知识图谱引导下的视频描述生成

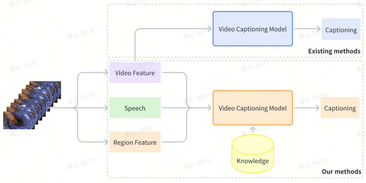
Text with Knowledge Graph Augmented Transformer for Video Captioning

Xin Gu[1,2], Guang Chen[3], Yufei Wang[3], Libo Zhang[1,2*], Tiejian Luo[1,2], Longyin Wen[3]

[1]智能软件研究中心，中国科学院软件研究所 [2]计算机学院，中国科学院大学 [3]字节跳动

会议: CVPR 2023 通讯作者: 张立波 联系方式: 18655882017 邮箱: libo@iscas.ac.cn

简 介



视频描述任务

定义 对于给定的视频，理解视频内容并生成自然语言描述。

挑战 (1) 视频中包含了复杂的语义关系和抽象概念，如何让模型 " 看懂 " 视频?
(2) 自然语言的语法结构和语义关系非常复杂，如何让模型学会 " 说话 " ?

现存问题 现有的视频描述方法通常存在长尾问题。即视频中的某些类别的事件、对象和场景出现频率较低，导致模型在描述这些 " 长尾 " 情况时表现较差。

贡献

我们提出了一种能够利用知识图谱的视频描述模型 TextKG，该模型能够自主学习知识图谱中的知识信息，并利用这些信息辅助视频描述的生成。TextKG 不仅能够生成更自然的描述文本，还能缓解长尾问题。

核心技术

构建知识图谱

通用知识图谱

包含与一般场景相关的通用知识信息。从现有的 ConceptNet 知识图谱中抽取而成。

特定知识图谱

包含与特定场景相关的专业知识信息。先从视频中提取文本，然后从文本中抽取知识。

双分支 Transformer 架构

外部分支 利用外部知识信息

内部分支 利用视频中的多模态信息

多模态特征融合

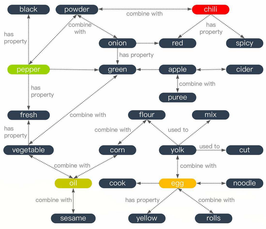
Early Fusion

将视频上下文特征、文本特征以及区域特征连接输入到模型中

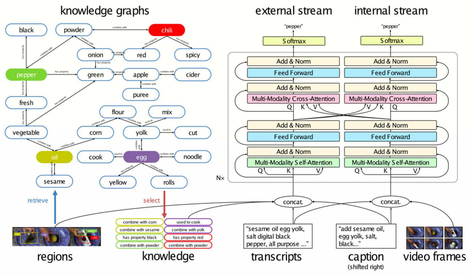
Intermediate Fusion

两个分支通过跨注意力模块交互，完成信息融合

Late Fusion 融合两个分支的输出，计算预测词概率向量



方 法



网络结构图

生成视频描述的流程

(1) 从知识图谱中检索与视频中目标物体相关的知识信息
(2) 根据与转录文本的语义相关性，筛选出与当前场景最相关的知识信息
(3) 将提取的多模态特征输入到两个分支中，这两个分支通过注意力模块交互
(4) 根据两个分支的输出概率向量预测单词

实例

输入:

输出:

枝藤上挂着红色的番茄

(1) 模型从视频中检测到 " 番茄 " 并从知识图谱中检索出 " 番茄是红色 "、" 番茄是甜的 "、" 番茄是一种食材 "、" 番茄长在枝藤上 " 等等知识信息；
(2) 通过计算这些信息与当前视频转录文本的相关性，保留 " 红色 " 和 " 长在枝藤上 " 等知识信息。
(3) 将提取到的多种视觉特征和知识信息输入到双分支网络中并生成描述 " 枝藤上挂着红色的番茄 "。

实 验

实验设置

1. 特征提取网络

(1) Faster-RCNN(ResNet-101) 模型提取区域特征
(2) Glove 模型提取文本特征
(3) Resnet-200 模型提取视频上下文特征
(4) C3D 模型提取视频动作特征

2.网络架构

2层Transformer模型

3.损失函数

交叉熵损失函数

$$\mathcal{L} = -\sum_{i=1}^{l}(\lambda_1 \log z_i^{\text{ext}} + \lambda_2 \log z_i^{\text{int}}),$$

和其他方法的比较

Method	B	M	R	C
OA-BTG [61]	41.4	28.2	-	46.9
POS-CG [51]	42	28.2	61.6	48.7
MGSA [4]	42.4	27.6	-	47.5
STG-KD [32]	40.5	28.3	60.9	47.1
ORG-TRL [64]	43.6	28.8	62.1	50.9
SGN [38]	40.8	28.3	60.8	49.5
MGRMP [5]	41.7	28.9	62.1	51.4
HMN [58]	43.5	29	**62.7**	51.5
TextKG	**43.7**	**29.6**	62.4	**52.4**

MSRVTT测试集

Method	B	M	R	C
OA-BTG [61]	36.9	36.2	-	90.6
POS-CG [51]	52.5	34.1	71.3	88.7
MGSA [4]	53.4	35	-	86.7
STG-KD [32]	52.2	36.9	73.9	93.0
ORG-TRL [64]	54.3	36.4	73.9	95.2
SGN [38]	52.8	35.5	72.9	94.3
MGRMP [5]	55.8	36.9	74.5	98.5
JCRR [14]	57.0	36.8	-	96.8
HMN [58]	59.2	37.7	**75.1**	104.0
TextKG	**60.8**	**38.5**	**75.1**	**105.2**

MSVD测试集

消融实验

V-F	R-F	Text	G-KG	S-KG	K-S	B	M	C	Rep
✓						7.4	15.7	32.1	4.1
✓	✓					9.5	17.7	45.9	5.2
✓	✓		✓		✓	9.7	17.8	48.9	4.2
✓	✓			✓	✓	9.7	18.0	48.5	3.3
✓	✓		✓	✓	✓	9.6	17.7	49.8	5.5
✓		✓				13.0	21.2	62.5	2.7
✓	✓	✓				13.9	**22.1**	71.3	2.9
✓	✓	✓	✓		✓	13.7	22.0	73.5	**2.0**
✓	✓	✓		✓	✓	13.5	21.9	74.8	2.1
✓	✓	✓	✓	✓		13.7	21.8	72.0	2.8
✓	✓	✓	✓	✓	✓	**14.0**	**22.1**	**75.9**	2.8

YouCookII 数据集

① "V-F" 和 "R-F" 代表视频特征和区域特征
② "Text" 代表语音转录特征
③ "G-KG" 和 "S-KG" 代表通用和特定知识图谱
④ "K-S" 代表知识选择机制

可视化结果

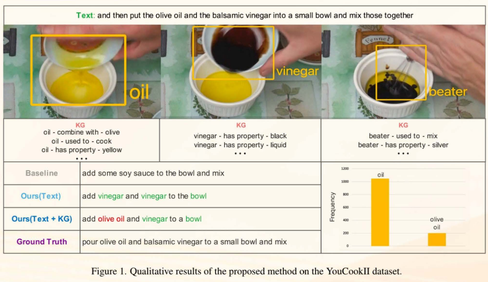


Figure 1. Qualitative results of the proposed method on the YouCookII dataset.

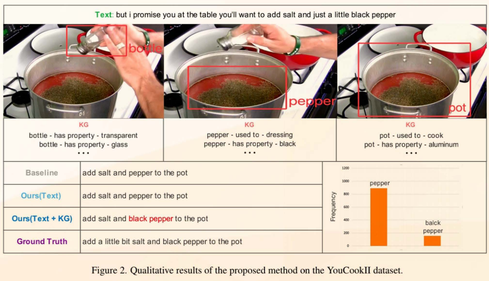


Figure 2. Qualitative results of the proposed method on the YouCookII dataset.

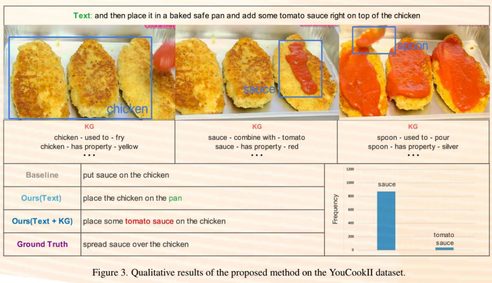


Figure 3. Qualitative results of the proposed method on the YouCookII dataset.

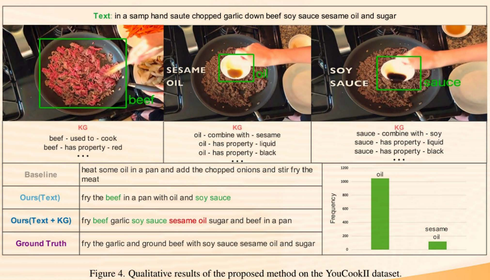


Figure 4. Qualitative results of the proposed method on the YouCookII dataset.