中国科学: 信息科学 2020 年 第 50 卷 第 6 期: 877–891
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
大数据分析专题 · 论文

基于长短时预测一致性的大规模视频语义识别算法

王铮[1,2], 翁泽佳[1,2], 王锐[1,2], 陈静静[1,2], 姜育刚[1,2*]

1. 复旦大学计算机学院, 上海 201203
2. 上海市智能信息处理重点实验室, 上海 200433
* 通信作者. E-mail: ygj@fudan.edu.cn

收稿日期: 2020–01–17; 接受日期: 2020–03–13; 网络出版日期: 2020–06–10

国家重点研发计划重点专项 (批准号: 2018YFB1004300) 资助项目

摘要 片段视频语义识别旨在识别视频中短小片段的语义概念, 是视频分析的一项重要任务. 由于片段视频的数量巨大且缺乏可参考的网络标签, 片段视频的标记十分困难, 通常只能对部分片段视频进行标记. 如何利用有限的语义标签提高片段视频语义识别的准确率是一项关键挑战. 因此本文提出了一种基于长短时预测一致性的视频语义识别算法. 该算法通过引入完整视频语义与片段视频语义一致性的约束, 对片段视频语义识别结果进行筛选, 以此提高片段视频语义识别的准确率. 本文提出的算法在大规模视频数据集 YouTube-8M 的片段视频语义识别任务上达到了 82.62% 的平均均值准确率 (mean average precision, MAP) 识别精度, 在第三届 YouTube-8M 比赛中排名第二.

关键词 大规模视频语义识别, 片段视频语义识别, 语义一致性, 特征聚合, 预测可靠性

1 引言

在信息技术高速发展的时代背景下, 大数据已经成为推动创新型国家建设的重要战略资源. 视频数据是大数据的主要组成成分, 占据了 80% 的互联网流量[1)]. 与其他数据相比, 视频数据具有体量大、内容丰富、标注稀疏等特点. 这也为视频数据分析带来了诸多挑战. 因此, 作为视频分析的重要任务, 视频语义识别一直以来都是计算机视觉领域和多媒体领域的研究热点.

近年来, 深度学习方法的运用大幅地推动了视频分析技术能力的提升. 其带来的直接影响是视频语义分析的结果越来越准确, 也间接带动了用于测评的视频数据规模的扩大. 这两方面的提升是相辅相成的. 大规模的视频数据集为训练更优的视频语义识别模型提供了充足的训练样本, 这对于依靠数据驱动的深度学习模型是十分必要的. 而越来越准确的视频语义分析结果反过来推动了视频数据集的

1) Cisco Study Reveals 80% of the World's Internet Traffic Will Be Video By 2019. 2020. https://www.purposefulfilms.com/cisco-study-reveals-80-of-the-worlds-internet-traffic-will-be-video-by-2019/.

引用格式: 王铮, 翁泽佳, 王锐, 等. 基于长短时预测一致性的大规模视频语义识别算法. 中国科学: 信息科学, 2020, 50: 877–891, doi: 10.1360/SSI-2020-0014
Wang Z, Weng Z J, Wang R, et al. Large-scale video semantic recognition based on consistency of segment-level and video-level predictions (in Chinese). Sci Sin Inform, 2020, 50: 877–891, doi: 10.1360/SSI-2020-0014

表 1 近十年发布的视频语义识别数据集列表

Table 1 A list of video semantic recognition datasets in recent ten years

Dataset	Year	#Videos	#Labels	Duration	Video-segment label
Hollywood2	2009	3669	12	Long	×
HMDB51	2011	7000	51	Short	×
UCF101	2012	13320	101	Short	×
Sports-1M	2014	1133158	487	Short	×
FCVID	2015	91233	239	Long	×
ActivityNet	2015	28000	203	Long	✓
YouTube-8M	2016	6100000	3,862	Long	✓
Charades	2016	9848	157	Short	×
Kinetics	2017	650000	700	Short	×
AVA	2017	1620000	80	Long	×
Something Something	2017	220847	174	Short	×
Moments in Time	2017	1000000	339	Short	×

扩容, 为深度学习模型从不同语义层面研究视频语义识别技术提供了可能. 从表 1 可以看到, 近十年发布的视频语义识别数据集在视频总量上和语义标签数量上总体呈上升趋势. 分析规模更大和语义更全面更细致的视频数据集, 要求设计出预测速度快且准确率高的视频语义识别算法.

为解决大规模视频数据的语义识别问题, 本文提出了基于长短时预测一致性的视频语义识别算法, 利用视频 (长时) 与其中的片段视频 (短时) 之间的语义一致性预测片段视频的语义类别, 从而提升片段视频预测的准确性和可靠性. 本文所提出的方法在目前最大规模的视频语义识别数据集 YouTube-8M 上进行了算法验证, 实验结果表明本文提出的算法是有效的.

本文剩余部分组织如下: 第 2 节介绍现有的视频分析数据集和深度视频分析模型, 第 3 节介绍所提出的基于长短时预测一致性的视频语义识别模型, 第 4 节给出实验设置和结果分析, 最后第 5 节总结.

2 视频语义分析技术简述

2.1 视频语义分析数据集

在过去的十年间, 已发布的视频语义分析数据集有数十个之多, 本文选择了其中具有代表性的 12 个数据集列在表 1 中. 这些数据集作为视频语义分析的基准数据集被广泛地使用, 大规模视频数据集极大地推动了视频语义识别准确率的提升, 同时为算法更全面地探索视频语义提供了丰富的数据.

Hollywood2 [1] 从 69 个电影中抽取了约 20 个小时的片段视频. 除了 12 个动作语义类别外, 还有 10 个场景类别的标注. 额外的场景标注允许模型探索动作语义类别和场景类别之间的相关性, 以在动作语义不明确时依靠场景信息进行推断.

HMDB51 [2] 包含 51 个人体动作类别, 包括脸部动作、嘴部动作、身体动作、身体与物体交互、人与人交互等 5 个大类. 同时, 该数据集概括性地描述了相机位置、身体关节、人的数量等.

UCF101 [3] 中的语义类别相比较于前两个数据集更为丰富, 共有 101 个语义类别, 近 27 个小时的视频, 包括身体动作、身体与物体交互、人与人交互、演奏乐器、体育等 5 个大类.

Sports-1M [4] 是第 1 个大规模视频语义数据集, 共有 487 类, 超一百万个 YouTube 上收集的体育类视频. 数据标签由 YouTube 话题应用接口自动获取, 未经过人工筛选. 大规模视频数据集的标签是由机器根据视频的描述、标注等自动生成的, 因而是有噪声的.

FCVID [5] 是国内发布的第 1 个大规模网络视频语义数据集, 共有 239 类, 9 万余个视频. 包括社会活动, 教程视频、物体、场景等 11 个大类. 视频总时长 4232 小时.

ActivityNet [6] 是一个多标签的视频语义数据集, 平均每个视频有 1.4 个标签. 共有 203 个类, 2.8 万个视频. 该数据集的标签是一个多层次化的结构.

YouTube-8M [7] 是目前视频数量最多的、标签种类最丰富的视频语义数据集. 共有 24 个大类, 3862 类, 610 余万个高质量的视频, 平均每个视频有 3 个语义标签. 主要包括活动、物体、场景、事件等多种类视频. 该数据集是本文主要采用的实验数据集.

Charades [8] 的独特之处在于其内容都是采集于日常的生活场景. 包含 157 类活动语义标签, 46 个物体标签, 15 个场景标签, 平均每个视频有 6.8 个动作标签, 大大超过 ActivityNet 平均 1.4 个的数量. 在此基础上发展出来的 Charades-Ego [9] 增加了一个主观视角.

Kinetics [10] 是目前经人工校验的最大的视频语义分析数据集. 该数据集包含 700 个类别, 所有标签均经过人工校验. 每个类别至少包括 600 个的视频, 所有视频的时长均为 10 s.

AVA [11] 是一个分析人类活动的视频语义数据集. 包含 80 个基本动作类别, 对 430 个 15 min 时长的视频以 3 s 为间隔做了密集标注, 在每个片段视频的中间一帧标注了人的边界框.

Something Something [12] 重点关注人与物体交互的动作, 特别是手部的交互动作. 除了标注动作类别外, 该数据集还标注了交互的对象名称.

Moments in Time [13] 收集了一百万个时长为 3 s 的视频分为 339 个类别. 该数据集最大的特点是类内样本差异较大, 体现了自然语言处理中一词多义的概念. 在其基础上衍生的 Multi-Moments in Time [14] 对每个视频标注了多个类别.

2.2 深度视频语义识别模型

近年来, 利用深度神经网络进行特征学习的技术受到了极大的关注. 深度学习在视频语义识别中的成功可以归功于利用了大量的标注数据在有千万级别参数量的深度神经网络模型上进行训练, 使得网络能够捕捉到视频中语义类别内的相似性和语义类别间的差异性.

2.2.1 循环网络

视频是由连续多个图像组成的视频帧序列, 一个直接的特征表示方法是, 先利用预训练的图像分类模型提取每一个视频帧的特征向量以捕捉空间信息, 再利用 LSTM [15] 之类的循环神经网络整合时序信息. Donahue 等 [16] 利用二维卷积神经网络产生视频帧特征并用 LSTM 建模时间序列, LSTM 的循环特性允许输入帧长度任意的视频来识别视频内容. 该网络结构同样也能用于生成文本描述. Wu 等 [17] 使用了两路 LSTM 编码视频特征序列, 并提取了视频的光流特征, 同样使用二维卷积神经网络和 LSTM 的组合方式建模光流信息. Ng 等 [18] 比较了多种简单的时序建模的方法和 LSTM 建模的方式, 并指出由于 LSTM 能够建模视频帧之间的先后顺序关系, 其准确率高于其他多种简单的时序建模方法.

2.2.2 融合网络

双流网络 [19] 的提出是为了融合静态表观特征和运动特征两种不同模态的特征. 在其基础上, 文

献 [20] 提出使用双流网络产生多尺度的特征图, 将特征图与运动轨迹合并在一起以计算以轨迹为中心的神经网络的响应, 该方法将深层特征转变为受运动轨迹约束的特征描述子. TSN [21] 提出将视频分成 3 个时间阶段, 每个阶段的视频分别提取静态表观特征和运动特征, 并将不同阶段的特征按模态组合在一起用于预测动作类别, 该方法能够建模更长的视频. 人体姿态特征是运动特征的一个特例, 该特征受镜头移动影响较小. PA3D [22] 用热力图表示视频中人体的多种姿态表示, 包括关节点、身体局部和全图特征, 多种特征融合后进行预测, 但该方法只适合有清晰人像的视频.

2.2.3 时空网络

时空网络是一种三维卷积神经网络, 在二维卷积操作的基础上增加一个时间维的卷积操作. 最早使用三维卷积识别动作的方法 [23] 用人工设计的 5 种三维卷积核分别建模灰度特征, 水平和垂直方向的图像梯度特征和光流特征. C3D [24] 是首个用大规模视频数据 Sport-1M 训练的时空网络, 使用了 5 个三维卷积层, 5 个最大池化层, 2 个全连接层和一个 softmax 损失函数层. C3D 在包括动作识别和物体识别的多个视频分析基准上都有非常出色的表现. 为了进一步提升三维网络的泛化能力, I3D [25] 提出将用于图像识别的二维深度神经网络膨胀成三维的时空网络. 通过在时间维度上重复二维卷积核, 膨胀后的三维时空网络可以复用 ImageNet 预训练的二维卷积核参数. I3D 的另一项发现是用 Kinetics 预训练的时空模型, 在 UCF101 和 HMDB51 上达到了更好的识别准确率. P3D [26] 从减少模型参数量的角度出发, 用一组 $1\times3\times3$ 的空间卷积核和 $3\times1\times1$ 的时间卷积核替代了 C3D 中 $3\times3\times3$ 的时空卷积核, 同时设计了 3 种时间卷积核和空间卷积核的组合方式. Non-local 操作 [27] 是一种可在现有网络结构中拓展的信息交互层, 每个时空位置的特征都会与其他位置的特征两两交互, 将其他位置的有效信息融入当前位置的特征中. 二维卷积操作的计算代价低但不能捕捉时序信息, 而三维卷积操作有很好的性能但计算代价很高, TSM [28] 提出的一种全新的视频识别结构, 通过在时间维度上平移部分特征图建模视频的时序变化, 卷积核仍采用二维卷积核. TSM 相比较于三维卷积网络有 3 个优势: 更快的测试速度, 更低的内存消耗和更多的时间尺度融合.

尽管时空卷积网络能在视频识别与分类上取得最优的效果, 但是该类模型往往由于网络结构复杂导致运算效率低等问题. 因此, 为了平衡效率和精度, 本文主要使用了循环网络和融合网络用于提取视频特征, 从而保证了单个网络的计算效率较低. 此外, 为了进一步提高识别准确率, 本文还进行了多模型识别结果融合, 以此提高识别准确率.

3 基于长短时预测一致性的视频语义识别模型

本节将详细介绍本文所提出的视频语义识别模型. 本文首先用视频预训练视频语义识别模型; 再利用片段视频对语义识别模型进行微调. 为了能够有效利用片段视频, 本文设计了一个针对片段视频的损失函数; 最后提出了长短时预测一致性的片段视频语义过滤算法. 模型示意图如图 1 所示.

3.1 视频特征表示

本文采用的 YouTube-8M 视频数据集的数据量非常大, 采用高效的视频帧特征聚合算法显得非常重要. 本小节将介绍视频帧的视觉特征聚合方式, 片段视频对应的音频特征采用与视觉特征一致的聚合方式. 图 2 展示本文的视频语义识别模型. 首先将视频解码成视频帧序列, 各个视频帧分别提取二维视觉特征和音频特征, 本文采用的视频帧特征提取网络是 Inception-V3 网络 [29], 音频特征提取网络采用的是, 减少全连接层特征维度和增加批量正则化的 VGG 网络 [30]. 特征聚合可使用不同的可学习

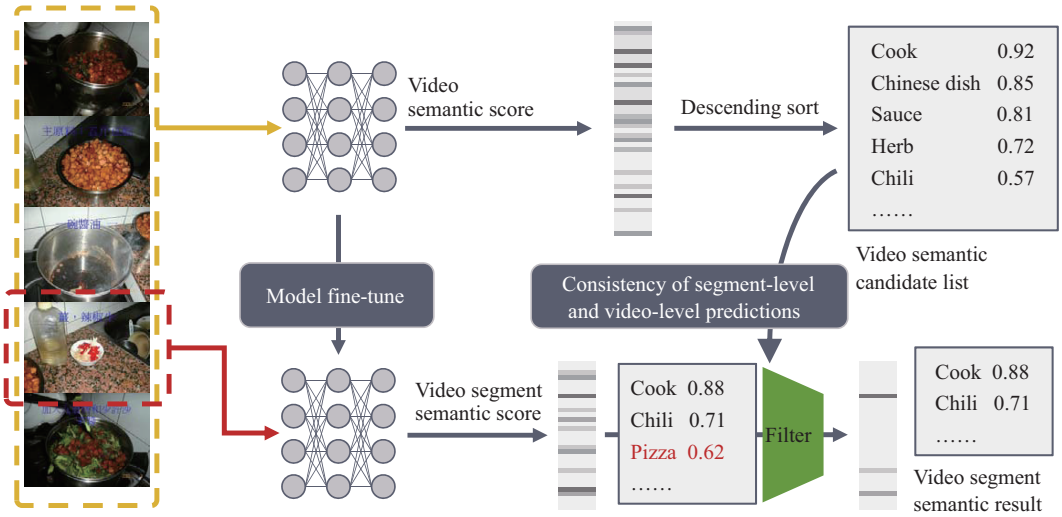


图 1 (网络版彩图) 基于长短时预测一致性的视频语义识别算法示意图

Figure 1 (Color online) The diagram of consistency of segment-level and video-level predictions

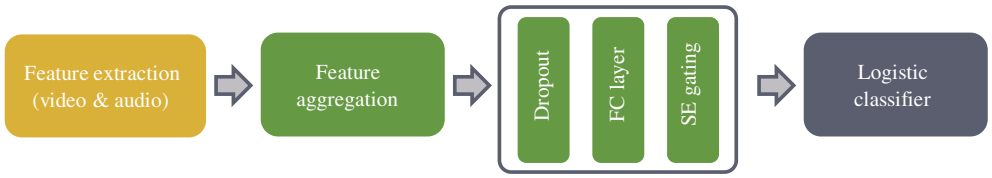


图 2 (网络版彩图) 视频语义识别模型

Figure 2 (Color online) The structure of video semantic recognition models

池化算法来实现, 将在下文详细介绍. 本文在特征聚合模块后连接了一个丢弃正则化项 (dropout)、一个全连接层和一个 SE 门控单元 [31], 最后连接到一个逻辑函数分类器完成语义预测. 以下具体介绍若干特征聚合算法的关键结构. 这些算法的目的是将多个视频帧特征向量 x_i 聚合为一个特征向量.

Vector of locally aggregated descriptors (VLAD) 最初是一种用于实例级检索 [32] 和图像分类 [33] 的局部特征池化算法, 捕获特征向量相对于局部聚类特征的变化:

$$\mathrm{VLAD}(j,k)=\sum_{i=1}^{N} a_k(\boldsymbol{x}_i)(\boldsymbol{x}_i(j)-\boldsymbol{c}_k(j)), \tag{1}$$

其中 $a_k(\boldsymbol{x}_i)$ 是将 $\boldsymbol{x}_i$ 分配给第 k 个聚类中心的指示函数, $\boldsymbol{c}_k(j)$ 是第 j 个聚类中心的特征向量. 由于分配聚类中心的操作是不可导的, VLAD 很难直接应用于神经网络. 为解决这个问题, Arandjelovic 等 [34] 提出了 NetVLAD, 使用软分配的方式构建局部特征描述子, 将式 (1) 中的 $a_k(\boldsymbol{x}_i)$ 修改为可导的 $\overline{a}_k(\boldsymbol{x}_i)$:

$$\overline{a}_k(\boldsymbol{x}_i)=\frac{\mathrm{e}^{\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{x}_i+b_k}}{\sum_{k'}\mathrm{e}^{\boldsymbol{W}_{k'}^{\mathrm{T}}\boldsymbol{x}_i+b_k}}. \tag{2}$$

本文采用了 NetVLAD 的变体, 包括 NeXtVLAD [35] 和 Non-local NetVLAD [36]. NetVLAD 为输出与输入向量相同维度的向量, 添加了一个全连接层用于降维, 计算量大且非常耗时. NeXtVLAD 受 ResNeXt 启发, 将输入特征分为 G 个组, 再分别计算聚类中心, 每个输出向量维度都变小了 G 倍, 大大降低了

全连接层的计算量实现了计算的加速, 最后多组的输出向量加和在一起. Non-local NetVLAD 在多个聚类中心之间交换了信息, 在视频和音频特征融合的不同阶段, 分别使用了 Early-fused Non-local NetVLAD (EFNL-NetVLAD) [36] 和 Late-fused Non-local NetVLAD (LFNL-NetVLAD) [36] 两种变体. 本文用 NeXtVLAD 替换了文献 [36] 中的 NetVLAD, 组合成 EFNL-NeXtVLAD 和 LFNL-NeXtVLAD.

Bag-of-Words (BOW) [37, 38] 和 Fisher Vector (FV) [39, 40] 是两种传统的聚合算法. 受软分配思想的启发, 将传统的 Bag-of-Frames (BOF) 和 FV 的硬分配修改为可学习的软分配, 分别构建 SoftDBOF [7] 和 NetFV [41]:

$$\text{SoftDBOF}(k) = \sum_{i=1}^{n} a_k(\boldsymbol{x}_i), \tag{3}$$

$$\text{NetFV}(j,k) = \sum_{i=1}^{N} a_k(\boldsymbol{x}_i)\left(\frac{\boldsymbol{x}_i(j) - \boldsymbol{c}_k(j)}{\sigma_k(j)}\right). \tag{4}$$

本文在 SoftDBOF 的输出上增加了一个门控单元 ($\boldsymbol{W}\boldsymbol{x} + b$) 得到 GatedDBOF.

Gated recurrent unit (GRU) [42] 和 LSTM [15] 一样, 可以自然地将向量序列聚合为一个向量:

$$\begin{aligned} r_i &= \sigma(\boldsymbol{W}_r \cdot [\boldsymbol{h}_{i-1}, \boldsymbol{x}_i]),\\ z_i &= \sigma(\boldsymbol{W}_z \cdot [\boldsymbol{h}_{i-1}, \boldsymbol{x}_i]),\\ \widetilde{\boldsymbol{h}}_i &= \tanh(\boldsymbol{W}_{\tilde{\boldsymbol{h}}} \cdot [r_t \boldsymbol{h}_{i-1}, \boldsymbol{x}_i]),\\ \boldsymbol{h}_i &= (1 - z_i)\boldsymbol{h}_{i-1} + z_i \tilde{\boldsymbol{h}}_i,\\ \text{GRU}(\boldsymbol{x}_i) &= \sigma(\boldsymbol{W}_o \cdot h_i). \end{aligned} \tag{5}$$

相比较于 LSTM, GRU 优势是在同等效果下更容易进行训练, 在很大程度上能够提高训练的效率, 因此 GRU 也作为本文的视频帧特征聚合算法之一.

ResNetLike [43] 的残差连接缩短了梯度传递路径, 深度网络能得到更充分的训练. 由于片段长度短且采样率低 (1 帧/秒), 本文还采用了平均池化和最大池化的方法来聚合多个特征向量. 本文采用 ResNetLike 网络结构及其两种简单变体来融合平均池化的视频帧特征和音频特征: 通过直接输入视频和音频的平均池化特征得到 ResNetLike 特征, 通过增加 ResNetLike 恒等块中视觉特征部分的通道数量得到 ResNetLike-Identity 特征, 通过拼接平均池化特征和最大池化特征作为输入得到 ResNetLike-Max 特征.

Squeeze and excitation gating (SE Gating) [31] 提取了特征的全局平均池化和由两个全连接层组成的门控权重向量. 输出特征是门控权重向量对原特征的加权输出.

3.2 单模型的混合结构表示

本文采用混合结构 [35] 加强单个模型的表达能力. 混合结构结合了知识蒸馏的思想, 并且还使用 KL 散度构造了一个额外的正则化项. 即时 (on-the-fly) 蒸馏避免了重新训练模型以进行知识蒸馏学习的麻烦. 本文将混合结构应用于 NeXtVLAD, EarlyNetVLAD, LightNetVLAD, GatedDBOF, SoftDBOF, NetFV, GRU 和 ResNetLike 的 3 种实现中.

3.3 长短时一致预测的模型训练方法

给定一个视频 Video_i, 片段视频 $\text{Segment}_{i,j}$ 是视频中第 j 个连续多帧序列, 例如连续 5 s 长度的片段视频. 片段视频语义预测任务预测 $\text{Segment}_{i,j}$ 的多个语义类别 C_s. 预测任务的标签是二值标签, 即

$\text{prediction}^s_{i,j} \in \{0,1\}$, 其中, 正例标签 1 指 $\text{Segment}_{i,j}$ 包含语义类别 C_s, 反之负例标签 0 指 $\text{Segment}_{i,j}$ 与语义类别 C_s 无关. 对于一个大规模的视频数据集来说, 要标注所有片段视频与所有语义类别的相关性是不现实的, 通常来说 Video_i 中大部分的 $\text{Segment}_{i,j}$ 的大部分的标注是空缺的. 视频中的片段可以被视为片段视频, 因此本文对视频训练的模型进行微调来识别片段视频. 对于从视频中提取的某个特定片段 $\text{Segment}_{i,j}$, C_s 是片段视频类别的集合, $Y^s_{i,j}$ 是 $\text{Segment}_{i,j}$ 被标记的片段视频类别, 而 Y^v_i 是被截取的视频的类别标签. 如下所示, 交叉熵函数 (cross entropy, CE) 仅涉及被标记的片段视频语义类别标签, 而与其他未标记的片段视频语义类别无关:

$$\text{CE}(Y^s_{i,j}) = -\sum_{y^s_{i,j} \in Y^s_{i,j}} (y^s_{i,j} \log(p) + (1 - y^s_{i,j}) \log(1-p)), \tag{6}$$

其中 y_s 是类别标签, p 是预测类别的概率.

实验结果表明上述损失函数最终导致片段视频所有语义类别的平均分数偏高, 由于只有被标注的负例类别得到了训练而更多未被标记的类别没有得到训练, 未标记类别的分数不受约束. 为解决该问题, 本文考虑用弱监督的方式来约束未标记类别的分数. 本文为片段视频语义识别模型设计了一项负例交叉熵损失函数用于训练片段视频中的未标记类别.

由于视频的正例类别的平均数量约为 3, 因此 $C_s - Y^s_{i,j} - Y^v_i$ 中的大多数类别都可以假设为负例类别. 本文对所有 $C_s - Y^s_{i,j} - Y^v_i$ 中的类别预测为负例类别, 以此近似一项负例交叉熵损失:

$$\text{CE}(C_s - Y^s_{i,j} - Y^v_i) = \frac{\Sigma_{\hat{y} \in (C_s - Y^s_{i,j} - Y^v_i)} \text{CE}(\hat{y})}{|C_s - Y^s_{i,j} - Y^v_i|}, \tag{7}$$

设定所有 $\hat{y} = 0$.

整体的损失函数计算如下:

$$\text{Loss} = \text{CE}(Y^s_{i,j}) + \alpha \text{CE}(C_s - Y^s_{i,j} - Y^v_i), \tag{8}$$

其中 α 是一个多项平衡系数, 此处简单地设置为 1.

3.4 长短时预测一致的模型测试方法

基于长短时预测一致性的预测算法如图 1 所示, 其基本思想是: 如果对一个介绍 “豆干制造过程” 视频的预测没有出现披萨的语义概念, 那么对该视频中截取的片段视频的语义预测也不会出现披萨的语义概念. 考虑以下事实: 如果一个语义实体出现在视频中的分数很小, 那么这个语义实体大概率不会出现在该视频的任何短时片段中. 基于上述长短时语义一致性的观察, 本文提出了一种长短时预测一致的模型预测算法. 该算法首先根据模型的预测结果, 建立一个视频语义类别的候选列表, 列表中的语义实体按预测分数大小排列. 假设视频的语义类别包含了所有的片段视频的语义类别. 在片段视频的预测结果上加一个掩模, 用候选列表筛选掉无关的语义类别, 只保留存在于候选列表中的语义类别. 这意味着如果片段视频的预测结果不存在于视频的候选列表中, 则将该项预测类别的概率置 0. 通过视频和片段视频在语义类别层面上的预测一致性, 不仅可以减少片段视频的假阳性预测, 还可以大幅减少后续分析片段视频预测结果的时间开销. 以下介绍 3 种候选列表的产生方法, 分别是分数排位约束、分数阈值约束和类别预测数量约束. 预测算法的具体步骤如算法 1 所示.

分数排名约束. 通过简单的统计, 本文发现采用视频预测结果中分数最高的前 20 个语义类别作为候选列表, 即可达到较高的片段视频预测标签召回率. 平均来看, 一个视频的分数最高的前 20 个预

算法 1 长短时预测一致的模型预测算法

输入: 视频 $\text{Video}_{i=0}^{I}$, 视频语义类别 C_v, 片段视频语义类别 C_s, 候选列表生成算法 algo;

for all Video_i **do**
  视频 Video_i 分段成多个片段视频 $\text{Segment}_{i,j=0}^{J}$;
  $\text{prediction}_i^v = \text{VideoPredictionModels}(\text{Video}_i)$;
  for all Segment_{ij} **do**
    $\text{prediction}_{ij}^s = \text{SegmentPredictionModels}(\text{Segment}_{ij})$;
  end for
end for
GenerateCandidates(algo**)**
for each $\text{Segment}_{i,j}$ **do**
  $\text{candidate}_{ij}^s = \{\text{prediction}_{ij}^s \cap \text{candidate}_i^v\}$;
  根据 candidate_{ij}^s 的分数由高到低预测 Segment_{ij} 包含语义类别 $c_s \in C_s$;
end for
计算所有类别 $c_s \in C_s$ 的均值准确率 (average precision, AP);
计算片段视频语义类别 C_s 的平均均值准确率 (mean average precision, MAP);

function GenerateCandidates(algo**)**
for each Video_i **do**
  if algo== 分数排名约束 **then**
    for each Video_i **do**
      $\text{candidate}_i = \{\text{prediction}_i^v | \text{prediction}_i^v.\text{score} \geqslant \text{sorted}(\text{prediction}_i^v.\text{score})[k-1]\}$ (k 是超参数);
    end for
  end if
  if algo== 分数阈值约束 **then**
    $\text{candidate}_i = \{\text{prediction}_i^v | \text{prediction}_i^v.\text{score} \geqslant \text{threshold}\}$ (threshold 是超参数);
  end if
  if algo== 类别预测数量约束 **then**
    设定最多有 m 个视频的预测结果为视频语义类别 $c_v \in C_v$, 若超出 M 则移除分数最小的预测结果 (m 是超参数);
    根据 prediction_i^v 的分数由高到低预测 Video_i 包含语义类别 $c_v \in C_v$ 并为该类别记录下 Video_i 的分数;
    更新所有类别 $c_v \in C_v$ 的预测结果中最小的分数 thresholds_{c_v};
    $\text{candidate}_i^v = \{\text{prediction}_i^v | \text{prediction}_i^v.\text{score} \geqslant \text{thresholds}_{c_v}\}$;
  end if
end for
end function

测类别可以覆盖其对应的 97% 的片段视频语义类别, 分数最高的前 100 个预测类别可以覆盖 99% 的片段视频语义类别. 要特别指出的是, 本文实验中视频的分数最高的前 20 个预测类别是从视频所有的 3862 个语义类别的预测中排序选出的, 而不仅仅是片段视频中的分数最高的前 1000 个语义类别. 因此, 本文的第一个想法是选择视频的分数最高的前 K 个预测结果来产生候选列表, 以此约束片段视频的候选类别. 同一个视频中包含的多个片段视频, 共享同一个视频语义候选列表.

分数阈值约束. 分数排名约束的算法忽略了一个问题: 一些视频包含有丰富多样的语义类别, 但同样存在一些视频的内容相比较而言更单一, 场景转换也更少, 其包含的语义类别数量相对较少. 而分数排名约束的算法忽略了这种视频之间类别数量多样性的差异. 因此, 本文提出的第 2 种候选列表产生办法是用视频的语义预测分数产生候选列表. 本文设置一个阈值, 如果视频在某项语义类别上的分数小于某个阈值, 则该项语义实体将不被加入语义候选列表, 而对片段视频进行预测时将不再考虑

该类别.

类别预测数量约束. 类别预测数量约束是限制每个语义类别可预测的视频数量. 由于不同的语义实体识别难度不一样, 不同语义类别的分数分布也不同, 直接根据一个全局的分数阈值产生候选列表的算法仍较为粗暴. 分数阈值约束的算法虽然规避了语义类别数量多样性的问题, 但未能解决分数本身可能存在的分布不一致的问题. 本文提出, 针对一定数量的测试数据, 限制每个语义类别能够预测的视频数量, 这个算法未对分数进行限制, 保证了不同的类别的样本总体的分数是灵活可变动的. 每个视频的候选列表由其存在于所有类别预测结果中的实体得到.

4 实验设置与分析

4.1 实验评价指标及数据集

评价指标是 100000 个片段视频样本的 MAP:

$$\text{MAP@100000} = \frac{1}{C_s}\sum_{c_s=1}^{C_s}\frac{\sum_{k=1}^{n}P(k)\times \text{rel}(k)}{N_{c_s}}, \tag{9}$$

其中 C_s 是类别数, $P(k)$ 是截止 k 时的精度, n 是每个类别预测的片段视频数, $\text{rel}(k)$ 是指示函数, 如果对第 k 个片段视频的预测是相关 (正确) 类别, 则 $\text{rel}(k)$ 为 1, 否则为 0. N_c 是每个类别的带正标签的片段视频数. 虽然本文应该在测试过程中预测测试视频中包含的所有片段视频, 但是在计算 MAP 时只使用有人工标注的片段, 而其他未明确标注的短时片段则在计算 MAP 之前从预测列表中删除.

本文采用 YouTube-8M 数据集进行算法验证. 该数据集有 610 余万个高质量的视频, 包含 3862 个视频语义类别, 数据集中的每个视频包含多个语义类别, 平均每个视频有 3 个语义类别, 所有的视频标签均是由机器根据网络标签自动生成的, 因此存在一定比例的标签噪声. 数据集有超过 23 万个片段视频标签, 这些标签经过人工验证, 共有 1000 个类, 每个视频截取 5 个片段视频. YouTube-8M 数据集提供 Inception 网络 [29] 提取的近 2 百万个视频帧的视觉特征, 以及相应的音频特征, 还单独为每个视频提供一个平均池化后的特征. 在测试中, 本文使用了 YouTube-8M 的官方保密测试数据集, 所有的实验结果均是由官方测试服务器计算的.

4.2 实验分析

基本模型. 用 YouTube-8M 中的视频数据训练的模型, 并将混合结构表示应用于几个单模型, 最终构建了 10 个混合结构表示的单模型, 如表 2 第 1 列所示.

首先测试的是视频标签预测. 如表 2 第 2 列所示, 大多混合结构表示的模型都能达到较高的精度, 对这些模型输出分数做简单的加权平均, 视频的 global average precision (GAP) [44] 可以达到 0.88932. 由于视频的类别包含了所有片段视频的类别, 本文将这些视频预训练的模型输出层加上一个掩模, 只保留片段视频的类别, 用于对片段视频的预测, 结果如表 2 第 3 列所示. 其中, Mix-SoftDBOF 模型实现了最好的 MAP, 其次是 Mix-GatedDBOF 和 Mix-NeXtVLAD 模型.

模型微调. 用 YouTube-8M 的片段视频数据集微调基本模型. 本文随机抽取片段视频数据集的 5/6 作为训练集, 其余作为验证集. 如表 2 第 4 列所示, 相比较于第 3 列的基本模型, 模型微调大幅提高模型的 MAP. 原基本模型的最佳模型: Mix-SoftDBOF 的 MAP 提升近 6%, 其他模型的 MAP 也有提升. 本文认为, MAP 得到大幅度提升的原因如下: 首先, 视频的标签是机器自动标注的, 包含的噪

表 2　多种片段视频语义预测模型的预测结果

Table 2　Segment MAP comparison for evaluating our approaches

Model	Video GAP	Video-segment MAP			
	–	–	Fine-tune	Fine-tune	Fine-tune
	–	–	–	Consistency	Consistency
	–	–	–	–	Using all data
Mix-NeXtVLAD*	**0.88433**	0.7373	0.78638	0.81127	**0.81548**
Mix-EFNL-NeXtVLAD*	0.88288	0.65238	0.77147	0.80857	0.81212
Mix-LFNL-NeXtVLAD	0.88142	0.69911	0.77579	0.80625	0.8093
Mix-SoftDBOF*	0.88071	**0.74305**	**0.80582**	**0.81237**	0.81421
Mix-GatedDBOF*	0.8802	0.73679	0.79963	0.81122	0.81327
Mix-NetFV	0.88251	0.73049	0.77949	0.80967	0.81235
Mix-GRU	0.87659	0.68541	0.77332	0.80436	0.8058
Mix-ResNetLike*	0.86499	0.71616	0.7835	0.8061	0.80928
Mix-ResNetLike-Identity*	0.86284	0.71958	0.78614	0.80796	0.81034
Mix-ResNetLike-Max*	0.86288	0.72541	0.78558	0.80825	0.81100
Essemble models (*)	**0.88932**	–	–	–	**0.8262**

表 3　多种一致性约束算法的结果

Table 3　Results of different consistency constraint algorithms

Constraint algorithm	Video-segment MAP
No constraint	0.80419
Score ranking constraint: top 100	0.82250
Score threshold constraint: >25E−5	0.82326
Prediction numbers per class constraint: 2000	**0.82620**

声比人工标注的片段视频标签更多. 其次, 视频的语义和片段视频的语义之间存在差异, 微调过程相当于迁移学习. 因此, 模型微调是很有意义的, 这使本文的模型更适合于分析片段视频的语义.

一致性预测. 在这一部分中, 本文将通过实验验证一致性预测算法的有效性. 本文使用了分数阈值约束的算法来生成候选标签列表, 并且选择了 Mix-NeXtVLAD 模型为所有视频片段生成候选标签列表, 所有预测分数大于 0.00025 的标签被添加到候选列表中. 在一致性预测过程中, 本文提出的算法将删除不在候选列表中的预测标签. 实验结果列于表 2 的第 5 列, MAP 的大幅提升证明了一致性预测算法的有效性. 基于一致性预测算法是有效的, 本文对多种一致性预测算法进行了验证, 表 3 说明利用类别预测数量约束的一致性预测方法取得了最大的 MAP.

模型选择和数据量效应. 训练模型使用全部数据可以提高最终的 MAP, 但是更多的训练数据同时意味着更少的用于选择最有模型的验证数据. 在本文的实验中, 首先随机采样了 5/6 的短时视频数据集作为训练集, 其余 1/6 的数据用作模型选择的验证集. 通过验证集上的 MAP, 可以找到训练时 MAP 最优的训练时长区间. 然后将所有数据用于模型训练, 并使用先前记录的时长区间来估算全部训练数据所需的训练时长. 本文使用随机权重平均 (stochastic weight averaging, SWA) [45] 将这些模型组合为一个模型. 一方面, 训练时长区间的估计提高了模型选择的宽容度, 另一方面, SWA 操作提高了模型的鲁棒性并可以获得更高的 MAP. 如表 2 最后一列所示, 通过增加可训练的数据量, 所有模型达到了

表 4 YouTube-8M 比赛前 5 名队伍的结果

Table 4 Results of top 5 teams in the YouTube-8M challenge

Team	Video-segment MAP
Layer6 AI	0.83292
BigVid (our team)	**0.82620**
RLin	0.82551
Bestfitting	0.81707
Last Top GB Model	0.80459

最好的性能.

模型组合. 本文选择了 Mix-NeXtVLAD, Mix-GatedDBOF, Mix-SoftDBOF, Mix-EarlyNet-VLAD 和 3 种 Mix-ResNetLike 模型变体用于模型组合. 根据文献 [41], 当组合的模型结构较为相似时, 该组合不会为最终模型带来太大的提升, 因此每个 Mix-ResNetLike 模型的权重都降低至原来的 1/3. 更多的模型组合和更优的权重分配可以为最终的 MAP 分数带来进一步的提升.

本文模型算法在第三届 YouTube-8M 比赛中排名第 2, 比赛结果如表 4 所示. 比赛第 1 名 Layer6 AI [46] 将片段视频的多标签预测问题转换为对于每个语义类别的二元预测, 并在分类器中增加了类别标签特征. 第 3 名 RLin [47] 同样用了筛选算法, 但其只使用多种 NeXtVLAD 的多种变体来预测, 并且只用了分数排名约束的候选列表筛选算法. 第 4 名 bestfitting [48] 融合了多种深度模型但未提出类别筛选算法. 第 5 名 Last Top GB Model 使用 LSTM 逐片段预测的算法, 利用上下文信息预测片段视频的语义类别.

图 3 展示了本文的算法在 3 个视频上的预测结果. 篇幅所限, 在此只展示若干片段视频的关键帧, 事实上这些视频有长有短. 图中展示了视频和片段视频的预测结果, 分别列于虚线框和实线框内. 本文只展示了每个预测结果的前十位, 事实上每个视频保留了前一百位的预测结果作为候选类别. 线框内打勾的类别是前十位的预测结果中预测正确的类别, 划去的类别是利用视频的预测结果筛选后去掉的类别, 换句话说, 被划去的类别没有出现在相应的视频的前一百位预测结果中. 单独看视频和片段视频的预测结果可以发现, 本文训练的模型在预测视频的语义类别时会给出更确定的判断, 只会对少数几个语义类别给予非常高的分数, 而其他大部分的语义类别分数都在 0.1 以下. 而本文训练的模型给出的前几位片段视频语义类别分数更加分散, 虽然前几个语义类别的分数也非常高, 但给出的后续若干类别的分数是逐渐减小的. 这说明相比较于视频, 片段视频的语义更为模糊, 尽管该片段视频的语义应当与所属视频保持一致, 但其视觉特征可能与其他的语义类别有一定的相似度. 另一个可能的原因是相比较于片段视频若干个二值标签的标注, 视频的多标签标注能让模型得到更充分的训练, 因此本文用视频的语义类别筛选片段视频的语义类别的方法是可靠的. 依次看单个视频的预测结果, 虽然部分视频预测结果不准确, 但利用所有结果筛选出来的片段视频语义类别大部分都不存在于片段视频中, 因而能够提高筛选过后的片段视频每个语义类别的识别准确率. 同时看每一个视频内的 3 个片段视频的预测结果, 不同的片段视频的预测结果区别较大, 说明视频语义会随着时间的进行而改变. 总体而言, 预测正确的视频语义会较高频率地出现在大多数片段视频的预测结果中.

5 总结

本文介绍了一种基于长短时预测一致性的大规模视频语义识别算法. 本文提出的算法利用视频和

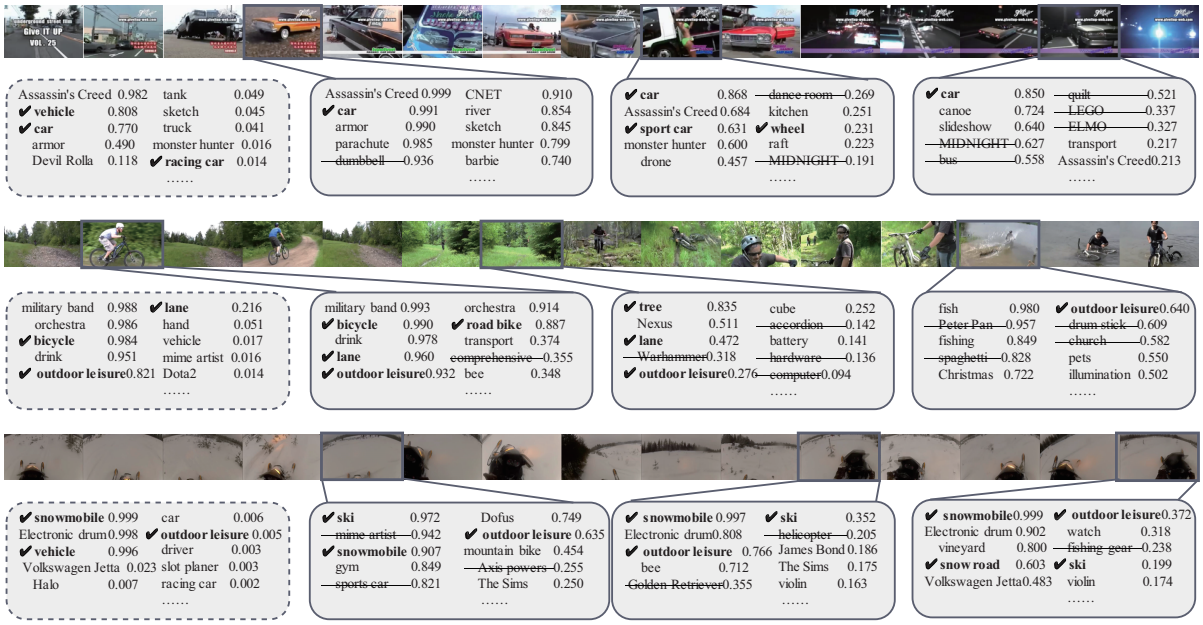


图 3 (网络版彩图) 预测结果样例示意图. 虚线框为视频预测结果, 实线框为片段视频预测结果, 预测正确的类别打勾, 一致性预测算法舍弃的类别划横线

Figure 3 (Color online) The diagram of prediction. Dotted frames are the video predictions, solid line frames are the video segment predictions, correct categories are checked, and categories discarded by our algorithm are crossed out

片段视频的语义一致性为片段视频语义预测提供指导. 本文所提出的算法包括用视频数据预训练识别模型, 用片段视频数据微调模型. 最后结合长短时预测一致性的预测算法, 得到了最优的预测结果. 本文所做的重要工作包括: 结合多种语义识别模型以获得最优的预测结果; 使用长短时预测一致损失函数减轻片段视频标签标注不充分可能引起的欠拟合问题; 利用长短时预测一致性筛选片段视频的识别结果, 提高片段视频的预测准确性. 本文提出的算法达到了 82.62% 的 MAP 识别精度, 在全球 283 个团队参加的第三届 YouTube-8M 比赛中排名第二. 在未来的方向上, 本文希望为视频语义理解引入因果推理 [49] 和符号学习 [50] 等具有可解释性的模型和算法.

参考文献

1 Marszalek M, Laptev I, Schmid C. Actions in context. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2009. 2929–2936

2 Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition. In: Proceedings of International Conference on Computer Vision, Barcelona, 2011. 2556–2563

3 Khurram S, Amir R Z, Mubarak S. UCF101: a dataset of 101 human action classes from videos in the wild. 2012. ArXiv: 1212.0402

4 Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks. In: Proceedings of Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 1725–1732

5 Jiang Y G, Wu Z, Wang J, et al. Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Trans Pattern Anal Mach Intell, 2017, 40: 352–364

6 Heilbron B G F C, Escorcia V, Ghanem B, et al. ActivityNet: a large-scale video benchmark for human activity understanding. In: Proceedings of Conference on Computer Vision and Pattern Recognition, Boston, 2015. 961–970

7 Abu-El-Haija S, Kothari N, Lee J, et al. YouTube-8M: a large-scale video classification benchmark. 2016. ArXiv: 1609.08675

8 Sigurdsson G A, Varol G, Wang X, et al. Hollywood in homes: crowdsourcing data collection for activity understanding. In: Proceedings of European Conference on Computer Vision, Las Vegas, 2016. 510–526

9 Sigurdsson G A, Gupta A, Schmid C, et al. Charades-ego: a large-scale dataset of paired third and first person videos. 2018. ArXiv: 1804.09626

10 Carreira J, Noland E, Hillier C, et al. A short note on the kinetics-700 human action dataset. 2019. ArXiv: 1907.06987

11 Murray N, Marchesotti L, Perronnin F. AVA: a large-scale database for aesthetic visual analysis. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2012. 2408–2415

12 Goyal R, Kahou S E, Michalski V, et al. The "Something Something" video database for learning and evaluating visual common sense. In: Proceedings of International Conference on Computer Vision, Venice, 2017

13 Monfort M, Andonian A, Zhou B, et al. Moments in time dataset: one million videos for event understanding. IEEE Trans Pattern Anal Mach Intell, 2020, 42: 502–508

14 Monfort M, Ramakrishnan K, Andonian A, et al. Multi-moments in time: learning and interpreting models for multi-action video understanding. 2019. ArXiv: 1911.00232

15 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780

16 Donahue J, Anne H L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 2625–2634

17 Wu Z, Wang X, Jiang Y G, et al. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: Proceedings of the 23rd ACM International Conference on Multimedia, 2015. 461–470

18 Ng Y J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 4694–4702

19 Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 4489–4497

20 Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2017. 6299–6308

21 Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks. In: proceedings of the IEEE International Conference on Computer Vision, Venice, 2017. 5533–5541

22 Lin J, Gan C, Han S. Tsm: temporal shift module for efficient video understanding. In: Proceedings of the IEEE International Conference on Computer Vision, Seoul, 2019. 7083–7093

23 Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell, 2012, 35: 221–231

24 Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 1933–1941

25 Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 4305–4314

26 Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 20–36

27 Wang X, Girshick R, Gupta A, et al. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Munich, 2018. 7794–7803

28 Yan A, Wang Y, Li Z, et al. PA3D: pose-action 3D machine for video recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 7922–7931

29 Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of AAAI Conference on Artificial Intelligence, San Francisco, 2017

30 Hershey S, Chaudhuri S, Ellis D P, et al. CNN architectures for large-scale audio classification. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, 2017. 131–135

31 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 7132–7141

32 Jégou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation. In: Proceedings of Conference on Computer Vision Pattern Recognition, San Francisco, 2010. 3304–3311

33 Gong Y, Wang L, Guo R, et al. Multi-scale orderless pooling of deep convolutional activation features. In: Proceedings

of European Conference on Computer Vision, Zurich, 2014. 392–407
34 Arandjelovic R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 5297–5307
35 Lin R, Xiao J, Fan J. Nextvlad: an efficient neural network to aggregate frame-level features for large-scale video classification. In: Proceedings of the European Conference on Computer Vision, Munich, 2018
36 Tang Y, Zhang X, Ma L, et al. Non-local netVLAD encoding for video classification. In: Proceedings of the European Conference on Computer Vision, Munich, 2018
37 Philbin J, Chum O, Isard M, et al. Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of Conference on Computer Vision and Pattern Recognition, Minneapolis, 2007. 1–8
38 Sivic J, Zisserman A. Video Google: a text retrieval approach to object matching in videos. In: Proceedings of International Conference on Computer Vision, 2003
39 Jegou H, Perronnin F, Douze M, et al. Aggregating local image descriptors into compact codes. IEEE Trans Pattern Anal Mach Intell, 2011, 34: 1704–1716
40 Perronnin F, Liu Y, Sánchez J, et al. Large-scale image retrieval with compressed fisher vectors. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2010. 3384–3391
41 Miech A, Laptev I, Sivic J. Learnable pooling with context gating for video classification. 2017. ArXiv: 1706.06905
42 Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014. ArXiv: 1412.3555
43 Ostyakov P, Logacheva E, Suvorov R, et al. Label denoising with large ensembles of heterogeneous neural networks. In: Proceedings of the European Conference on Computer Vision, Munich, 2018
44 Lee J, Reade W, Sukthankar R, et al. The 2nd YouTube-8M large-scale video understanding challenge. In: Proceedings of the European Conference on Computer Vision, 2018
45 Izmailov P, Podoprikhin D, Garipov T, et al. Averaging weights leads to wider optima and better generalization. 2018. ArXiv: 1803.05407
46 Ma J, Gorti S K, Volkovs M, et al. Cross-class relevance learning for temporal concept localization. 2019. ArXiv: 1911.08548
47 Lin R C, Xiao J, Fan J P. MOD: a deep mixture model with online knowledge distillation for large scale video temporal concept localization. 2019. ArXiv: 1910.12295
48 Dai S. A segment-level classification solution to the 3rd YouTube-8M video understanding challenge. 2019. https://research.google.com/youtube8m/workshop2019/c_09.pdf
49 Cheng C, Zhang C, Wei Y, et al. Sparse temporal causal convolution for efficient action modeling. In: Proceedings of the 27th ACM International Conference on Multimedia, Nice, 2019. 592–600
50 Dai W Z, Xu Q L, Yu Y, et al. Tunneling neural perception and logic reasoning through abductive learning. 2018. ArXiv: 1802.01173

Large-scale video semantic recognition based on consistency of segment-level and video-level predictions

Zheng WANG[1,2], Zejia WENG[1,2], Rui WANG[1,2], Jingjing CHEN[1,2] & Yu-Gang JIANG[1,2*]

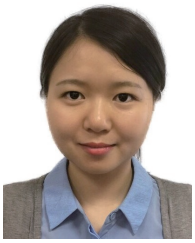
1. *School of Computer Science, Fudan University, Shanghai* 201203*, China*;
2. *Shanghai Key Lab of Intelligent Information Processing, Shanghai* 200433*, China*
* Corresponding author. E-mail: ygj@fudan.edu.cn

Abstract Segment-level video semantic recognition, which known to be an important task in video analysis, attempts to identify the semantic concepts in short video clips. Labeling video segments is difficult because there is an extremely large number of segments and there are no network tags; consequently, only a portion of the video segments are labeled. Determining how to improve the accuracy of semantic recognition of fragmented videos with limited semantic labels is a key challenge in video semantic recognition. This paper proposes a video semantic recognition algorithm based on the consistency of video- and segment-level predictions. The proposed algorithm introduces the constraint of consistency between complete video semantics and fragmentary video semantics. The proposed algorithm can be applied to filter the video segment semantic results to improve recognition accuracy. The proposed algorithm achieved 82.62% mean average precision on the video segment semantic recognition task using the large-scale video dataset YouTube-8M and ranked second in the third YouTube-8M competition.

Keywords large-scale video semantic recognition, segment-level semantic recognition, semantic consistency, feature aggregation, reliable prediction

Zheng WANG was born in 1995. He received a B.S. degree in computer science from Zhejiang University of Technology, China, in 2017. He is currently working toward a Ph.D. degree at the School of Computer Science, Fudan University. His research interests include computer vision and multimedia retrieval.

Zejia WENG was born in 1998. He is currently working toward a bachelor degree at the School of Computer Science, Fudan University. His current research interest is large-scale video recognition.



Jingjing CHEN was born in 1990. She received her Ph.D. degree in Computer Science from City University of Hong Kong. She is now a pre-tenured associate professor at the School of Computer Science, Fudan University. Previously, she was a postdoctoral research fellow in the School of Computing at the National University of Singapore. Her research interests are diet tracking and nutrition estimation based on multi-modal processing of food images.

Yu-Gang JIANG was born in 1981. He received a Ph.D. degree in computer science from City University of Hong Kong. From 2008 to 2011, he was with the Department of Electrical Engineering, Columbia University, New York, USA. He is currently a professor of computer science at Fudan University, Shanghai, China. His research interests include multimedia content analysis and computer vision.