Classifying Basketball Plays

吴浩贤 1600012771 朱文韬 1600012785

Introduction

视频语义理解是计算机视觉中一个重要的的话题，其中体育比赛是一个重要的分析对象。其特点主要在于：大量高质量的视频，明确的实际应用，以及视频内容高度模式化。以篮球比赛为例，从解说员对比赛内容的描述不难发现，一场比赛中发生的事件不过传球、抢断、投篮（1/2/3分，成功/失败）等十几类。因此，对体育比赛视频的分析理解很大程度上就归结到对每个回合的识别（Recognition ）或分类（Classification ）这一基本任务，即将比赛回合的视频片段归到相应事件所属的类中。

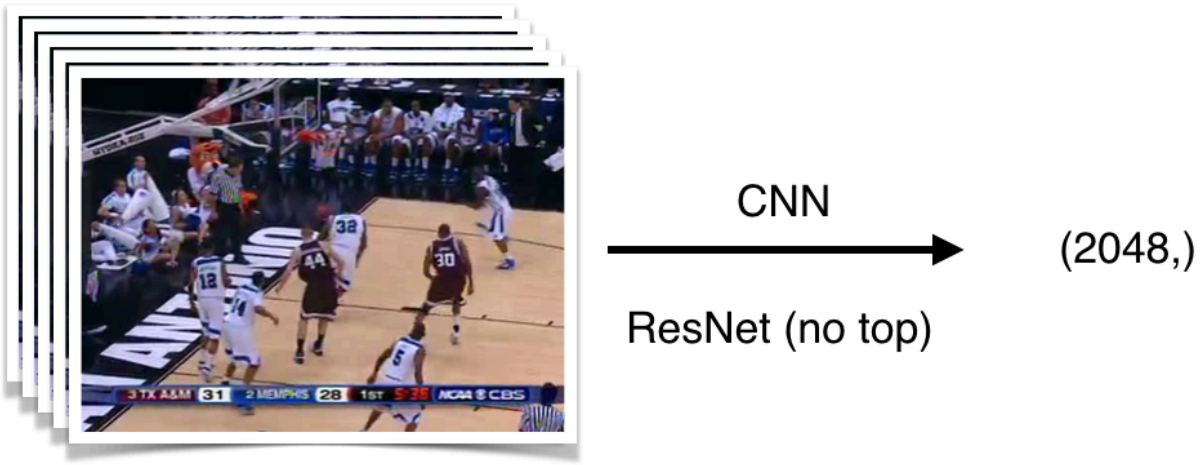
我们采用的数据集是Google Basketball Dataset，来源于2016年的CVPR Paper Detecting events and key actors in multi-person videos。数据集包含1988年-2011年251场NCAA赛事全国转播视频的14238个回合片段，分属7个事件类别。其中，人工标注了所有事件的结束时刻，采用向前采用一定帧数的方法获得事件样本（实际意义：比赛中比分变动/球权变动等回合结束信息是易于获得的）。此外，数据集还提供了球员位置信息（人工标注+KLT Tracker）和回合开始时球的位置信息供选择使用。数据集本身的挑战主要来自于：

- 大多数视频拍摄于20年前，因此清晰度低，噪声大。
- 不均衡的分类。篮球赛事本身特点决定了类间样本数有巨大差别，即使在做了更均衡合理的类别划分后样本最多和最少的类相差依然达到十倍。
- 真实转播拍摄视角变化大，由多个变化的摄影机位剪辑而成，且往往只能包含场上的一部分运动员，直接跟踪难度大。

Implementation

考虑到模型表现、计算资源等因素，我们采用的模型基于LRCN(Long-term Recurrent Convolutional Networks)，即先对时间轴上各帧采样并用卷积神经网络提取视觉特征，再通过循环神经网络刻画时间序列信息并给出分类。LRCN作为一种较为通用的处理视频信息的模型，当在我们面临Fine-grained和高度模式化的任务时，考虑在这个基础的模型中以合适的方式引入球和球员的信息。

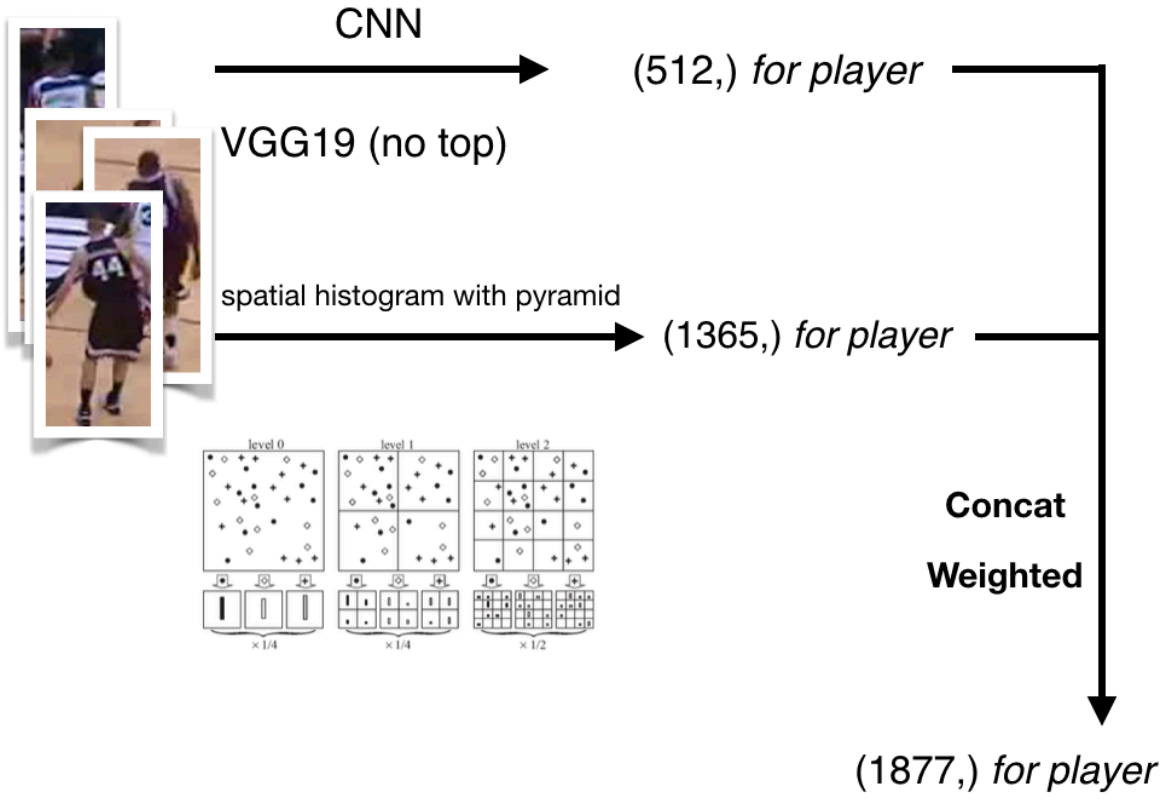
Feature Extraction



1. **Frame Feature**：整个场景的画面细节特征提取：从事件结束时刻向前每隔0.2s取1帧，一共抽取十帧，每一帧过基于`ImgeNet`预训练的`Resnet50`，取过了`max-pooling`后最后一层`fc`的2048维输入，作为当前时间的`frame_vec`（如上图）。
2. **Player Feature**：球员的特征对区分回合种类有重要意义，不仅在于球员的动作姿态，更在于球员间的空间位置关系。此外，同一时刻不同球员对回合的影响不可能是完全相同的：例如，控球队员和直接防守球员的特征对区分回合种类的意义显然比远离球的旁观队员更显著。因此，我们对每个

球员提取两种视觉特征，并关于球员和球的距离负相关关系做了加权处理（平方反比）：

- Appearance：根据`player_boundingbox`截取当前帧的`player_crop`，并送到基于`ImageNet`预训练的`VGG-19`，取过了`max-pooling`后最后一层`fc`的512维输入作为`player_appearance`
- Spatial：每个运动员的boundingbox位置信息抽取出1365维`spatial histogram`和`spatial pyramid`。`spatial histogram`将帧分成32*32，根据boundingbox 占每一块的面积作为直方图数值，`spatial pyramid`是抽取多尺度下的空间信息。（如下图）。



最终，我们对每帧得到2048维`frame_vec`，对每个球员得到1877维`player_vec`，对球的位置加权归一化作为每帧球员信息的总体表示，以应对球员人数变化的问题并反映待分类事件发生时球场的关注焦点信息。

Ball Detection

篮球的运动是视频事件中最直接和显著的特征，我们可以获得篮球位置作为特征并指导`Player Feature`的加权。

无监督方法

考虑到篮球的形状、颜色等相对固定且有明显的特征，采用方法：Canny 边缘检测 + Hough变换 + OpenCV object-tracker。Hough圆形检测出来多个候选区域后，通过限制尺寸、模板匹配找到最佳候选圆。

监督学习方法

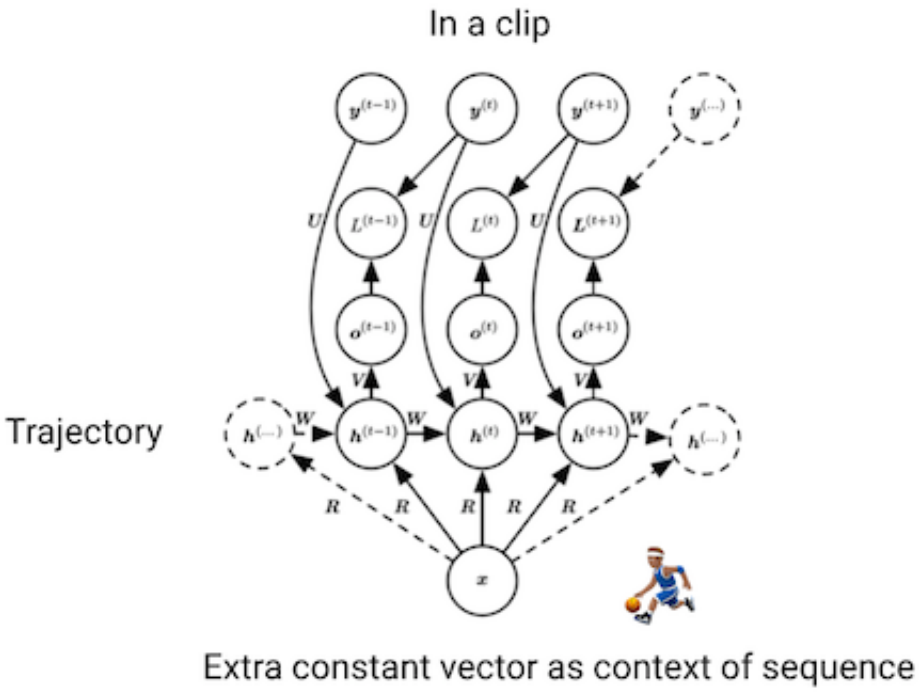
尝试用在COCO数据集上训练好的开源目标检测代码及模型检测篮球，没有取得很好的效果（～30%准确率）。鉴于数据集的特殊性，对该数据集单独采样并手工标注篮球区域进行训练。考虑到速度和标注数据量要求等，我们采用YOLO(You Only Look Once)神经网络结构，在手工标注数据集（200张用于训练）上达到55%测试集准确率。

Model Architecture

基于对问题不同的假设和强调，我们实现了多种不同的循环神经网络模型，并得到较符合预期的结果。

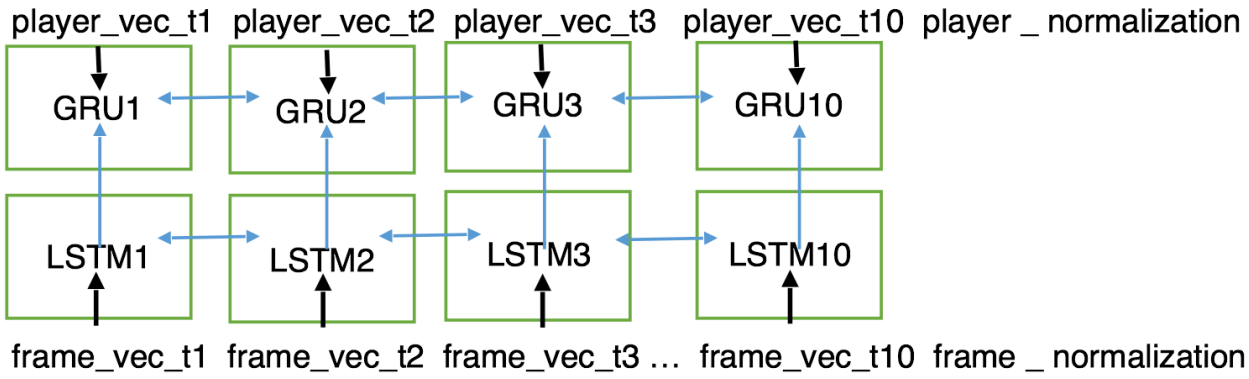
Context Model

在事件发生的较短时间（2秒）内忽略`player_vec`的变化，而是作为`frame_vec`的Context纳入模型中。采用类似Image Caption的经典模型中的网络结构（如下图）。



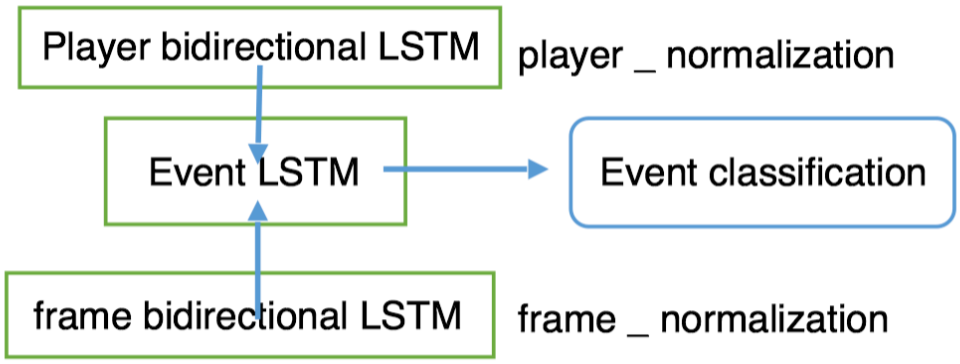
Player GRU Model

对`player_vec`单独训练一个RNN（GRU），用GRU的前后向hidden_state和刻画`frame_vec`的LSTM concatenate起来过fc层再过softmax分类器 。



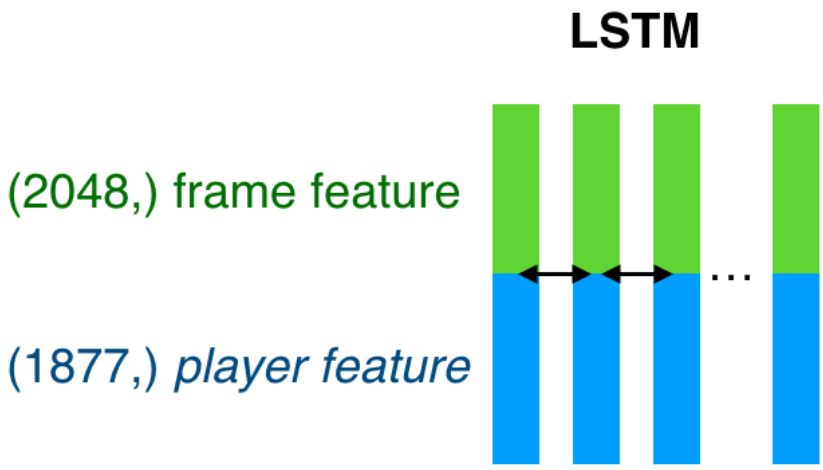
Concat Model

对`frame_vec`和`player_vec`分别经过双向`LSTM`获得clip-level的信息感知后再拼接经过`Event LSTM`给出最终分类（如下图）。



Unified Model

在每一个采样时刻，都将归一化的`frame_vec`和`player_vec`直接concat在一起，在统一的循环神经网络中训练（如下图）

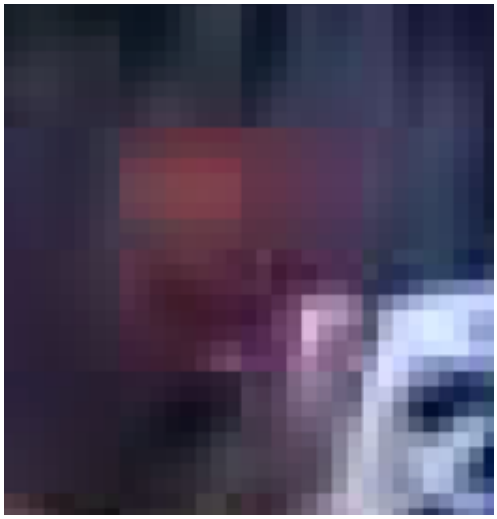


Results

Ball Detection

Hough变换对于形状要求比较苛刻，而且球在复杂背景下边缘模糊，球在快速运动的过程中，形变色变严重。只有当球没有被遮挡，以及球和背景对比度明显的时候才有良好的检测和跟踪效果。

目前ball tracking有一些基于抛物线模型的轨迹追踪，但是必须是在球投出去的时候有良好的效果，对于我们的数据集而言，由于球的像素比较低，以及很多时候球都被遮挡，碰撞频繁，运动不规则，因此没有很好的应用空间（如图）。



监督学习的方法在单帧图片上达到了一定的准确率，但应用于画质较低的视频片段时不足以产生连贯的轨迹坐标流。同时，因为目标物体同时具有模糊、高速、体积小三个严重影响追踪的特征，使得许多成熟的检测算法也不能达到这一目的。因此，我们最终选取事件开始的球的坐标作为球员加权依据，而没有将球的轨迹信息显式地纳入特征。

Classifying Results

我们测试了这些模型的精确度和Top2精确度，并和Baseline方法作了比较：

	Acc	$Top2Acc$
LRCN	0.35	0.59
Spatial Only	0.44	0.69
Context	0.47	0.70
Concat	0.49	0.73
Player GRU	0.53	0.80
Unified	0.57	0.83

得到的主要结论有：

1. 我们引入的与视频特定语义下的特征的确提高了LRCN模型结构下对篮球比赛回合的分类能力。
2. 对问题加入较强的假设后，结果与只利用提取到的球员位置信息没有太大的区别，显著低于其他方法（`Spatial`）。
3. 训练神经网络时将单向改为双向，可以更整体地认识事件过程，普遍提高2-3个点；其他训练手段如Batch Normalization，Gradient Clipping也对模型表现和稳定性带来显著提高。
4. 不均衡的类划分及相似的类影响明显：训练样本最多的一个类`2-pointer failure`的$(Acc, Top2Acc)$可以达到$(0.7, 0.9)$以上，而影响了相似类`2-pointer success`（尽管样本绝对数量不低）的Acc。同时，后者具有较相当高的$Top2Acc$。$Top2Acc - Acc$主要包含了对分值和命中情况中有一项判断错误的情形，体现相似类的区分能力。

Future Works

1. 考虑对每个球员单独构建一个RNN，再和`frame_vec`结合。这样就实现了更好的球员信息跟踪捕捉，能更精细地衡量回合中画面重点的变化。
2. 如果球在每一帧的定位难以实现，那么可以考虑给每一个球员添加一个

attention机制，自适应地调整球员信息的重要权重。
3. 对不均衡的类划分，考虑对小样本的类进行上采样（对视频数据增强）。
4. 在对球轨迹信息的获取上，通用方法不能起到很好作用，可以加入特定语义下的特征如`player_vec`辅助训练。

Comments

由于关于Paper和数据集本身没有开源代码，因此整个过程是从头实现的。其中YOLO的实现参考了开源框架Turi Create和Darknet。

吴浩贤主要负责：特征提取、无监督的篮球检测算法、Concat和Player GRU的模型搭建和调节

朱文韬主要负责：数据集处理、有监督的篮球检测算法、其他网络模型的搭建和调节

References

1. Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR2015

   Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell

2. Detecting events and key actors in multi-person videos, CVPR 2016

   Vignesh Ramanathan, Jonathan Huang,

   Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, Li Fei-Fei

3. Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR 2015

   Karen Simonyan, Andrew Zisserman

4. Deep Residual Learning for Image Recognition, CVPR 2016
   Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun
5. You Only Look Once: Unified, Real-Time Object Detection, CVPR 2016
   Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi