跨语言知识蒸馏的视频中文字幕生成

侯静怡 齐雅昀 吴心筱 贾云得

(北京理工大学计算机学院智能信息技术北京市重点实验室 北京 100081)

摘 要 视频字幕生成(video captioning)在视频推荐、辅助视觉、人机交互等领域具有广泛的应用前景.目前已有大量的视频英文字幕生成方法和数据,通过机器翻译视频英文字幕可以实现视频中文字幕的生成.然而,中西方文化差异和机器翻译算法性能都会影响中文字幕生成的质量.为此,本文提出了一种跨语言知识蒸馏的视频中文字幕生成方法.该方法不仅可以根据视频内容直接生成中文语句,还充分利用了易于获取的视频英文字幕作为特权信息(privileged information)指导视频中文字幕的生成.由于同一视频的英文字幕与中文字幕之间存在语义关联关系,本文方法从中学习到与视频内容相关的跨语言知识,并利用知识蒸馏将英文字幕包含的高层语义信息融入中文字幕生成.同时,通过端到端的训练方式确保模型训练目标与视频中文字幕生成任务目标的一致性,有效提升中文字幕生成性能.此外,本文还对视频英文字幕数据集 MSVD 扩展,给出了中英文视频字幕数据集 MSVD-CN.

关键词 中文字幕生成;视频理解;知识蒸馏;视频中英字幕数据集;特权信息

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2021.01907

Cross-Lingual Knowledge Distillation for Chinese Video Captioning

HOU Jing-Yi QI Ya-Yun WU Xin-Xiao JIA Yun-De

(*Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing* 100081)

Abstract Video captioning aims to automatically generate the natural language descriptions of a video, which requires understanding the visual content and describing it with grammatically accurate sentences. Video captioning has wide applications in video recommendation, vision assistance, human-robot interaction and many other fields, and has attracted growing attention in the fields of computer vision and natural language processing. Although remarkable progress has been made on English video captioning, using other languages such as Chinese to describe a video remains under-explored. In this paper, we investigate Chinese video captioning. However, the insufficiency of paired videos and Chinese captions makes it difficult to train a powerful model for Chinese video captioning. Since there exist many English video captioning methods and training data, it is a feasible method to perform Chinese video captioning by translating the English captions via machine translation. However, the difference between Chinese and Western cultures and the performance of machine translation algorithms will both affect the quality of generated Chinese captions. To this end, we propose a cross-lingual knowledge distillation method for Chinese video captioning. Based on a two-branches structure, our method does not only directly generate Chinese captions according to the video content, but also takes full advantage of the easily accessible English video captions as the privileged information to guide the generation of

收稿日期:2020-09-28;在线发布日期:2021-02-18.本课题得到国家自然科学基金(62072041)资助.**侯静怡**,博士,主要研究方向为计算机视觉、视频分析.E-mail:b2082893@ustb.edu.cn.**齐雅昀**(共同一作),硕士研究生,主要研究方向为计算机视觉、视频分析.E-mail:qiyayun@bit.edu.cn.**吴心筱**(通信作者),博士,副教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、图像视频内容理解.E-mail:wuxinxiao@bit.edu.cn.**贾云得**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、认知计算和智能系统.

Chinese video captions. Since the Chinese and English captions are semantically correlated with respect to the video content, our method learns cross-lingual knowledge from them and utilizes knowledge distillation to integrate the high-level semantic information in English captions into Chinese captions generation. Meanwhile, the consistency between the training target and the captioning target is guaranteed by the end-to-end training strategy, thus effectively improving the performance of Chinese video captioning. Benefit from the mechanism of knowledge distillation, our method only utilizes English captions data during the training stage, and after training it can directly generate Chinese captions from the input video. To verify the universality and flexibility of our cross-lingual knowledge distillation method, we use four mainstream visual captioning models for evaluation, covering the CNN-RNN structure, RNN-RNN structure, CNN-CNN structure and model based on Top-Down attention mechanism. These models are widely used as the backbone models in a large number of visual captioning methods. Moreover, we extend the English video captioning dataset MSVD into a cross-lingual video captioning dataset with Chinese captions, called MSVD-CN. MSVD-CN contains 1970 video clips collected from the Internet and 11758 Chinese captions besides the original 41 English captions per video in MSVD. In order to reduce the annotation mistakes caused by annotators' typos or misunderstandings of the video contents, we propose two automatic inspection methods to perform semantic and syntactic checks, respectively, on the collected manual annotations in the data collection stage. Extensive experiments are carried out on the MSVD-CN dataset, via four widely used evaluation metrics for video captioning including BLEU, METEOR, ROUGE-L, and CIDEr. The results demonstrate that the superiority of proposed cross-lingual knowledge distillation on Chinese video captioning. Furthermore, we also report some qualitative experiment results to show the effectiveness of our method.

Keywords Chinese video captioning; video understanding; knowledge distillation; cross-lingual video captioning dataset; privileged information

1 引 言

视频字幕生成旨在自动生成描述视频内容的语句,是计算机视觉与自然语言处理两大领域的交叉融合课题,在视频推荐、辅助视觉、人机交互等领域具有广泛的应用前景.视频字幕生成模型既需要具有理解视频内容的能力,也需要具备充足的语言知识,从而生成忠实流畅的描述语句.

现有的大多数视频字幕生成方法集中于英文字幕生成,而中文字幕生成方法较少,可获取的视频中文字幕数据资源也较少.一种简单的视频中文字幕生成方法是先生成英文字幕,然后利用机器翻译,将英文字幕翻译成中文字幕.Boroditsky[1]的研究指出,不同的语言塑造不同的抽象思想.无论是语言表述习惯还是抽象思维方式,中英文两种语言都因中西方文化差异而不同.而视频字幕生成方法的目标是不仅要准确描述视频内容,还需要母语使用者认同语句的流畅性.因此,机器翻译不是解决视频中文字幕生成的最佳途径.

图 1 是视频中文字幕的直接生成与机器翻译生成的结果比较.图中第一行的"演奏音乐"通常形容类似音乐会的场景.与之相比,"男孩在弹钢琴"对该视频的描述更加忠实准确.这两种中文字幕的差异主要由中英文对应的不同抽象思想在视频理解以及表达方式上的不同所导致,换言之,机器翻译的结果

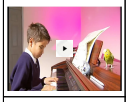
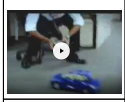
视频	英文字幕	机器翻译	中文字幕
	A boy is playing music.	一个男孩正在演奏音乐.	男孩在弹钢琴.
	A man talking on the mobile.	一个男人在移动电话上说话.	一个男人在玩遥控汽车.
	The puppy's are playing the bear.	小狗在玩熊.	小狗在照镜子.

图 1 视频中文字幕的机器翻译生成和直接生成比较

更多是英文抽象思想的一种体现.此外,由于视频英文字幕生成与机器翻译两个步骤训练的目标均与视频中文字幕生成任务的目标不一致,即没有以生成忠实流畅的视频中文字幕为目标,从而导致中文字幕结果会积累英文字幕生成和机器翻译的损失,还可能会在表达视频语义以及结构信息时产生不完整、不流畅的情形,最终影响中文字幕生成的效果.

因此,本文研究直接由视频生成中文字幕的模型与方法.引入较易获取、信息量丰富的英文字幕作为额外信息,提升视频中文字幕生成模型的性能.鉴于英文字幕表达方式的差异和可能包含的误差信息,本文通过知识蒸馏(Knowledge Distillation,KD)技术[2]将英文字幕建模为特权信息(privileged information),并从中提炼大量知识,用于训练视频中文字幕生成模型.其中,知识蒸馏通过教师-学生(teacher-student)双分支结构把从特权信息中抽取知识和在任务中运用知识分割为两个模块,利用双分支结构分别处理表达结构不同的输入信息;通过将蒸馏英文字幕知识和直接学习生成中文字幕设为目标,在增加信息熵的基础上避免误差的影响.另外,英文字幕仅供教师模型在训练阶段使用,不会增加视频中文字幕生成模型的数据需求以及计算量.使用英文字幕作为特权信息除了其生成方式比较成熟、容易获取外,其包含的高层语义信息能够以一种更加凝练的形式为视频中文字幕生成提供指导.英文字幕所能提供的知识源自同一视频中英文字幕之间会产生的一种语义关联、但不严格对应的关系.例如,对于同一视频,中英文字幕标注内容可能分别为“一个男人正在处理做饭要用到的土豆”与“a man is cutting potatoes in a kitchen”,二者虽均为视频内容的正确描述、且语义相关,但具体的词法与句法却不完全一致.这种关系是视频内容的跨语言知识体现,可以通过知识蒸馏进行提炼与迁移,实现利用额外英文字幕增益视频中文字幕生成性能的目的.

具体地,本文提出一种跨语言知识蒸馏的视频中文字幕生成模型,该模型同时挖掘中文字幕与英文字幕之间的关联关系,以及中文字幕与视频之间的对应关系.该模型基于教师-学生网络,共有两个分支:英文-中文分支作为教师模型以及视频-中文分支作为学生模型.英文-中文分支学习同一视频英文字幕与中文字幕的关联关系,视频-中文分支学习视频与中文字幕的对应关系.通过知识蒸馏的方式将跨语言知识从英文-中文分支传递给视频-中文分支,对中文字幕生成过程进行指导,从而提升视频中文字幕生成的性能.在训练阶段,首先进行英文-中文分支的学习,随后在英文-中文分支的指导下进行视频-中文分支的学习,最后逐渐减少英文-中文分支的指导信息,直至视频-中文分支收敛.在测试阶段,模型只使用视频-中文分支对输入视频进行中文字幕生成,不再需要视频的英文字幕数据.

为了实现上述学习过程,我们对常用的视频英文字幕数据集 MSVD[3] 进行扩展,构建一个包含中英文字幕的跨语言视频字幕数据集,称为 MSVD-CN.MSVD-CN 由 196 个中文母语者手工标注数据集中 1970 个网络视频的中文字幕.其中,每个视频至少包含 5 句中文字幕,整个数据集共收集到 11 758 句中文字幕内容.为确保标注内容在语义上的准确性,本文提出一个自动检测方法计算每一个标注与视频的相关性得分,去除打分较低的句子,即与视频内容无关的干扰语句.与此同时,为了确保标注内容在句法上的准确性,对其规定五条应满足的约束条件,并针对其中较难解决的错别字问题采用一种基于序列的无监督方法,检测并更正标注语句中的字符键入错误.该方法建模句子语义的连贯性,将错别字看作影响语义连贯的异常字符.

本文的主要贡献包括:提出跨语言知识蒸馏的视频中文字幕生成模型,充分提炼英文字幕中的有效知识指导模型的训练,将英文字幕建模为特权信息,确保最终的视频中文字幕生成过程仅需要视频数据作为输入数据,且不增加额外的计算量;将 MSVD 数据集扩展为包含中英字幕的跨语言视频字幕数据集 MSVD-CN,可用于验证模型在视频中文字幕生成、跨模态检索、跨语言迁移等多种任务上的性能;实验结果表明,本文提出的模型在视频中文字幕生成任务的性能明显高于只利用视频与中文字幕的单语言模型,证明利用资源丰富的语种数据能有效增益于资源匮乏语种的视频字幕生成.

2　相关工作

2.1　视频字幕生成模型

现有的视频字幕生成模型大致分为两类:基于模板的模型[4-7]和基于序列的模型[8-13].基于模板的模型首先根据视频创建出句子的结构作为模板(例如:主语-谓语-宾语),然后将对应的词填入模板中.这类模型因为仅能生成固定结构的句子导致得到的结果缺乏语言多样性.基于序列的模型受神经机器翻译(Neural Machine Translation,

NMT)的启发,通过编码器将视频转化为特征表示,随后通过解码器生成描述语句.这类模型不再受限于固定的语句模板,能够更灵活地生成准确多样化的字幕,是近年来视频字幕生成任务的主流模型.序列模型最早可以追溯到 Donahue 等人[8]提出的 LRCNs 模型,该模型对 Rohrbach 等人[7]的方法进行扩展,将基于概率的机器翻译方法替换为由长短时记忆网络(Long Short-Term Memory,LSTM)组成的解码器来生成视频字幕. Venugopalan 等人[14]最早提出用于视频字幕生成任务的端到端序列模型 CNN-LSTM.他们利用 CNN 提取视频关键帧的特征,并利用平均池化进行整合,然后输入 LSTM 来生成描述视频的语句.对于视觉编码,考虑到视频的时序信息,Yao 等人[15]对视频包含的两类时序结构进行建模,利用 3D-CNN 探索代表视频细粒度运动信息的局部结构,而代表视频帧序列顺序的全局结构则通过拥有软注意力机制的 LSTM 进行建模. Pan 等人[16]提出 HRNE 编码器对视频进行多粒度建模,探索视频从局部细节再到整体内容的层级结构. Baraldi 等人[17]提出一种具有时间边界感知的 LSTM 单元(time boundary-aware LSTM cell)来捕捉视频中动作之间的不连续,从而对包含多个动作的视频进行更好的编码.与上述利用固定视频特征表示的模型不同,Venugopalan 等人[9]提出了 S2VT 模型,目前已经成为视频字幕生成的经典模型.该模型利用双层的 LSTM 结构,可以处理变化长度的视频帧序列并获取视频的时序信息.对于语言解码,Gan 等人[18]提出语义集成网络(semantic composition network),为每个语义概念提供一层 LSTM 参数,使得视频中检测出的所有语义概念参与每一时刻的解码过程. Pan 等人[19]提出融合迁移语义属性的 LSTM 模型,将视频帧与视频序列中包含的语义属性整合至序列学习中,提升生成句子的质量.

视频所包含的各种信息对生成字幕的影响程度不同,因此,注意力机制(attention mechanism)被成功引入视频字幕生成任务. Yu 等人[20]提出视线编码注意力网络(gaze encoding attention network),利用视线跟踪(human gaze tracking)数据学习权重,实现对视频信息的时间-空间注意力分配.相较于对已有的视频帧特征进行加权融合,Chen 等人[21]直接利用注意力机制有效选择信息量更多的视频帧,代替了等间隔抽帧的视频特征提取策略,从而在编码阶段更好地利用视频数据.

目前已经有学者研究跨语言图像字幕生成方法. Miyazaki 等人[22]通过共享图像编码器的方式,首先利用图像英文字幕数据集训练,然后移除训练后的 LSTM 解码器并添加一个新的 LSTM 来生成日语字幕,实现英文字幕生成过程到日语字幕生成过程的知识迁移,有效提升图像日语字幕的生成质量.我们的方法与该方法的动机相似,都关注利用一种资源丰富的字幕数据来提高另一种语言的字幕生成模型性能.不同的是我们学习的是同一视频中英文字幕之间基于视频内容的跨语言知识,并通过知识蒸馏的思想指导视频中文字幕的生成,而不是单纯地使用英文数据对模型进行预训练.

2.2 视频字幕数据集

视频字幕生成方法常用的数据集有 MSVD[6]、MSR-VTT[23]、M-VAD[7]等. MSVD 数据集包含 1970 个从 YouTube 网站上收集的视频片段.该数据集虽然包含多种语言的视频字幕标注,但大部分为英语字幕.每个视频大约标注有 40 个英文字幕,而整个数据集却只有 493 个中文字幕,并且为广东话而非普通话,远不足以作为视频中文字幕生成的数据集.另一个常用的数据集是 MSR-VTT 数据集,其包含 10000 个网络视频,每个视频有 20 个标注语句,是一个大规模数据集. M-VAD 和 MPII-MD[24]数据集中的视频采集自电影片段,每个视频由描述视频服务(Descriptive Video Service,DVS)半自动地标注一个句子.

上述数据集提供的标注数据均为英文,缺少其它语种的字幕数据,无法支撑其它语种字幕生成模型的训练,例如本文所关注的视频中文字幕生成任务.因此,本文在 MSVD 视频英文字幕数据集的基础上进行扩展,为数据集中的视频标注中文字幕,构建视频中英文字幕数据集,称为 MSVD-CN.

2.3 知识迁移

知识迁移作为迁移学习的中心思想,致力于将某个领域或任务上学习到的知识运用到不同但相关的领域或任务中.根据任务的特点、领域之间数据分布假设、模型结构设计的差异,往往采用不同的方法实现知识迁移.

基于特征的迁移学习方法[25-26]将源域和目标域数据投影到公共特征空间,然后在公共特征空间最小化源域和目标域数据特征分布的差异,从而将源域知识迁移到目标域上来提高目标域任务的性能.基于样本的迁移[27-29]采用一种特定的权值调整策略,假设源域中的部分数据可以通过重新分配权重

在目标域实现重用. 基于参数/模型的迁移[30-31]复制在源域训练好的部分网络结构与参数，用作目标域模型的一部分.

作为实现知识迁移的一种有效方式，知识蒸馏最早由 Buciluǎ 等人[32]定义，随后经 Hinton 等人[2]推广. 相比于迁移学习探索两种数据的公共特征空间，或复用模型结构与参数的方式，知识蒸馏方法中的教师-学生模型双分支结构，能够分别利用适合自身输入数据特点的网络结构对数据进行学习，并在教师模型的输出处利用软目标进行知识迁移.

受益于信息从教师到学生模型的有效传递，知识蒸馏广泛应用于模型压缩[33-35]中，规模较小的学生模型在规模更大的教师模型或是组合教师模型的指导下，拥有可以媲美大规模模型的性能. 近年来，越来越多的工作[36-40]不再局限于用于模型压缩的知识迁移，转而关注利用知识蒸馏处理各种情况下的知识迁移. 例如 Lopez-Paz 等人[41]将特权信息的概念引入知识蒸馏，在训练过程中，通过教师模型对特权信息进行提炼，并通过知识蒸馏的方式为学生模型提供帮助，学生模型则不会接触特权信息. Zhao 等人[40]在视频在线运动检测中针对模型执行任务时只能获取当前时刻以及之前时刻视频帧的特点，构建输入数据不同的教师模型与学生模型. 将只能在训练阶段获取到的未来视频帧建模为特权信息，通过知识蒸馏将离线教师模型中特权信息的知识传递给在线学生模型. Gupta 等人[36]则将有大量标注的模态(RGB 图)作为训练另一种无标注模态(深度图、光流图)的监督信号，实现图像不同模态之间监督信息的迁移过程. 本文关注利用知识传递过程中，知识蒸馏能够在增加信息熵的基础上避免误差影响的特性，首次提出利用知识蒸馏完成借助资源丰富的语种数据增益资源相对较少语种的视频字幕生成质量的任务.

3　视频中文字幕生成模型

本节提出跨语言知识蒸馏的视频中文字幕生成模型，将易于获取的视频英文字幕建模为特权信息，并学习同一视频中英文字幕之间的关联关系，利用知识蒸馏的方式将习得的跨语言知识迁移至中文字幕的生成过程. 与已有的视频英文字幕生成模型只学习视频到某种语言描述的过程不同，本文模型的整体结构拥有两个分支，分别学习属于同一视频的英文字幕与中文字幕的关联关系(即英文-中文分支)，以及视频内容与中文字幕的对应关系(即视频-中文分支)，其中英文-中文分支视作教师模型，视频-中文分支视作学生模型.

如图 2 所示，模型整体可分为三个关键模块，分别是图中上方虚线框内的英文-中文分支，下方虚线框内的视频-中文分支，以及中间部分的知识蒸馏操作. 知识蒸馏作为一道桥梁，利用教师模型经过温度参数处理后的概率分布输出，将原本训练过程相互独立的两个分支联系起来，实现跨语言知识由英文-中文分支到视频-中文分支的迁移. 由于模型对两个分支的具体结构没有特殊要求，因此分支内结构具有较高的灵活性.

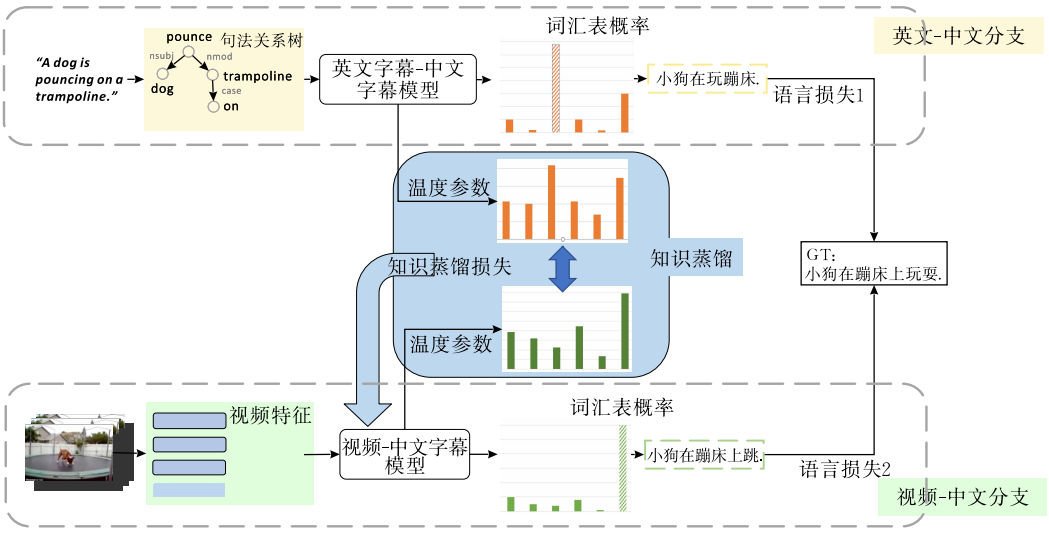


图 2　跨语言知识蒸馏的视频中文字幕生成模型的结构图(图中上方的分支为英文-中文分支，下方的分支为视频-中文分支，中间部分为知识蒸馏过程的示意图)

如图 3 所示，生成模型的训练阶段中，视频-中文分支与英文-中文分支相互配合，分别输入视频以及视频对应的英文字幕，利用知识蒸馏进行跨语言知识的迁移. 而英文字幕作为特权信息仅在训练阶段作为教师模型的输入，学生模型在测试时不再需要教师模型提供的跨语言知识指导，仅使用视频-中文分支的单分支结构即可根据输入的视频直接生成对应的中文字幕.

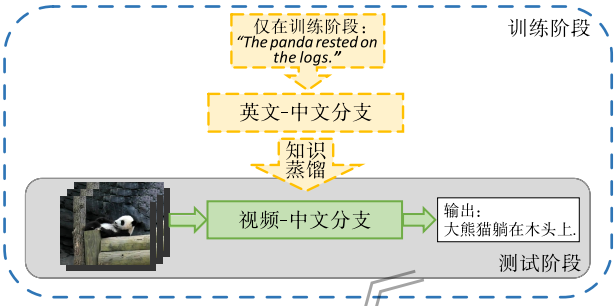


图 3 跨语言知识蒸馏的视频中文字幕生成模型的示意图

3.1 英文-中文分支

作为对同一段视频的描述表示，中文字幕和英文字幕之间存在着一种基于视频内容的关联关系，这其中包含着关于视频内容的跨语言知识. 此外，视频与语言属于两种不同模态，二者存在较大的模态分布差异，而英文字幕与中文字幕同属于自然语言，作为同一模态的事物会具有很多本质上的相似之处. 基于中英文字幕的上述特点，英文-中文分支首先学习从同一视频的英文字幕生成中文字幕的过程. 值得注意的是，这里要学习的关联关系与机器翻译中关注两种语言的语句在词法句法上的严格对应关系不同，中英文字幕之间关联关系的唯一约束是要求二者为描述同一视频的语句. 英文-中文分支在训练过程中将描述同一视频的英文字幕与中文字幕分别作为输入数据与输出真值.

为构建英文字幕的句法图结构，本文利用斯坦福大学 Manning 等人[42]提出的句法依存关系分析工具(Stanford dependency parser)对英文字幕的句法依存关系进行分析以得到句法依存树. 其中句法树的节点代表英文单词实体，而边则代表句法关系，例如关系“nsubj”表示头节点对应的单词为尾节点单词的主语. 将句法依存树所包含的句法关系，整理为句法图结构，其中每一组句法关系都利用形如〈词语$_1$-关系-词语$_2$〉的三元组进行建模. 经过上述操作，英文字幕被表示为对应的句法图结构，该结构中既包含词法信息(即句子描述的具体内容)，也包含句法信息(即句子的句法结构).

针对上述句法图结构，采用方法[43]中的图结构信息更新方式，通过图卷积网络(Graph Convolution Network，GCN)实现图中的信息传递. 经过图卷积操作后，句法关系三元组由更新后的两个词语实体向量以及句法关系向量组成. 将当前英文字幕的所有关系特征向量组合起来得到英文字幕的词法-句法特征表示 $\boldsymbol{s}_i \in \mathbb{R}^{N_i^l \times D_{lang}}$：

$$\boldsymbol{s}_i = [\boldsymbol{s}_{i,1}, \boldsymbol{s}_{i,2}, \cdots, \boldsymbol{s}_{i,N_i^l}] \tag{1}$$

其中 N_i^l 为第 i 句英文字幕所具有的句法关系数量，而 D_{lang} 表示词法-句法特征的维度.

在由英文字幕的词法-句法特征向量生成中文字幕的过程中，引入注意力机制在不同时刻关注不同三元组. 字幕生成模型结构由两层功能不同的 LSTM 组成，分别是拥有自上而下注意力机制的句法关系注意力 LSTM 以及生成中文字幕的语言 LSTM. 句法关系注意力 LSTM 在不同时刻为词法-句法特征中包含的 N_i^l 项关系赋予不同的权重. 其输入特征以及隐藏状态的计算过程如下：

$$\bar{\boldsymbol{s}} = \frac{1}{N_i^l} \sum_{j=1}^{N_i^l} \boldsymbol{s}_{ij} \tag{2}$$

$$\boldsymbol{h}_t^{l,1} = \mathrm{LSTM}([\boldsymbol{h}_{t-1}^{l,2}, \bar{\boldsymbol{s}}, \boldsymbol{W}_e \boldsymbol{\Pi}_t], \boldsymbol{h}_{t-1}^{l,1}) \tag{3}$$

其中 $\boldsymbol{h}_{t-1}^{l,2}$ 为语言 LSTM 上一时刻的输出，$\bar{\boldsymbol{s}}$ 为一句英文字幕对应句法关系特征的平均值，$\boldsymbol{W}_e \in \mathbb{R}^{E \times |\Sigma|}$ 为词汇表 Σ 的词嵌入向量矩阵，$\boldsymbol{\Pi}_t$ 为 t 时刻输入单词的单热点(one-hot)编码. 根据句法关系注意力 LSTM 的输出，计算词法-句法特征中每项关系所分配的权重：

$$a_{i,t} = \boldsymbol{w}_a^{\mathrm{T}} \tanh(\boldsymbol{W}_{va} \boldsymbol{v}_i + \boldsymbol{W}_{ha} \boldsymbol{h}_t^{l,1}) \tag{4}$$

$$\boldsymbol{\alpha}_t = \mathrm{softmax}(\boldsymbol{a}_t) \tag{5}$$

其中 $\boldsymbol{w}_a$，$\boldsymbol{W}_{va}$ 以及 $\boldsymbol{W}_{ha}$ 为可学习的参数.

利用得到的权重 $\boldsymbol{\alpha}_t$ 与句法关系特征进行凸组合，计算得到语言 LSTM 所需要的词法-句法关系特征，该特征代表了当前英文字幕包含的信息，定义为

$$\hat{\boldsymbol{s}}_{i,t} = \sum_{j=1}^{N_i^l} \alpha_{j,t} \boldsymbol{s}_{i,j} \tag{6}$$

而语言 LSTM 则是根据输入的特征来生成相应时刻的词汇表分布，并选择概率最大的单词输出，计算过程可表示为

$$\boldsymbol{h}_t^{l,2} = \mathrm{LSTM}([\hat{\boldsymbol{s}}_{i,t}, \boldsymbol{h}_t^{l,1}]; \boldsymbol{h}_{t-1}^{l,2}) \tag{7}$$

$$P_t = \mathrm{softmax}(\boldsymbol{W}_z \boldsymbol{h}_t^{l,2} + \boldsymbol{b}_z) \tag{8}$$

其中 $\boldsymbol{W}_z$，$\boldsymbol{b}_p$ 分别为可学习的权重以及偏移量.

对于英文-中文分支生成的中文字幕，使用视频字幕生成任务常用的交叉熵损失计算其与数据集标注真值之间的差异作为损失函数来训练该分支，损

失值记为$\mathcal{L}_{en}$.

3.2 视频-中文分支

视频-中文分支实现根据输入的视频内容生成中文字幕.由于双分支模型的灵活性,视频-中文分支可以采用任意端到端视觉字幕生成模型.由于对该分支模型结构的研究并不是本文的重点,因此我们选择了四种已有的经典模型,分别是 *NIC*[44]、*S2VT*[9]、*ConvCap*[45]以及 *Top-Down*[46].这四种模型涵盖了视觉字幕生成模型中最有代表性的 CNN-RNN 结构,RNN-RNN 结构,CNN-CNN 结构以及基于自上而下注意力机制的模型,并在大量字幕生成方法中作为主干模型使用,因此具有可靠性.本文使用这些模型,既可以实现视频-中文分支的作用,也不为实验分析引入额外的干扰因素,避免因字幕生成模型本身性能不稳定造成的影响.除了本文选择的四种经典模型外,在实际使用过程中还可以利用其他视频字幕生成模型作为本文模型的视频-中文分支.对于输入视频,提取其帧级别视觉特征$\boldsymbol{v}_i\in\mathbb{R}^{T\times D_{vis}}$作为该视频的特征表示,其中 T 为视频提取的关键帧数量,D_{vis}为视频特征的维度.下面扼要介绍这四种基础模型的结构.

NIC 是结构为编码器-解码器的图像字幕生成模型,该模型将输入的视觉数据先编码为视觉特征,然后输入由 LSTM 构成的解码器,生成与视觉内容对应的字幕.在本文中,对提取得到的视频帧级别特征进行平均池化操作,将其结果作为 LSTM 解码器初始时刻的视觉信息输入 $\boldsymbol{s}_{-1}$.在每个时刻 t,解码器将上一时刻生成的词向量 $\boldsymbol{s}_{t-1}$表示作为输入并得到当前时刻的词向量 $\boldsymbol{s}_t$,表示为

$$\boldsymbol{s}_{-1}=\frac{1}{N_i^v}\sum_{j=1}^{N_i^v}\boldsymbol{v}_{ij} \tag{9}$$

$$\boldsymbol{s}_{t+1}=\mathrm{LSTM}(\boldsymbol{W}_e s_t),\ t\in\{0,\cdots,N_s-1\} \tag{10}$$

其中 $\boldsymbol{W}_e$是词语的词向量映射矩阵.*NIC* 模型的训练损失为模型生成字幕与数据集给定的标注真值之间的负对数似然函数,定义为

$$L(V,S)=-\sum_{t=1}^{N}\log p_t(S_t) \tag{11}$$

S2VT 由两个串联的 LSTM 组成.由于采用 LSTM 结构对视频帧序列进行处理,*S2VT* 能建模视频中的时序信息.视频字幕的整个生成过程分为两个阶段,分别是编码阶段和解码阶段.在编码阶段,第一层 LSTM 将输入的视频帧特征序列进行编码,并将当前时刻输出的隐藏层状态与空填充标识符(零向量)进行拼接操作,作为第二层 LSTM 的输入数据.在该阶段并不生成字幕内容.而在所有视频帧特征都被送入 LSTM 后,模型进入解码阶段.在解码阶段,第二层 LSTM 输入字幕起始标识符,并继续根据自身上一时刻的隐藏层状态开始输出字幕,此时第一层 LSTM 的输入为空填充标识符.与 *NIC* 相同,*S2VT* 在训练过程中最小化负对数似然函数.

ConvCap 是基于 CNN 的模型,其对帧级别视频特征的编码方式与 *NIC* 一致,同样为平均池化操作.与基于 LSTM 的模型不同,*ConvCap* 在每个时刻中将预测结果与编码的特征相连接作为输入来继续预测下一时刻的结果.*ConvCap* 同样使用负对数似然函数进行模型训练.

Top-Down 根据当前时刻生成的字幕内容,为图像的不同区域分配不同的空间注意力权重.本文针对视频字幕生成任务,为 *Top-Down* 模型增加时间-空间注意力分配,使得 *Top-Down* 在不同时刻既选择关注视频空间信息中的不同区域,又选择关注视频时序信息中的不同帧.该操作由两层功能不同的 LSTM 完成,分别是进行注意力分配的视频注意力 LSTM 与生成中文字幕的语言 LSTM.

视频注意力 LSTM 的输入为平均池化操作后的视频帧级别特征,记作 $\bar{\boldsymbol{v}}$:

$$\bar{\boldsymbol{v}}=\frac{1}{N_i^v}\sum_{j=1}^{N_i^v}\boldsymbol{v}_{ij} \tag{12}$$

根据视频注意力 LSTM 当前时刻的输出,分别计算视频在时间上以及在空间上的注意力分配情况.

时间注意力分配,即对关键帧序列权重 $\beta_{t,i}$的计算过程为

$$\beta_{t,i}=\mathrm{softmax}(\boldsymbol{w}_{\mathrm{time}}^{\mathrm{T}}\tanh(\boldsymbol{W}_{\mathrm{time}}\boldsymbol{v}_i+\boldsymbol{U}_{\mathrm{time}}\boldsymbol{h}_t^1)) \tag{13}$$

其中 $\boldsymbol{w}_{\mathrm{time}}$,$\boldsymbol{W}_{\mathrm{time}}$以及 $\boldsymbol{U}_{\mathrm{time}}$均为可学习参数.根据该权重对视频关键帧特征向量进行加权求和,得到经过时间注意力分配后的视频时间特征表示.类似的,在时间注意力分配后计算空间注意力权重,并通过对视频空间特征加权求和得到视频空间特征表示.利用上述过程得到的视频时间、空间特征作为视频语言生成 LSTM 的输入,生成视频内容对应的中文字幕.

上述四种视频字幕生成模型均采用生成的中文字幕与数据集给定的标注真值之间的损失函数 $\mathcal{L}_{vid}$,进行视频-中文分支的训练.

3.3 知识蒸馏

考虑到英文字幕与中文字幕同属于自然语言,二者数据差异相比跨模态的视觉数据和语言数据的差异更小.我们将英文字幕建模为特权信息在训练阶段作为英文-中文分支的输入数据.并将利用了特权信息的英文-中文分支视作教师模型,而视频-中

文字幕视作学生模型.通过引入知识蒸馏的思想提炼同一视频的中英文字幕所包含的基于视频内容的跨语言知识,并迁移至由视频生成中文字幕的过程中,实现利用学习效果更好的英文-中文分支帮助视频-中文分支在视频中文字幕任务上取得更优结果.受启发于利用特权信息的知识蒸馏工作[40],本文与传统的知识蒸馏不同,教师模型与学生模型之间的差异主要体现在输入数据的不同,而不是模型结构与大小的区别.

我们将英文-中文和视频-中文这两个分支输出单词概率分布的 KL 散度(Kullback-Leibler divergence)作为训练知识蒸馏模型的损失函数,在训练过程中最小化该损失函数.其中计算词语概率分布时采用的 softmax 操作引入了额外的温度参数,可以视作一种更加平滑的概率分布计算方式.这种概率分布与数据集标注数据提供的词语标签真值作用不同,前者是一种软目标,包含更大的信息量,后者是硬目标,无法给出除了正确词语信息之外的知识.例如,一段视频描述的内容大致是"一个男人在厨房处理绿色的食材",英文-中文分支当前时刻已经输出的内容为"一个男人在切",由于先前训练过程的积累以及对于数据集描述内容的理解,该分支下一时刻生成的词汇表概率分布将会趋向于蔬菜类的词语,例如"蔬菜"、"黄瓜"、"花椰菜"等等,而无关的词语比如"风筝"、"蝴蝶"、"摩托车"概率则相对较低.与视频-中文分支注重生成结果与数据集真值一致相比,英文-中文分支可以为模型提供数据集内容的跨语言理解以及通用的中文语言知识,指导模型能够更好地生成中文字幕.

对 t 时刻语言 LSTM 最后一层生成的特征进行处理:

$$Q_{ti}=\frac{\exp(z_{ti}/T_{\text{temp}})}{\sum_j \exp(z_{tj}/T_{\text{temp}})} \tag{14}$$

$$z_t=\boldsymbol{W}\boldsymbol{h}_t^2+\boldsymbol{b} \tag{15}$$

其中 T_{temp} 表示温度参数,当 T_{temp} 趋向于 0 时,Q_t 将收敛为一个单热点向量,而 T_{temp} 越大,Q_t 的分布将越为平滑,而 $\boldsymbol{W}$ 与 $\boldsymbol{b}$ 分别是可学习的权重与偏移.得到两个分支对于视频及其英文字幕所输出的两个软 softmax 概率分布后,计算这两个概率分布的 KL 散度作为知识蒸馏部分的损失函数 $\mathcal{L}_{distill}$:

$$\mathcal{L}_{distill}=\sum_{x\in L,y\in V} Q_l(x)\log\left(\frac{Q_l(x)}{Q_v(y)}\right) \tag{16}$$

其中 $Q_l(x)$ 代表英文-中文分支的软 softmax 概率分布,$Q_v(y)$ 则表示视频-中文分支的软 softmax 概率分布.在该损失函数的指导下,视频-中文分支可以充分利用英文-中文分支中,由中英文字幕关联关系蒸馏得到的跨语言知识,提升视频中文字幕生成模型的性能.

3.4 模型训练策略

如 3.1 小节所述,由于数据类型的差异导致两个分支拥有不同的学习难度.本文根据分支学习难易程度制定训练策略,来对跨语言视频中文字幕生成模型进行训练.整体训练思路遵循一个由易到难的学习过程,由跨语言知识学习效率更高的英文-中文分支作为教师网络来指导后续视频-中文分支模型的学习.在初始训练过程中,先利用英文字幕与中文字幕数据训练英文-中文分支,随后在英文-中文分支的指导下对视频-中文分支进行训练.随着训练进行,逐步减少英文-中文分支的指导信息,直至视频-中文分支收敛.

在两个分支单独训练的时刻,分别最小化各自分支交叉熵损失 $\mathcal{L}_{en}$ 与 $\mathcal{L}_{vid}$ 对自身模型参数进行优化调整.在联合训练阶段,训练损失 $\mathcal{L}$ 由两部分组成:

$$\mathcal{L}=\mathcal{L}_{vid}+\lambda_d\mathcal{L}_{distill} \tag{17}$$

其中 λ_d 代表知识蒸馏指导信息比重的超参数.在指导视频-中文分支训练的过程中,我们固定英文-中文分支的参数,无需通过上述损失进行优化更新.

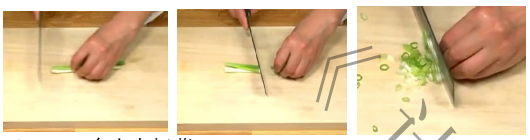
联合训练阶段的训练损失 $\mathcal{L}$ 由 $\mathcal{L}_{vid}$ 与 $\mathcal{L}_{distill}$ 共同组成,意味着同时利用视频中文字幕硬目标与知识蒸馏得到的软目标进行训练.软目标通过知识蒸馏方法引入温度参数,利用信息熵更高,即包含信息量更大的软目标,辅助信息熵较小的硬目标进行训练,可以增加信息量,弥补监督信号不足的问题.而硬目标指导视频-中文分支直接学习视频中文字幕的生成,可以减轻英文字幕数据误差的影响.软目标与硬目标相互协助,在增加知识信息熵的基础上避免误差的影响.

4 MSVD-CN 数据集

视频中文字幕生成模型的训练需要中文数据的支持,因此需要构建一个与已有的英文数据集对应的视频中文字幕数据集.本文构建一个拥有中英文字幕的跨语言视频字幕数据集,称为 MSVD-CN.该数据集由现有常用的视频英文字幕数据集 MSVD 扩展得到.MSVD-CN 数据集包含 1970 个网络视频,除了 MSVD 平均每个视频原有的 41 句英文字幕之外,每个视频至少包含 5 句中文字幕,整个数据集共收集到 11758 句中文字幕内容.图 4 中展示了 MSVD-CN 数据集中部分中文字幕标注的示例.

GT1：守门员扑住一个球.
GT2：足球场上一个球被守门员扑倒了.
GT3：足球场上守门员防住了敌方的进球.
GT4：一个守门员防住了敌方踢过来的足球.
GT5：一个足球门将守住了地方射来的足球.
GT6：守门员将球扑了出来.
GT7：一个守门员扑出了一计射门.

GT1：一个小男孩在弹吉他.
GT2：一个小孩在弹吉他.
GT3：一个男孩在演奏电吉他.
GT4：一个男孩在舞台上弹吉他.
GT5：一个小男孩在弹吉他.



GT1：一个人在切葱.
GT2：一个人在把葱切成葱花.
GT3：一个人在切葱.
GT4：一个人在切葱花.
GT5：一个人在切大葱.

图 4　MSVD-CN 数据示例

为了能够得到准确可靠的中文字幕标注，我们设计了一个视频标注系统，并邀请 196 位受过良好教育且普通话标准的志愿者利用该系统对视频进行标注. 系统的用户界面如图 5 所示，在每次标注前，标注者都需要阅读标注的说明规则.

(a) 用户弹出界面　　(b) 用户标注界面

图 5　用户界面

4.1　标注质量控制

在收集数据阶段，为了减少由于用户对视频内容理解有误或字符错误输入而导致的标注误差，我们提出两种自动检查方法，分别对收集的手工标注进行语义和句法检查.

4.1.1　语义检查

语义检查的目的是过滤出与视频内容无关的标注. 通常情况下，同一视频的大多数标注句子与视频内容是相关的. 基于这个特点，我们提出一个自查法来衡量句子与视频内容的相关性. 该方法衡量当前句子与当前视频的其余句子之间的相似度. 具体地，将当前进行检查的字幕视为候选句子，当前视频的其余字幕视为参考句子. 首先计算候选句子与参考句子中的 n 元组(n-grams)词频-逆文档频率(Term Frequency-Inverse Document Frequency，TF-IDF). 令 $\{c_{mi}\}_{i=1}^{N_c}$ 表示数据集中第 m 个视频的所有描述句子，总数为 N_c. 在句子 c_{mi} 中第 k 个 n 元组 ω_k^n 出现次数为 $\tau_k^n(c_{mi})$，则关于该 n 元组的 TF-IDF 表示为

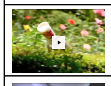
$$g_k^n(c_{mi})=\frac{\tau_k^n(c_{mi})}{\sum_{\omega_l^n\in\Omega}\tau_l^n(c_{mi})}\log\left\{\frac{|V|}{\sum_{V_p\in V}\min\left(1,\sum_q\tau_k^n(c_{pq})\right)}\right\} \tag{18}$$

其中 Ω 是所有 n 元组的集合，V 是 MSVD-CN 数据集所有视频的集合.

将 n 元组词频-逆文档频率与相同长度的句子 c_{qp} 组合得到向量 $\boldsymbol{g}^n(c_{qp})$，则候选句子 c_{mj} 得分为

$$\text{Score}(c_{mj})=\frac{1}{N(N_c-1)}\sum_{n=1}^{N}\sum_{i=1,i\neq j}^{N_c}\cos(\boldsymbol{g}^n(c_{mj}),\boldsymbol{g}^n(c_{mi})) \tag{19}$$

其中 N 是 n 元语法的数量，$\cos(\cdot,\cdot)$代表计算两个向量的余弦相似度. 从得分值的计算过程可以看出，出现频率高的单词，其词频-逆文档频率相对较低从而分数也会降低. 因为大多数的高词频词汇通常无实际意义，如代词、助词等. 因此，该分数很有效地衡量了句子的语义信息，从而可以筛选出语义与大多数描述当前视频的句子不相关的句子. 图 6(a)中是使用上述方法检测出的与视频内容不相关的标注示例.

视频	英文字幕	中文字幕
	A girl is playing the violin.	犹抱琵琶半遮面.(A line from a Chinese poem.)
	A squirrel has a yogurt cup stuck on its head.	红红火火恍恍惚惚.(An Internet slang.)
	A drill sergeant is speaking to some soldiers.	那人在切西兰花.(That man is chopping broccoli.)

(a) 语义检查结果示例

一个男人在举一个很重地（的）哑铃.
一个人在锅里扁（煸）蒜末.
一个老年人在赶（擀）面团.
小猫咪再（在）玩耍.
两个女还（孩）在一起跳舞.

(b) 字符键入错误检查结果示例

图 6　错误标注示例

4.1.2　句法检查

本文从单词数量、字符类型、动词是否存在、命

名实体是否存在以及字符键入错误这五个方面对用户手工标注的中文字幕进行句法检查.其中较难的一项就是要修正标注中文字幕的错别字.考虑到中文错别字通常是因为打字失误出现的同音汉字或是相近汉字,而其本身是正确的汉字,只有通过理解它在句子中的语义才能判断其对错.为了解决这个问题,我们采用基于序列的无监督策略,将错别字看作影响语义连贯的异常字符,来检查标注语句中的字符键入错误.

针对句法检查的五个方面,我们采取下列五项具体措施:

(1)单词数量.筛选出少于5个汉字或多于25个汉字的句子,来保证标注内容的有效性.

(2)字符类型.包含除简体中文之外字符类型的句子需要移除.

(3)动词的存在.使用斯坦福大学Manning等人[42]提出的CoreNLP工具包中的POS(Part-Of-Speech)标签工具分析每句字幕的词语成分,且保证每个句子至少包含一个动词.

(4)命名实体的存在.过于具体的内容,例如一个人的名字,一个地点或一件艺术品的名称是不应该出现在字幕语料库中的,因此利用CoreNLP工具包中的命名实体检测工具移除句子中检测到的命名实体.

(5)字符键入错误.受到异常事件检测方法的启发,我们建立一个端到端的字符键入错误检测方法.如图7所示,该方法将字符的键入错误视为句子中的异常字符来进行检测.训练时使用均方误差(MSE)作为方法的损失函数.测试时,计算输出内容和真实字幕之间的余弦相似度,然后选择相关性较低的字符作为字符错误的候选项进行更正.图6(b)展示了一些检测出的字符错误的示例.

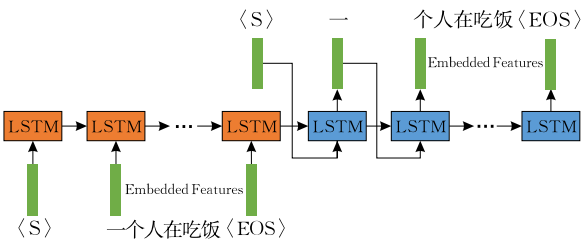


图7 端到端字符键入错误检测器

5 实 验

为了验证本文方法的有效性,我们在MSVD-CN数据集上进行实验,采用BLEU[47]、METEOR[48]、ROUGE-L[49]和CIDEr[50]四种常用的视频字幕任务的评价指标对生成的中文字幕进行性能评价.

5.1 实验设置

在视频特征提取方面,对每个视频随机采样10帧关键帧,不足10帧的用全零向量进行填充.视频关键帧的视觉特征利用ResNet-152[51]网络作为特征提取器,选择网络最后一个平均池化层之前的输出作为特征表示,其特征维度 $D_{vis}=2048$.对于中文描述语句,将每句中文字幕使用CoreNLP工具包中的CRF-based分词器[42],将字幕分割成有意义的词.然后通过预先定义好的词典将每一个词表示为单热点向量,并利用映射矩阵将这些词的单热点向量映射至低维空间(300维)中.对于英文-中文分支所使用的英文字幕,利用NLTK工具[52]将英文句子分割成单词并利用维度为300的Glove词嵌入向量作为英文单词的向量表示,$D_{lang}=300$.在测试阶段,生成中文字幕以及间接方法英文字幕的过程均采用贪心搜索的策略.

5.2 实验结果

本文提出的跨语言知识蒸馏的视频中文字幕生成模型对于视频-中文分支的要求较为灵活,可以采用许多现有的端到端视频语言描述模型或者图像语言描述模型,不要求模型必须包含特定的技术或框架.从验证跨语言知识蒸馏方法有效性的角度出发,我们选择已有的四种经典视频/图像字幕生成模型*NIC*[44]、*S2VT*[9]、*ConvCap*[45]以及*Top-Down*[46],并对其进行细微调整,作为视频-中文分支.

通过分别测试单独视频-中文分支在MSVD-CN上得到的结果,以及在英文-中文分支指导下的视频-中文分支结果,来对本文提出方法的有效性进行评估.值得一提的是,该模型仅在训练阶段用到英文字幕,而在测试阶段与一般的视频字幕生成任务一样,只需要输入视频数据.实验结果如表1所示,其中"NIC-KD"、"S2VT-KD"、"ConvCap-KD"以及"Top-Down- KD",分别代表引入了跨语言知识蒸馏的四种视频中文字幕生成方法.可以看到,对于四种不同的视频-中文分支,本文提出的跨语言知识蒸馏的视频中文字幕生成方法均有提升,验证了本文将英文字幕建模为特权信息并通过知识蒸馏的方式增益视频中文字幕生成的思路是有效的.

表 1　跨语言知识蒸馏视频中文字幕生成模型在 MSVD-CN 数据集上的结果

模型	B@1/%	B@2/%	B@3/%	B@4/%	METEOR/%	ROUGE-L/%	CIDEr/‰
NIC	50.9	37.3	28.3	18.8	21.0	46.5	40.1
NIC-T	44.4	29.6	20.4	11.1	17.7	41.7	16.4
NIC-KD	**51.2**	**38.9**	**30.5**	**21.3**	**21.6**	**47.6**	**42.1**
S2VT	51.5	38.4	29.6	20.6	21.3	47.3	46.3
S2VT-T	41.8	26.8	17.4	9.3	16.4	39.2	18.9
S2VT-KD	**52.1**	**39.5**	**30.7**	**21.1**	**21.6**	**48.0**	**46.8**
ConvCap	49.8	35.7	26.3	16.8	21.1	46.3	40.3
ConvCapp-T	41.9	25.5	16.0	7.1	16.4	39.0	18.3
ConvCap-KD	**50.8**	**37.4**	**28.3**	**19.1**	**21.5**	**47.0**	**42.2**
Top-Down	51.7	38.4	29.2	19.7	22.2	48.2	55.2
Top-Down-T	42.3	27.5	18.2	9.7	17.0	39.6	23.0
Top-Down-KD	**52.9**	**39.8**	**30.4**	**21.1**	**22.9**	**49.4**	**57.4**

我们还将本文方法与间接方法进行对比. 在实现间接方法时，利用英文数据以及前面提到的四种视频-中文分支，分别训练得到相应的视频英文字幕生成模型，并将生成的英文字幕利用谷歌翻译转化成中文字幕. 结果如表 1 所示，其中“NIC-T”、“S2VT-T”、“ConvCap-T”以及“Top-Down-T”，分别代表利用机器翻译将四种视频字幕生成模型输出的英文字幕转换为中文的结果. 可以看到，与直接利用视频与中文字幕数据对进行训练的方法相比，间接方法的评价结果明显降低. 导致该情况的主要原因是间接方法不仅累积了视频英文字幕生成任务以及机器翻译任务的损失，还会因为两个任务的训练目标与视频中文字幕生成任务不一致，使得模型不是直接向着生成忠实流畅的中文字幕这个目标进行学习，从而为方法在视频中文字幕生成任务上的表现引入很多不确定性. 该结果可以表明，本文将视频英文字幕数据集 MSVD 扩展为包含中文字幕的跨语言视频字幕数据集 MSVD-CN，对于视频中文字幕生成任务的研究是有意义的. 与此同时，该结果还验证了跨语言知识蒸馏的视频中文字幕生成方法的有效性. 为了更加直观地展现出方法生成结果的质量，图 8 给出部分定性实验结果. 该图包含同一视频字幕生成模型利用三种不同训练方法所得到的结果以及数据集标注的中文字幕真值，其中图像下方第一行的“Top-Down-KD”表示以 Top-Down 模型为视频-中文分支的跨语言知识蒸馏的视频中文字幕生成方法的结果，而第二行的“Top-Down”表示不引入英文-中文分支的普通视频中文字幕生成方法结果，“Top-Down-T”则表示利用 Top-Down 方法先生成英文字幕再翻译得到中文字幕的结果，“GT”为视频对应中文字幕真值. 从图中可以看出，本文提出的跨语言知识蒸馏视频中文字幕生成方法取得了更优的实验结果，其生成的中文字幕更接近视频内容，语言更为通顺流畅.

Top-Down-KD：一个人在往锅里倒调料.
Top-down: 一个人在往锅里倒鸡蛋.
Top-down-T: 一个男人在锅里加一些鸡.

GT1: 一个人在往锅里倒调料.
GT2: 一个人在往锅里的香肠上撒调料.
GT3: 一个人在香肠里加调料.

Top-Down-KD：一个人骑着摩托车在路上行驶.
Top-down: 一个人在骑自行车.
Top-down-T: 一个男人骑着摩托车.

GT1: 一个男人背着包骑着摩托车.
GT2: 一个人骑着摩托车在路上行驶.
GT3: 一个男人在土地上骑摩托车.

Top-Down-KD：一个人在剥洋葱.
Top-down: 一个人在切洋葱.
Top-down-T: 一个男人正在剥一块肉.

GT1: 一个人在剥洋葱.
GT2 一个人在剥洋葱皮.
GT3: 女人在给切了一半的洋葱剥皮.

Top-Down-KD：一个女人在做虾.
Top-down: 一个女人在做鸡蛋.
Top-down-T: 一个女人正在把鱼放进锅里.

GT1: 一个女人在做饭.
GT2: 一位女士给虾上糊了面粉后扔进锅里炸.
GT3: 一个女人在炸虾.

图 8　部分定性实验结果展示(包含三种方法的生成结果以及数据集标注的中文字幕内容)

图 9 展示了一些失败的定性实验例子. 图 9 中例子(a)生成的字幕均没有准确地描述视频的内容,可以看到引入跨语言知识蒸馏方法的生成结果与基础方法相比,能够额外生成出与“倒”这个动作相关的内容. 但是由于模型所使用的视觉特征未能很好的捕捉到牛奶的颜色与质地,所以将其混淆为“油”. 在未来的工作中,可以使用更多样的视觉特征,或者自己设计提取特征的网络封装进模型中,改善这类问题. 例子(b)中,模型在“猴子”这个词语的生成上,受到数据集的长尾效应(long-tailed problem)的影响,即词语出现频率存在一定的差异,使得生成结果是数据集中动物类词语出现次数最多的“狗”. 例子(c)中,方法中的知识蒸馏在学习跨语言知识时,忽略了小孩子在吃冰激凌的细节,原因可能是数据集中的英文字幕标注以及中文字幕标注均包含两类描述,一类单纯描述孩子在笑,而另一类描述孩子在笑的同时还关注到了孩子与冰激凌之间的互动. 由于在教师模型学习跨语言知识的过程中,不会限制中英文字幕的两类描述一定是对应的,所以模型倾向于学习所有字幕中都出现的内容,也就是该视频中的“小孩子在笑”,保证生成字幕是准确可靠的,但也在一定程度上损失了一些细节描述. 之后的工作中可以继续探索如何权衡两种描述之间的选择.

Top-Down-KD: 一个人在往碗里的鸡蛋里倒油.
Top-down: 一个人在搅拌鸡蛋.

GT1: 一个女人在向装有打好的鸡蛋的碗里倒牛奶.
GT2: 一个女人在向一盆鸡蛋中倒牛奶.

(a)

Top-Down-KD: 一只狗在逗一只狗.
Top-down: 一只狗在地上跑.

GT1: 猴子和小狗在玩耍.
GT2: 一只猴子在戏弄一只狗.

(b)

Top-Down-KD: 一个小孩在笑.
Top-down: 一个孩子在吃东西.

GT1: 一个小孩吃着冰激凌开心地笑着.
GT2: 一个小孩子坐在沙发上在开心地笑.
GT3: A baby laughs while eating an ice cream cone.
GT4: A baby is laughing.

(c)

图 9 部分失败的定性实验结果展示

6 结 论

本文提出了跨语言知识蒸馏的视频中文字幕生成方法. 该方法可以通过学习同一视频中英文字幕之间的关联关系,将蒸馏得到的基于视频内容的跨语言知识用于提高视频中文字幕生成的质量. 我们提出英文-中文分支以提升视频-中文分支的性能,能够有效地利用资源更为丰富的视频英文字幕并将其建模为特权信息,对资源相对较难获取的视频中文字幕生成提供帮助. 实验表明,在多种经典视频字幕生成模型中引入英文-中文分支进行知识蒸馏都能够有效提升模型性能. 在未来的工作中,我们将跨语言知识蒸馏扩展到其它多语种的视频字幕生成方法.

参 考 文 献

[1] Boroditsky L. Does language shape thought?: Mandarin and English speakers' conceptions of time. Cognitive Psychology, 2001, 43(1): 1-22

[2] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015

[3] Chen D L, Dolan W B. Collecting highly parallel data for paraphrase evaluation//Proceedings of the 49th Annual Meeting on Association for Computational Linguistics. Portland, USA, 2011: 190-200

[4] Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of actions. International Journal of Computer Vision, 2002, 50(2): 171-184

[5] Krishnamoorthy N, Malkarnenkar G, Mooney R J, et al. Generating natural-language video descriptions using text-mined knowledge//Proceedings of the 27th AAAI Conference on Artificial Intelligence. Washington, USA, 2013: 541-547

[6] Guadarrama S, Krishnamoorthy N, Malkarnenkar G, et al. YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 2712-2719

[7] Rohrbach M, Qiu W, Titov I, et al. Translating video content to natural language descriptions//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 433-440

[8] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and

description//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2625-2634

[9] Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to sequence — Video to text//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4534-4542

[10] Wang X, Chen W, Wu J, et al. Video captioning via hierarchical reinforcement learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4213-4222

[11] Pu Y, Min M R, Gan Z, et al. Adaptive feature abstraction for translating video to text//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 7284-7291

[12] Pan Y, Yao T, Li H, et al. Video captioning with transferred semantic attributes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 984-992

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

[14] Venugopalan S, Xu H, Donahue J, et al. Translating videos to natural language using deep recurrent neural networks//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies. Colorado, USA, 2015: 1494-1504

[15] Yao L, Torabi A, Cho K, et al. Describing videos by exploiting temporal structure//Proceedings of the IEEE International Conference on Computer Vision. Colorado, USA, 2015: 4507-4515

[16] Pan P, Xu Z, Yang Y, et al. Hierarchical recurrent neural encoder for video representation with application to captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1029-1038

[17] Baraldi L, Grana C, Cucchiara R, et al. Hierarchical boundary-aware neural encoder for video captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3185-3194

[18] Gan Z, Gan C, He X, et al. Semantic compositional networks for visual captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5630-5639

[19] Pan Y, Yao T, Li H, et al. Video captioning with transferred semantic attributes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6504-6512

[20] Yu Y, Choi J, Kim Y, et al. Supervising neural attention models for video captioning by human gaze data//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 490-498

[21] Chen Y, Wang S, Zhang W, et al. Less is more: Picking informative frames for video captioning//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 367-384

[22] Miyazaki T, Shimizu N. Cross-lingual image caption generation//Proceedings of the 54th Annual Meeting on Association for Computational Linguistics. Berlin, Germany, 2016: 1780-1790

[23] Xu J, Mei T, Yao T, et al. MSR-VTT: A large video description dataset for bridging video and language//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5288-5296

[24] Rohrbach A, Rohrbach M, Tandon N, et al. A dataset for movie description//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3202-3212

[25] Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 97-105

[26] Long M, Zhu H, Wang J, et al. Deep transfer learning with joint adaptation networks//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 2208-2217

[27] Dai W, Yang Q, Xue G R, et al. Boosting for transfer learning//Proceedings of the International Conference on Machine Learning. Montreal, Canada, 2007: 193-200

[28] Wan C, Pan R, Li J. Bi-weighting domain adaptation for cross-language text classification//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 1535-1540

[29] Liu X, Liu Z, Wang G, et al. Ensemble transfer learning algorithm. IEEE Access, 2017, 6: 2389-2396

[30] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1717-1724

[31] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks?//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 3320-3328

[32] Buciluǎ C, Caruana R, Niculescu-Mizil A. Model compression//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 535-541

[33] Sau B B, Balasubramanian V N. Deep model compression: Distilling knowledge from noisy teachers. arXiv preprint arXiv:1610.09650, 2016

[34] Kim J, Park S U, Kwak N. Paraphrasing complex network: Network compression via factor transfer//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2018: 2760-2769

[35] Lan X, Zhu X, Gong S. Knowledge distillation by on-the-fly native ensemble//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2018: 7517-7527

[36] Gupta S, Hoffman J, Malik J. Cross modal distillation for supervision transfer//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2827-2836

[37] Do T, Do T T, Tran H, et al. Compact trilinear interaction for visual question answering//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea 2019: 392-401

[38] Afouras T, Chung J S, Zisserman A. ASR is all you need: Cross-modal distillation for lip reading//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Virtual Barcelona, Spain, 2020: 2143-2147

[39] Perez A, Sanguineti V, Morerio P, et al. Audio-visual model distillation using acoustic images//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Snowmass, USA, 2020: 2854-2863

[40] Zhao P, Wang J, Xie L, et al. Privileged knowledge distillation for online action detection. arXiv preprint arXiv: 2011.09158, 2020

[41] Lopez-Paz D, Bottou L, Scholkopf B, et al. Unifying distillation and privileged information. arXiv preprint arXiv: 1511.03643, 2015

[42] Manning C D, Surdeanu M, Bauer J, et al. The Stanford CoreNLP natural language processing toolkit//Proceedings of the 52nd Annual Meeting on Association for Computational Linguistics. Baltimore, USA, 2014: 55-60

[43] Johnson J, Gupta A, Feifei L, et al. Image generation from scene graphs//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1219-1228

[44] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3156-3164

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

[46] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6077-6086

[47] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318

[48] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language//Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, USA, 2014: 376-380

[49] Lin C. ROUGE: A package for automatic evaluation of summaries//Proceedings of the ACL Workshop on Text Summarization Branches Out. Baltimore, USA, 2004: 74-81

[50] Vedantam R, Zitnick C L, Parikh D, et al. CIDEr: Consensus-based image description evaluation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4566-4575

[51] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[52] Wagner W, Steven B, Ewan K, et al. Natural language processing with python, analyzing text with the natural language toolkit. Language Resources and Evaluation, 2010, 44(4): 421-424

HOU Jing-Yi, Ph. D. Her research interests include computer vision and video content analysis.

QI Ya-Yun, M. S. candidate. Her research interests include computer vision and video content analysis.

WU Xin-Xiao, Ph. D., associate professor, Ph. D. supervisor. Her research interests include computer vision, image and video analysis, and understanding.

JIA Yun-De, Ph. D., professor, Ph. D. supervisor. His research interests include computer vision and cognitive computing, and intelligent systems.

Background

Video captioning is a challenging task that combines computer vision and natural language processing, which generating natural language sentences according to a given video. The video caption generation model not only needs to have the ability to understand the video content, but also needs to have sufficient language knowledge to generate faithful and smooth sentences. While video captioning has made a great process in recent years, most methods and datasets mainly focus on English video captioning. When it comes to video captioning tasks in other languages, it will be limited by the lack of corresponding data.

In this paper, we focus on Chinese video captioning and take full advantage of the easily accessible English captions as the privileged information to guide the generation of Chinese video captions. Our method learns cross-lingual knowledge contained in Chinese and English captions and utilizes knowledge distillation to enhance the quality of generated Chinese video captions. Benefit from the mechanism of knowledge distillation, our method only utilizes English captions data during the training stage, and after training it can directly generate Chinese captions from the input video. Moreover, we build a cross-lingual video captioning dataset MSVD-CN based on the English video captioning dataset MSVD by manually adding Chinese captions. In total, MSVD-CN contains 1970 video clips collected from the Internet and 11758 Chinese captions besides the original 41 English captions per video in MSVD. The experimental results on MSVD-CN show that our approach can effectively improve the performance of Chinese video captioning.

This work is supported by the National Nature Science Foundation of China (No. 62072041).