计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(5):252−256 [doi: 10.15888/j.cnki.csa.007410]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多粒度视频信息和注意力机制的视频场景识别①

袁韶祖, 王雷全, 吴春雷

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通讯作者: 吴春雷, E-mail: wuchunlei06@163.com

摘　要: 视频场景识别是机器学习和计算机视觉一个重要的研究领域. 但是当前对于视频场景识别的探索工作还远远不够, 而且目前提出的模型大都使用视频级的特征信息, 忽略了多粒度的视频特征关联. 本文提出了一种基于多粒度的视频特征的注意力机制的模型架构, 可以动态高效的利用各维度视频信息之间存在的丰富的语义关联, 提高识别准确度. 本文在中国多媒体大会 (CCF ChinaMM 2019) 最新推出的 VideoNet 数据集上进行了实验, 实验结果表明基于多粒度的视频特征的注意力机制的模型与传统方法相比具有明显的优越性.

关键词: 视频场景识别; 注意力机制; 多粒度视频信息; 卷积神经网络; 检测网络

引用格式: 袁韶祖,王雷全,吴春雷.基于多粒度视频信息和注意力机制的视频场景识别.计算机系统应用,2020,29(5):252–256. http://www.c-s-a.org.cn/1003-3254/7410.html

Video Scene Recognition with Multi-Granularity Video Features and Attention Mechanism

YUAN Shao-Zu, WANG Lei-Quan, WU Chun-Lei

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Video scene recognition has attracted much attention in the field of machine learning and computer vision. It is not only an important practical application, but also a challenge for image understanding in the field of computer vision. Nevertheless, current exploration of video scene recognition has not been unable to meet the needs of production environment. And most proposed models only use video-level feature information, while ignore association of multi-granularity video feature. In this study, we propose an architecture of attention mechanism with multi-granularity video features, which can make use of the rich semantic association among the various dimensions of video information dynamically and efficiently, and improve the performance of the model. The experiments are conducted on the latest VideoNet dataset released by CCF China MM 2019. The result shows that the proposed model based on attention mechanism model with multi-granularity video features outperforms the previous methods.

Key words: video scene recognition; attention mechanism; multi-granularity video features; CNN; detection network

1　引言

近年来, 随着深度学习技术的发展, 大量针对物体、人脸、动作等维度的识别技术涌现出来. 而随着监控技术和短视频 APP 的广泛应用, 视频场景识别已成为一项极具科研价值和应用价值的技术. 它的具体任务是给定一个特定的视频进行镜头分割, 通过提取关键帧, 输出场景的类别. 目前主流的算法是使用视频级别的特征直接进行场景分类. 然而这种方法只考虑到了视频级的全局特征, 却忽略了富含更多信息的局部特征以及其中存在的关联. 针对以上问题, 本文提出

① 基金项目: 山东省重点研发计划 (2019GGX101015); 中央高校自主创新科研计划 (17CX02041A, 18CX02136A)
Foundation item: Key Research and Development Program of Shandong Province (2019GGX101015); Innovative Research Program of the Central Universities of China (17CX02041A, 18CX02136A)
收稿时间: 2019-10-16; 修改时间: 2019-11-15, 2019-11-20; 采用时间: 2019-11-22; csa 在线出版时间: 2020-05-07

了一种新的模型, 该模型利用视频级别的全局信息和物体级别的局部信息, 提供更加丰富的推断信息. 同时, 本文采用了注意力机制来筛选对于视频场景识别重要程度高的特征, 这一过程既增强了全局信息和局部信息的关联, 同时也实现了对于特征的降维, 有效地加速了模型的收敛. 与官方开源的模型相比, 本文提出的模型在准确率上取得了非常大的提升, 这进一步说明了该模型的有效性.

本文中, 创新点可以总结归纳为如下 3 点:

1) 本文在视频场景分类中构造了全局和局部的多粒度的特征.

2) 本文提出全新的注意力机制的场景分类模型, 该模型可以很好的通过注意力机制将两种粒度的特征融合, 并对结果进行降维.

3) 新模型准确率比官方发布的基于 CNN 网络的模型提高了 12.42%, 这进一步证明我们的模型的有效性和优越性.

2 相关工作

2.1 视频级特征和物体级特征

特征在计算机视觉领域中扮演着重要的角色, 选择合适的特征可以极大的提升模型的性能. 早期视频特征主要使用 VGG 特征, 该模型由 Simonyan K 等提出, 也大量应用在图像识别领域. 后来何凯明通过残差的思想实现了 101 层的 CNN 模型, 得到了拟合更强的网络[1]. Resnet 作为特征提取网络被广泛应用于视频识别和图像描述等领域[2]. Jiang YG 等使用 resnet 作为视频级特征实现了视频场景分类的基础模型[3]. 使用 Resnet 提取的视频级特征也被称作 RGB 特征. 然而视频帧之间是存在时空关系的, 采用 RGB 特征无法表征出这种时序关系[4]. 为了解决这一问题, Tran D 等提出了空间卷积 (C3D) 的网络来获取时空的信息[5]. Sun DQ 等提出利用帧之间的差异性计算时空信息的“光流法”[6]. 这两种跨时空特征被广泛的应用于视频是被, 动作识别等领域[7]. 以上特征都可以被视作视频级别的特征, 未从更细的粒度考虑视频内部的语义特征联系. Ren SQ 等认为, 细粒度的特征有利于增强模型对于视觉信息的理解, 为了得到这种信息, 他们在较大的视觉检测数据集上训了 Faster-RCNN[8]用于识别目标图像中的物体, 同时提出检测模型标识每个物体的中间特征, 并将所有特征级联起来作为图像的总体特征[9]. 该模型首次提出后被应用于图像描述和图像问答领域, 并取得了不错的成绩. 我们认为, 该特征同样可以应用于视频理解领域.

2.2 注意力机制

注意力机制在深度学习领域有着极为重要和深远的影响, 被广泛应用各个领域中. 在机器翻译领域, 早期的 Encode-Decoder 模型不能很好的解码源语言中的重点信息, 为了解决这一问题, Bahdanau 等将注意力机制最早应用于机器翻译的解码阶段[10]. 受到这种思维的启发, Xu K 等意识到图像领域也存在需要重点关注的区域, 于是他们将注意力机制引入到图像描述中来, 并创造性的提出了两种注意力机制: 软注意力和基于强化学习的硬注意力[11]. 在这之后注意力机制在各个领域大放异彩, 陆续出现了很多新式的注意力机制. 在图像描述领域, Lu JS 等提出了 when to look 注意力, 去决定在图像描述过程中应该注意图像还是注意文本[12]. 在图像问答中, Lu JS 等 提出公用注意力机制, 从理论层面将注意力矩阵逆置之后用于两种模态[13], Kim JH 提出双线性注意力[14], 相当于给注意力矩阵降维, 但是最终的结果不变, 两种注意力都可以降低运算复杂度, 有利于采用更深的注意力网络, 从而提升效果. 在对抗生成领域, Kim J 将注意力机制引入到了生成对抗网络, 通过网络自适用的决定应该更注重哪一区域的生成, 用来生成更高质量的图[15]. 即便是在最新谷歌提出的 Transfromer 和 Bert 中, 也采用了自注意力机制, 用来解决自然语言中超远距离词的依赖问题, 该模型在自然语言界引起了极大轰动[16]. 由于注意力机制在人工智能领域的出色表现, 因此在实验中也会用注意力机制来提升本文所提出模型的能力.

3 视频场景识别方法模块介绍

3.1 基于 Resnet 和 Faster-RCNN 的多粒度特征构造

Resnet 是深度卷积神经网络的一种, 它在原有的较浅层次的卷积神经网络的基础上添加了“残差”机制, 因此再反向传播的过程中可以保证导数不为 0, 从而避免了深层网络出现梯度弥散的现象,有效的增加了卷积的拟合性. Resnet 的残差过程可由式 (1) 表示:

$$y = F(x) + Wx \tag{1}$$

其中, x 是输入的特征图, F 代表卷积, W 是用来调整 x 的 channel 维度的, y 是当前残差的输出.

由于 Resnet 的输出可以作为对图片信息的一个较强的表征, 本文采用这种特征作为视频场景的一个全局表示, 即粗粒度特征.

Faster-RCNN 是一种比较新且准确率较高的检测模型, 其原理和 SPPnet[6]和 Fast-RCNN[17]这些模型有很大差别, 这些模型虽然减少了检测网络运行的时间, 但是计算区域建议依然耗时依然比较大. Faster-RCNN 采用了区域建议网络 (region proposal network) 用来提取检测物体的区域, 它和整个检测网络共享全图的卷积特征, 极大的降低区域建议网络所花时间, 从而提升了检测的效率和质量.

在本文中, Faster-RCNN 作为检测器标识出视频图片中的物体信息, 每一个物体区域分别作为改物体的特征表示, 这种检测得到的特征作为细粒度的特征表示.

3.2 多粒度特征的注意力融合模型

图 1 是本文所提出的场景识别模型, 这里所采用的的注意力机制是一种典型的注意力架构[10], 并在此基础上设计了多粒度特征的注意力融合模型. 在 3.1 中检测模型 Faster-RCNN 提取提取到的检测特征 S 是一个 $n\times D$ 维的向量, 即对应于 n 个不同物体的子区域, 每个区域都是一个 D 维的向量,可由如下字母表示:

$$S=\{S_1,S_2,\cdots,S_i,\cdots,S_n\},\ S_i\in R^D \tag{2}$$

其中, R^D 表示属于 D 维度, S_i 表示第 i 个物体的图像区域. 对于每个物体的特征表示, 式 (3) 中本文借鉴注意力分配函数$\emptyset$ [18]根据细粒度检测特征 S_i 和全局特征 I_i 生成一个权重分布 α_i:

$$\alpha_i=\emptyset(I_i,S_i),\ i\in[1,n] \tag{3}$$

这里的分配函数是一种映射关系, 它将两种粒度的视觉信息通过单层神经元映射到同一个维度空间, 再相加得到权重, 这个权重分布就包含了两种粒度特征的融合信息. 同时, 该权重分布和 S_i 的维度是一致的, 通过后续的加权操作, 既实现了对于多个物体特征的降维, 又得到两种信息融合的一个强表征信息.

在 (4) 式中, Softmax 函数对权重分布 α_i 作归一化处理得到注意力权重 a_i, 这时 a_i 介于 0 到 1 之间:

$$a_i=\frac{\exp(\alpha_i)}{\sum_{k=1}^{n}\exp(\alpha_i)} \tag{4}$$

其中, a_i 表示视觉注意力模型中第 i 个物体的图像对应区域的权重.

最后, 将注意力权重和相对应的视频图像区域加权求和, 得到该视频场景的最终表示 att, 如式 (5) 表示:

$$att=\sum_{k=1}^{n}S_i\alpha_i \tag{5}$$

式中, S_i 为视频图像的区域, α_i 为式 (4) 中 attention 学习得到的权重, 这个权重是神经网络根据当前输入视觉信息自动生成的.

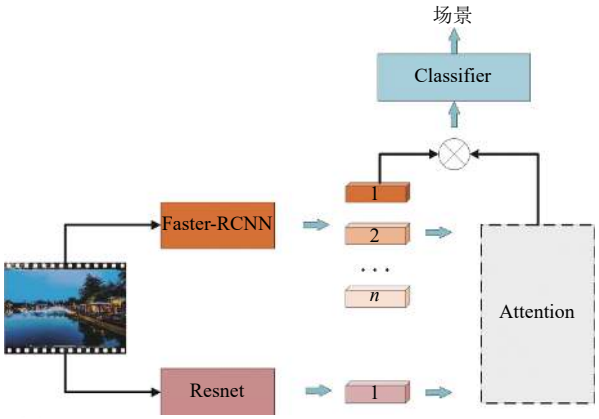

图 1 我们的模型架构

3.3 新模型整体架构

在视频场景识别中, 首先将给定的视频切割成一个视频帧序列 $T_i(i=1,2,\cdots,m)$, 模型要对这 m 个特定视频帧进行场景分类的 $p_i\ (i=1,2,\cdots,m)$. 接下来两种特征的提取: 我们使用深度卷积神经网络 Resnet 提取视频帧全局的视觉特征 $I_i\ (i=1,2,\cdots,m)$,这同时也是即将进行场景分类的帧的粗粒度的表示, 该表示是一个 D(2048) 维的向量; 同样的, 通过预训练的 Faster-RCNN 提取视频场景中的物体区域, 也就是检测特征, 该特征是物体级别的细粒度信息, 可以表示为 $S=\{S_1,S_2,\cdots,S_N\}$, 其中 n 代表检测模型提取的物体区域个数, 实践中 n 被设置为 36. 这个过程可用下面两个公式表示:

$$I_i=f(T_i) \tag{6}$$

$$S_i=g(T_i) \tag{7}$$

为了示意方便, 这里 f 代表深度卷积网络 Resnet, g 代表检测网络 Faster-RCNN.

得到多粒度的视觉特征后, 新模型使用全局特征作为注意力机制的键值, 通过注意力单元的计算得到 n 个注意力权重 α. 这里的权重 α 是由注意力模型根据不同物体重要程度学习得到的: 物体重要程度越大, 其权重值约接近于 1; 如果物体对于场景推断越不重要要甚至起到干扰作用, 其权重越接近于 0. 最后通过物体特征和注意力机制生成的权重加权计算得到融合多粒度信息表示的视觉特征 att, 这同时也实现了对于细粒度特征的降维, 即从 $n\times D$ 维降维成 D, 所以 att 是一个

D 维的向量. 这部分流程图如图 1 所示, 可以由式 (8)、式 (9) 概括:

$$\alpha_i = attention(I_i, S_i) \tag{8}$$

$$att = S_i \times \alpha_i \quad (i = 1, 2, \cdots, n) \tag{9}$$

最终, 融合多粒度信息表示的视觉特征被输入到一个分类器中. 该分类器由一个两层的神经网络, 和一个激活函数构成, 它的作用是将 D 维表示向量映射为 d, d 代表了场景分类的总数, 选取其中值对应的最大的索引, 该索引所对应的场景表示就是最后输出的场景分类的结果. 分类器部分可以用式 (6), 式 (7) 表示:

$$logit = W_2(W_1 \times att + b_1) + b_2 \tag{10}$$

$$p = Softmax(logit) \tag{11}$$

式中, W_1, W_2 代表两层神经网络的可学习权重, $logit$ 是未经过激活函数的值, p 为最终的分类概率, 概率最大的索引所对应的场景即为神经网络的输出结果.

3.4 总结

和已有的方法[3]相比, 本文摒除了只采用单维度的 CNN 特征或者将几种 CNN 特征简单连接的方法. 本模型通过已有的深度卷积和检测的方法构建了两种不同粒度的特征. 特别的, 本文采用注意架构将两种粒度的信息巧妙融合在了一起, 既实现了对信息的降维, 同时增强了全局信息和局部信息的关联.

4 实验

4.1 数据集和评估方法

本文采用了在 ChinaMM 大会上极链科技与复旦大学联合推出全新视频数据集 VideoNet. 该数据集具备规模大、维度多、标注细三大特点. VideoNet 包含近 9 万段视频, 总时长达 4000 余小时. VideoNet 数据集对视频进行了事件分类标注, 并针对每个镜头的关键帧进行了场景和物体两个维度的共同标注. 考虑到算力等因素, 该实验从中抽取了 100 000 个视频样本的镜头分割和关键帧结果, 推断每个镜头的关键帧对应的场景类别. 为了保证模型的训练和测试效果, 本实验按照 6:2:2 的比例切随机分数据集, 即使用 60 000 数据训练, 20 000 用于验证, 20 000 用于测试.

4.2 评估方法

模型的目标是对给定的测试视频样本和镜头关键帧结果, 推断每个镜头的关键帧对应的场景类别. 因此可以通过以下公式判读模型是否分类正确:

$$I_i = \begin{cases} 1, & 当 p_i \in G_i \\ 0, & 其他 \end{cases} \tag{12}$$

其中, G 为关键帧场景类别的 ground-truth, p_i 为场景预测输出. 如果该关键帧未出现训练集中任何一类场景, 则 G_i=–1. 因此, 准确率公式可以定义为:

$$A_C = \sum_{i=1}^{NS} \frac{I_i}{N} \mathrm{p} \tag{13}$$

训练过程中该模型使用了交叉熵[19]作为损失, 因此也可以通过交叉熵损失的变化判断模型的优化程度和模型训练是否收敛. 损失函数可用公式表示为:

$$L(\theta) = -\sum G_i \log p_i \tag{14}$$

4.3 实验分析

本文采用了准确率和 log 损失来评测模型的质量和训练情况. 在图 2 中,我们绘制了测试损失和迭代次数的相关折线图, 不难看出本文提出的方法可以快速的收敛, loss 值在训练的过程中稳定的下降, 最终迭代次数为 20 时得到最好的效果. 结合图 3 的准确率曲线, 通过观察可以看出随着训练损失的下降, 模型的测试准确率也在不断提升, 最高可以达到 67.71%. 由于模型训练了 25 个迭代, 通过图 3 表所示, 在超过 20 个迭代次数的时候, 模型的测试准确率会有小幅度的下降, 说明模型出现了过拟合现象. 在表 1 中, 我们列举了模型迭代次数 19 到迭代次数 25 之间的准确率, 通过对比发现, 迭代次数为 23 的时候模型得到最好的效果, 准确率为 67.71%.

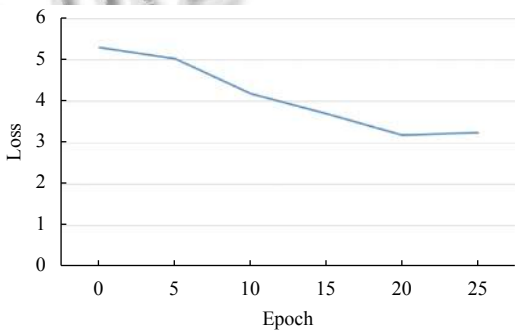

图 2 交叉熵损失变化

通过表 1, 可以看出, 本文提出的模型准确率大幅度优于 VideoNet 官方开源的 Baseline 模型. 与我们提出模型训练取得的最好的效果相比, 新模型准确率比官方 baseline 提升了 12.42%. 这些数据证明: 本文提出的模型可以在较少的训练迭代次数下收敛. 基于多粒度视觉特征和注意力机制的模型有效的提升了视频场

景识别的质量. 相比于传统的使用 C3D 特征等方法, 多粒度视觉信息可以大幅度提升识别的准确率, 因为不同粒度的信息不但补充了更加丰富的识别信息, 同时还使用注意力机制将不同粒度的信息联系在一起, 更加充分的利用了信息.

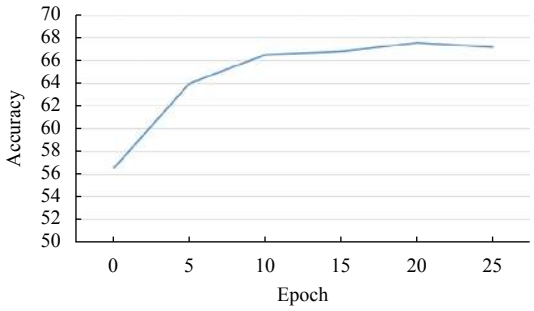

图 3 准确率损失变化

表 1 模型的准确率对比 (%)

模型	准确率
官方 Baseline[2]	55.29
新模型 (19 epoch)	67.54
新模型 (21 Epoch)	66.99
新模型 (23 Epoch)	**67.71**
新模型 (25 Epoch)	67.18

5 结论与展望

本文提出了使用多粒度视频特征信息基于注意力架构的视频场景检测模型, 并在 VideoNet 数据集上取得优异的成绩. 该算法的亮点在于使用全局性的信息引导下, 通过注意力机制自适应的对场景中重要的局部信息加权, 从而达到更加精准的识别效果. 和官方开源的模型基线相比, 本文考虑了全局特征和局部特征, 很好的利用了多个粒度视频信息. 并且在模型中采用了注意力模型, 既完成了对特征的降维, 又能很好的将多个粒度的信息联系起来. 在未来的工作中, 我们将进一步探索多维度的视频信息和不同注意力机构对于场景识别的影响.

参考文献

1 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. arXiv preprint arXiv: 1512.03385, 2015.

2 Vinyals O, Toshev A, Bengio S, *et al*. Show and tell: A neural image caption generator. arXiv preprint arXiv: 1411.4555, 2015.

3 Jiang YG, Wu ZX, Wang J, *et al*. Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(2): 352–364. [doi: 10.1109/TPAMI.2017.2670560]

4 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014.

5 Tran D, Bourdev L, Fergus R, *et al*. C3D: Generic features for video analysis. arXiv preprint arXiv: 1412.0767, 2014.

6 Sun DQ, Yang XD, Liu MY, *et al*. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. arXiv preprint arXiv: 1709.02371, 2017.

7 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv: 1406.2199, 2014.

8 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv preprint arXiv: 1506.01497, 2015.

9 Anderson P, He XD, Buehler C, *et al*. Bottom-up and top-down attention for image captioning and visual question answering. arXiv preprint arXiv: 1707.07998, 2017.

10 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv: 1409.0473, 2014.

11 Xu K, BA J, Kiros R, *et al*. Show, attend and tell: neural image caption generation with visual attention. arXiv preprint arXiv: 1502.03044, 2015.

12 Lu JS, Xiong CM, Parikh D, *et al*. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. arXiv preprint arXiv: 1612.01887, 2016.

13 Lu JS, Yang JW, Batra D, *et al*. Hierarchical question-image co-attention for visual question answering. arXiv preprint arXiv: 1606.00061, 2016.

14 Kim JH, Jun J, Zhang BT. Bilinear attention networks. arXiv preprint arXiv: 1805.07932, 2018.

15 Kim J, Kim M, Kang H, *et al*. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv: 1907.10830, 2019.

16 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. arXiv preprint arXiv: 1706.03762, 2017.

17 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805, 2018.

18 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv: 1409.0473, 2014.

19 Deng LY. The cross-entropy method: A unified approach to combinatorial optimization, monte-carlo simulation, and machine learning. Technometrics, 2004, 48(1): 147–148.