微软研究院于 2010 年公开的公共数据集，该数据集由 1 970 个视频片段构成，平均每个视频片段包含 40 个人工标注语句。MSR-VTT 是由微软于 2016 年公开的一个用于测试视频描述模型的公共数据集。该数据集由 10 000 个视频片段构成，平均每个视频片段包含 20 个人工标注语句。模型训练均采用上述数据集中的英文标注语句。

3.2 评估指标

对模型结果的评价采用 METEOR、BLEU（bilingual evaluation understudy）、ROUGE-L 和 CIDEr 共 4 种指标。

METEOR 指标基于 wordnet 同义词库，预先给定一组校准 Z，通过最小化对应语句中连续有序的块 C 来得出，计算式为

$$P_n = \gamma\left(\frac{C}{Z}\right)^{\varphi} \tag{11}$$

$$F_{\text{mean}} = \frac{P_Z R_Z}{\alpha P_Z + (1-\alpha) R_Z} \tag{12}$$

$$P_Z = \frac{|Z|}{\sum_k h_k(c_i)} \tag{13}$$

$$R_Z = \frac{|Z|}{\sum_k h_k(s_{ij})} \tag{14}$$

$$\text{METEOR} = (1 - \gamma \text{frag}^{\beta}) F_{\text{mean}} \tag{15}$$

其中，α、γ 和 φ 均为用于评价的默认参数，$h_k(c_i)$、$h_k(s_{ij})$ 分别是一个 N 元模子（n-gram）出现在候选句子 c_i 和标注句子 s_{ij} 中的次数；METEOR 标准基于单精度的加权调和平均数和单字召回率，其结果和人工判断的结果有较高的相关性。

BLEU 是由 IBM 提出的一种常用的机器翻译评测方法，BLEU 定义 n 元词的个数来度量生成结果和目标语句之间的语义相似度。因此该方法首先统计候选语句和参考语句中 n 元词的个数，然后相除即为精确率结果。实验采用 BLUE-4 评测方法，其中 4 为 n 元词中 n 的个数。计算式为

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C'} \text{Count}_{\text{clip}}(n\text{-gram})} \tag{16}$$

$$\text{BP} = \begin{cases} 1, c > r \\ \mathrm{e}^{\left(1-\frac{r}{c}\right)}, c \leqslant r \end{cases} \tag{17}$$

$$\text{BLEU} = \text{BP} \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{18}$$

其中，BP 为惩罚因子，p_n 为计算所预测的 n-gram 的精确率，N 为元组个数。

ROUGE-L 相比于 BLEU，不需要预先指定 n-gram 的值，因为其考虑的是单词的顺序性，更能评价句子层级的意义。但 ROUGE-L 只考虑了一个最长子序列，其他序列不能对其产生影响。ROUGE-L 的评价计算式为

$$\text{ROUGE-L} = \frac{(1+\beta^2) R_{\text{lcs}} P_{\text{lcs}}}{R_{\text{lcs}} + P_{\text{lcs}}} \tag{19}$$

CIDEr 是 Vedantam 等[28]于 2015 年提出的专门用于图像或视频描述的一种评价指标，旨在度量待评价语句和真实语句间的匹配程度。其主要原理是，将待评价与真实语句表示成词频和逆向词频（TF-IDF, term frequency-inverse document frequency）的向量形式，然后求其余弦相似度来对句子进行打分。

3.3 实验过程

在 UltraLab 图形工作站上进行方法验证，采用 Pytorch 深度学习架构。工作站配备 24 核的 Intel Xeon Gold 6146 CPU 和 8 块英伟达公司的 GTX2080 图形处理器，工作站的内存大小为 200 GB。将数据集 MSVD 和 MSR-VTT 均划分为训练集、验证集和测试集。其中 MSVD 训练集为 1 300 个视频片段，验证集为 200 个视频片段，测试集为 470 个测试片段；MSR-VTT 的训练集为 6 513 个视频片段，验证集为 609 个视频片段，测试集为 2 878 个视频片段。

首先，构建用于描述视频内容的字典，分别统计 MSVD 和 MSR-VTT 数据集语料库中的单词频

率。去除语料库中的生僻字和标点符号，保留单词频率大于或等于 5 的单词，对于不可识别的符号用 < UNK > 代替，并添加开始标识符 < BOS > 和结束标识符 < EOS > 。通过以上限定条件，MSVD 和 MSR-VTT 数据集构建的字典大小分别为 7 562 和 9 729。然后，根据第 2 节所述方法进行视频的多维度和多模态特征的提取和融合来完成编码。解码采取贪心搜索方法，即解码时的 LSTM 每个时间步取概率最大的单词作为句子成员，直到解码完成。

训练模型的批量处理和学习率分别设定为 128 和 0.001。模型生成句子的最大长度设定为 28 个单词，目标函数采用负对数似然损失函数，用于度量人工标注句子和模型生成句子之间的距离。采用 Pytorch 架构中的 Adam 优化器对模型的参数进行优化，训练的 epoch 设定为 40。测试时对模型生成的句子用 3.2 节所述的方法进行评估。评估的参考句子为目标视频的人工标注语句，计算结果取其中最大值。

3.4　结果分析

实验结果如表 1 所示。由表 1 中结果可知，本文方法在 MSVD 和 MSR-VTT 这 2 个公共数据集上的得分最高。这是因为本文方法融合了多维度信息，在 Kinetics 上预训练的 3DResNet152 网络可以有效提取数据的动态信息，在 Imagenet 物体分类数据集上预训练的 2DResNet152 网络可以有效提取视频中的视觉场景中的背景和目标信息，在 Place365 场景识别数据集上预训练的 2DResNet152 网络可以有效提取视频的场景信息。除了上述 3 种视觉信息，本文方法还针对视频中的关键帧来提取语义信息，获取的语义信息本质是在 COCO 数据集上预训练的静态图像语言描述迁移到视频的语言描述，然后将这些丰富的视觉和语义 2 种模态信息通过有效方法融合，进而获得更加准确的视频描述。

视频描述的目标是生成更加符合人类语言习惯的语言描述，所提模型对 MSVD 和 MSR-VTT 数据集的视频描述实例如表 2 所示，展示了部分视频的原手工标注语句和模型自动生成的语句。由表 2 可知，与视频片段的原人工描述相比，模型自动生成的描述语句包含的语言要素更加丰富。一方面由于模型获得的特征更加多样化，使其能够适应内容多种多样的视频片段；另一方面是人工标注的语言会受到人们自身知识、兴趣和语言能力的限制，所以所提模型产生了效果较好的描述。

4　结束语

为解决视频的多维度信息的表示和提取，提高语言描述的质量，采用多种视觉任务预训练模型有效地提取视频中丰富的静态和动态视觉信息，并结合视频关键帧的语义信息，构建模型融合多维度和多模态信息，进而生成整个视频的语言描述。3DResNet152 深度卷积神经网络具有良好的时空特征表示的特点，可以提取目标视频的动态信息；而在 Imagenet 和 Palce365 公共数据集上预训练的 2DResNet152 深度卷积神经网络，可以对场景中的背景和物体等静态信息进行特征表示。多维度的动态信息和静态信息形成互补，为视频的语言描述提供丰富的视觉特征。此外，采用在 COCO 数据集预训练的静态图像语言生成模型对视频的关键帧进行语言描述，可为视频语言描述提供较丰富的语义信息。用多层 LSTM 模型将获得的多维度视觉和语义 2 种模态信息进行融合，生成较高质量的视频语言描述。通过在 MSVD 和 MSR-VTT 这 2 个数据集上实验验证，所提方法在常用的 METEOR、BLEU、ROUGE-L 和 CIDEr 评价指标上均获得了

表 1　MSVD 和 MSR-VTT 的测试集在 BLEU4、METEOR、ROUGE-L、CIDEr 上的测试结果

方法	MSVD				MSR-VTT			
	METEOR	BLEU4	ROUGE-L	CIDEr	METEOR	BLEU4	ROUGE-L	CIDEr
MPool[12]	29.1	33.3	—	—	23.7	30.4	52	35
SA[14]	29.6	41.9	—	51.7	25	28.5	53.3	37.1
S2VT[13]	29.8	—	—	—	25.7	31.4	55.9	35.2
VidLAB[29]	—	—	—	—	27.7	39.1	60.6	44.4
TDDF[30]	27.8	37.3	59.2	43.8	—	—	—	—
PickNet[17]	33.1	46.1	62.9	76	27.2	38.9	59.5	42.1
本文方法	33.6	46.7	65	76.8	28.5	39.3	61.2	44.6

表 2　　MSVD 和 MSR-VTT 数据集的视频描述实例

数据集	视频	原手工标注	本文方法描述结果
MSVD		A group of children on stage. A group of children reciting a poem onstage. A group of children singing a song on a stage.	A group of children sang to the audience on a stage in a studio.
		A man gettin flushed in the water. A man is going into a tunnel. A man is playing in pool.	A man slides down a swimming slide.
		A group of men run a track race. Men are running a race. Athletes are running down the track. Men are racing around a track.	A group of athletes raced on a playground.
MSR-VTT		A man is constructing a model. A man is standing. A workplace environment people doing work.	A group of people were doing experiments in a laboratory.
		A lecturer is talking to his classroom. A lot of people are waiting for a lecture. A man is giving lecture in class.	A teacher was teaching a group of students in a classroom.
		Man in wheelchair with broken arm and leg sits near to a table with movable top. The man in the wheel chair is setting up a prank. where the bow on the table will spill on him.	A nanny was serving a man with injured arms and legs.

良好的效果。

在今后的研究中，将继续探索多维度和多模态信息的特征表示和融合方法，并且将探讨注意力机制，让模型能够聚焦更加有用的信息，进一步提升视频描述的质量。

参考文献：

[1] KOJIMA A, IZUMI M, TAMURA T, et al. Generating natural language description of human behavior from video images[C]//15th International Conference on Pattern Recognition. ICPR, 2000: 728-731.

[2] ZHAO B, LI X, LU X. CAM-RNN: co-attention model based RNN for video captioning[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5552-5564.

[3] PARK J, SONG C, HAN J. A study of evaluation metrics and datasets for video captioning[C]//2017 International Conference on Intelligent Informatics and Biomedical Sciences. ICIIBMS, 2017: 172-175.

[4] YI B, YANG Y, FUMIN S, et al. Describing video with attention-based bidirectional LSTM[J]. IEEE Transactions on Cybernetics, 2018, 49(7): 1-11.

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

[6] SHEN Z, LI J, SU Z, et al. Weakly supervised dense video captioning[C]//IEEE Conference on Computer Vision and Pattern Recognition. CVPR, 2017: 1916-1924.

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

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

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

[10] THOMASON J, VENUGOPALAN S, GUADARRAMA S, et al. Integrating language and vision to generate natural language descriptions of videos in the wild[C]//International Conference on Computational Linguistics. ICCL, 2014: 1218-1227.

[11] JOHNSON M, SCHUSTER M, LE Q V, et al. Google's multilingual neural machine translation system: enabling zero-shot translation[J]. Transactions of the Association for Computational Linguistics, 2017, 5(2): 339-351.

[12] VENUGOPALAN S, XU H, DONAHUE J, et al. Translating videos to natural language using deep recurrent neural networks[C]//2015 Confe-

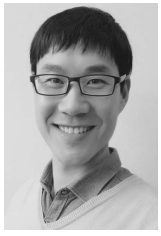
rence of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1494-1504.
[13] VENUGOPALAN S, ROHRBACH M, DONAHUE J, et al. Sequence to sequence-video to text[C]//IEEE International Conference on Computer Vision. ICCV, 2015: 4534-4542.
[14] YAO L, TORABI A, CHO K, et al. Describing videos by exploiting temporal structure[C]//IEEE International Conference on Computer Vision. ICCV, 2015: 4507-4515.
[15] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[16] JIN Q, CHEN J, CHEN S, et al. Describing videos using multi-modal fusion[C]//The 24th ACM International Conference on Multimedia. ACM, 2016: 1087-1091.
[17] CHEN Y, WANG S, ZHANG W, et al. Less is more: picking informative frames for video captioning[C]//The European Conference on Computer Vision. ECCV, 2018: 358-373.
[18] CHEN T H, LIAO Y H, CHUANG C Y, et al. Show, adapt and tell: adversarial training of cross-domain image captioner[C]//IEEE International Conference on Computer Vision. ICCV, 2017: 521-530.
[19] CHEN D L, DOLAN W B. Collecting highly parallel data for paraphrase evaluation[C]//The 49th Annual Meeting of the Association for Computational Linguistics. ACL, 2011: 190-200.
[20] XU J, MEI T, YAO T, et al. MSR-VTT: a large video description dataset for bridging video and language[C]//IEEE Conference on Computer Vision and Pattern Recognition. CVPR, 2016: 5288-5296.
[21] XU K, BA J, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[J]. Computer Science, 2015, 2(1): 2048-2057.
[22] DENG J, DONG W, SOCHER R, et al. Imagenet: a large-scale hierarchical image database[C]//IEEE Conference on Computer Vision and Pattern Recognition. CVPR, 2009: 248-255.
[23] ZHOU B, LAPEDRIZA A, KHOSLA A, et al. Places: a 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1452-1464.
[24] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//The IEEE Conference on Computer Vision and Pattern Recognition. CVPR, 2017: 6299-6308.
[25] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning. ICML, 2015: 448-456.
[26] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. CVPR, 2016: 770-778.
[27] CHEN X, FANG H, LIN T Y, et al. Microsoft coco captions: data collection and evaluation server[J]. arXiv Preprint, arXiv:1504.00325, 2015.
[28] VEDANTAM R, ZITNICK C L, PARIKH D, et al. Cider: consensus-based image description evaluation[C]//IEEE Conference on Computer Vision and Pattern Recognition. CVPR, 2015: 4566-4575.
[29] RAMANISHKA V, DAS A, PARK D H, et al. Multimodal video description[C]//The 24th ACM International Conference on Multimedia. ACM, 2016: 1092-1096.
[30] ZHANG X, GAO K, ZHANG Y, et al. Task-driven dynamic fusion: Reducing ambiguity in video description[C]//IEEE Conference on Computer Vision and Pattern Recognition. CVPR, 2017: 3713-3721.

[作者简介]

丁恩杰（1962− ），男，山东青岛人，博士，中国矿业大学教授，主要研究方向为工业物联网、模式识别、人员定位等。

刘忠育（1985− ），男，河南辉县人，中国矿业大学博士生，主要研究方向为计算机视觉、自然语言处理等。

刘亚峰（1985− ），男，江苏徐州人，博士，中国矿业大学助理研究员，主要研究方向为机器学习、计算机视觉、行为识别等。



郁万里（1987− ），男，江苏徐州人，博士，不来梅大学在站博士后，主要研究方向为工业物联网、网络优化、移动边缘计算等。