基于视觉的视频分类入门

Introduction to Visual-based video classification

互联网上图像和视频的规模日益庞大，据统计 Youtube 网站每分钟就有数百小时的视频产生，这使得急切需要研究视频相关算法帮助人们更加容易地找到感兴趣内容的视频。这些视频分类算法能实现自动分析视频所包含的语义信息、理解其内容，对视频进行自动标注、分类和描述，达到与人媲美的准确率。**大规模视频分类是继图像分类问题解决后下一个急需解决的关键问题。**

视频分类的主要目标是理解视频中包含的内容，确定视频对应的几个关键主题。视频分类（Video Classification）算法将基于视频的语义内容如人类行为和复杂事件等，将视频片段自动分类至单个或多个类别[1]。视频分类不仅仅是要理解视频中的每一帧图像，更重要的是要识别出能够描述视频的少数几个最佳关键主题。**视频分类的研究内容主要包括多标签的通用视频分类和人类行为识别等。**与之密切相关的是，**视频描述生成（Video Captioning）试图基于视频分类的标签，形成完整的自然语句，为视频生成包含最多动态信息的描述说明。**

虽然融合多种特征如文本-图像融合、声音-视频融合对提高视频分类的性能有所帮助，但是本文主要**关注研究融合视频本身的空间和时间特征，也称为基于视觉的视频分类。**

一、 传统视频分类方法研究

在深度学习方法广泛应用之前，大多数的视频分类方法采用基于人工设计的特征和典型的机器学习方法研究行为识别和事件检测。

传统的视频分类研究专注于采用对局部时空区域的运动信息和表观（Appearance）信息编码的方式获取视频描述符，然后利用词袋模型（Bag of Words）等方式生成视频编码，最后利用视频编码来训练分类器（如 SVM），区分视频类别。视频的描述符依赖人工设计的特征，如使用运动信息获取局部时空特征的梯度直方图（Histogram of Oriented Gradients，HOG），使用不同类型的轨迹的光流直方图（Histogram of Optical Flow, HOF）和运动边界直方图（Motion Boundary Histogram，MBH）。通过词袋模型或 Fisher 向量方法，这些特征可以生成视频编码。

当前，基于轨迹的方法（尤其是 DT 和 IDT）是最高水平的人工设计特征算法的基础[2]。许多研究者正在尝试改进 IDT，如通过增加字典的大小和融合多种编码方法，通过开发子采样方法生成 DT 特征的字典，在许多人体行为数据集上取得了不错的性能。

然而，随着深度神经网络的兴起，特别是 CNN、LSTM、GRU 等在视频分

类中的成功应用，其分类性能逐渐超越了基于 DT 和 IDT 的传统方法，使得这些传统方法逐渐淡出了人们的视野。

二、 深度网络方法研究

深度网络为解决大规模视频分类问题提供了新的思路和方法。近年来得益于深度学习研究的巨大进展，特别是卷积神经网络（Convolutional Neural Networks，CNN）作为一种理解图像内容的有效模型，在图像识别、分割、检测和检索等方面取得了最高水平的研究成果。卷积神经网络 CNN 在静态图像识别问题中取得了空前的成功，其中包括 MNIST、CIFAR 和 ImageNet 大规模视觉识别挑战问题。CNN 采用卷积与池化操作，可以自动学习图像中包含的复杂特征，在视觉对象识别任务中表现出很好的性能。基于 CNN 这些研究成果，国内外开始研究将 CNN 等深度网络应用到视频和行为分类任务中。

与图像识别相比，视频分类任务中视频比静态图像可以提供更多的信息，包括随时间演化的复杂运动信息等。视频(即使是短视频)中包含成百上千帧图像，但并不是所有图像都有用，处理这些帧图像需要大量的计算。最简单的方法是将这些视频帧视为一张张静态图像，应用 CNN 识别每一帧，然后对预测结果进行平均处理来作为该视频的最终结果。然而，这个方法使用了不完整的视频信息，因此使得分类器可能容易发生混乱。

（1） 监督学习方法

i. 基于图像的视频分类：将视频片段视为视频帧的集合，每个视频帧的特征通过 ImageNet 数据集上预先训练的最高水平的深度模型（如 *AlexNet，VGGNet，GoogLeNet，ResNet*）进行获取。*最终，帧层特征汇聚为视频层特征，作为标准分类器（如 SVM）识别的输入。*

ii. 端到端的 CNN 网络：关注于利用 CNN 模型学习视频隐含的时空模式，如 3D CNN，Two-stream CNN，TSN 模型等。

iii. 双流（Two-stream）法中的时间 CNN 只能获取很短时间窗口内的运动信息，难以处理长时间多种行为组成的复杂事件和行为。因此，**引入 RNN 来建模长期时间动态过程**，常用的模型有 LSTM，GRU-RNN 等。LSTM 避免了梯度消失的问题，在许多图像和视频摘要、语音分析任务中非常有效。

iv. 视频中包含了很多帧，处理所有的视频帧计算代价很大，也会降低识别那些与类别相关的视频帧的性能。因此，**引入视觉注意力机制**来识别那些与目标语义直接相关的最有判别力的时空特征

（2） 非监督学习方法

采用非监督学习的方法，整合空间和时间上下文信息，是发现和描述视频结构的一种很有前途的方法。

三、 视频分类数据集

图像数据集基准对图像分类问题解决起到了非常重要的推动作用。从最早的小规模的带标注的数据集 Caltech101/256, MSRC, PASCAL，当更大的数据集如 ImageNet 和 SUN 发布后，图像理解的视觉算法研究进展很快。特别是 ImageNet 及其大规模视觉识别挑战赛（ImageNet Large Scale Visual Recognition Challenge, ILSVRC）极大地促进了深度特征学习技术的发展，陆续出现了 AlexNet、VGGNet、Inception、ResNet 等网络架构，最终使得识别错误率低于人眼，说明 CNN 已经基本解决了 ImageNet 数据集上的图片分类问题。

近年来为推动视频分类的研究，也陆续发布了相关的视频数据集。小型标注良好的数据集如 KTH，Hollywood2，Weizmann；中型的数据集如 UCF101，Thumos'14 和 HMDB51，这些数据集超过了 50 类行为类别；大型数据集如 Sports-1M，YFCC-100M，FCVID 数据集，ActivityNet 数据集，YouTube-8M 等。

数据集统计表

数据集	视频数	分类数	发布年	背景
KTH	600	6	2004	干净 静态
Weizmann	81	9	2005	干净 静态
Kodak	1,358	25	2007	动态
Hollywood	430	8	2008	动态
Hollywood2	1,787	12	2009	动态
Olympic Sports	800	16	2010	动态
HMDB51	6,766	51	2011	动态
CCV	9,317	20	2011	动态
UCF-101	13,320	101	2012	动态
THUMOS-2014	18,394	101	2014	动态
MED-2014	约 31,000	20	2014	动态
Sports-1M	1,133,158	487	2014	动态
MPII Human Pose	20,943	410	2014	动态
ActivityNet	27,901	203	2015	动态
EventNet	95,321	500	2015	动态
FCVID	91,223	239	2015	动态
YouTube-8M	8,264,650	4,800	2016	动态

其中比较有代表性的有 YouTube-8M（2016）、ActivityNet（2015）、Sports-1M（2014）、**UCF-101（2012）**、**HMDB51（2011）**等。

YouTube-8M 的提出标志着视频分类朝大规模通用多标签视频分类的方向发展。

当前的研究结果表明：

- HMDB51 数据集上，DOVF+MIFS 方法最高水平的准确度为 75%，在该数据集上还有较大的性能提升空间[3]；
- UCF101 数据集上，TLE 方法达到最高水平的准确率为 95.6%[4]。

四、 当前主要研究方向

- 大规模多标签视频分类与标注（large-scale multi-label video classification / annotation）
- 视频的时间/序列模型和池化方法（temporal / sequence modeling and pooling approaches for video）
- 时间注意力模型机制（temporal attention modeling mechanisms）
- 视频描述学习，如分类性能 vs.视频描述符大小（video representation learning e.g., classification performance vs. video descriptor size）
- 多模型（声音-视觉）建模和融合方法（multi-modal (audio-visual) modeling and fusion approaches）
- 从噪声/不完整的人工标注标签中学习（learning from noisy / incomplete ground-truth labels）
- 多重实例学习 multiple-instance learning (training frame-/segment-level models from video labels)
- 迁移学习，领域适应和泛化（transfer learning, domain adaptation, generalization）
- 衡量：性能 vs.训练数据和计算量（scale: performance vs. training data & compute quantity）

五、 相关会议和期刊

1. ICCV：International Conference on Computer Vision
2. IJCAI：International Joint Conference on Artificial Intelligence
3. AAAI：American Association for Artificial Intelligence
4. CVPR：Conference on Computer Vision and Pattern Recognition
5. ICML：International Conference on Machine Learning
6. ICLR：International Conference on Learning Representations
7. IJCV：International Journal of Computer Vision
8. ECCV：European Conference on Computer Vision

六、 小结

当前视频分类的主流方法主要是深度学习方法。**这些方法主要源于图像和语音领域中流行的深度模型。**视频数据的复杂特性，包括大量的空间、时间和音频信息，使得现有深度模型不足以处理视频相关任务。这使得强烈需要新的模型来有效获取视频的空间和音频信息，最重要的是建模空间的动态过程。除此之外，训练 CNN/LSTM 模型需要大量带标签的数据，这些数据通常昂贵，并且获取耗时，因此，**充分利用未标注的数据和丰富的上下文信息建立更好的视频描述模型是**

一个很有希望的研究方向。

七、 附录-术语表

Video classification：视频分类
Video captioning：视频描述生成
Clip：片段
Volume：域
Frame：帧
spatial-temporal：时空
sequence：序列
appearance：表观
video representations：视频描述
feature representation：特征描述
descriptor：描述符
frame-level：帧层
video-level：视频层
segmentation:分割
hand-crafted feature：人工设计的特征
state-of-the-art：最高水平
off-the-shelf：现有
Untrimmed Video Classification：videos can contain more than one activity 暂未找到合适的中译文
Trimmed Activity Classification：a trimmed video clip that contains a *single* activity instance 暂未找到合适的中译文
图像表示：Image Representation
运动检测与跟踪：Motion Detection and Tracking
边缘：edge
图像分割：Image segmentation
纹理特征提取：feature extraction
局部特征：local features
人工标注：Ground-truth
自动标注：Automatic Annotation
运动检测与跟踪：Motion Detection and Tracking

[1] Wu Z, Yao T, Fu Y, *et al.* Deep Learning for Video Classification and Captioning. arXiv:1609.06782, 2016.
[2] Wang H, Schmid C. Action recognition with improved trajectories[C]. ICCV, 2013,
[3] Lan Z, Zhu Y, Hauptmann A G. Deep Local Video Feature for Action Recognition. arXiv:1701.07368, 2017.
[4] Diba A, Sharma V, Gool L V. Deep Temporal Linear Encoding Networks. arXiv:1611.06678, 2016.