2018 年秋-第 15 周

一、数据集

1、HollywoodHeads dataset：21 部电影的视频剪辑。原始数据集为所有帧提供头部注释。

2、Extend 处理：结合检测来形成轨道；包含 331746 个与头部轨道的完整注释的视频帧，分割成 216719，67181 和 47846 的帧分别用于训练，验证和测试

3、用于比较评估以前工作：保留含有 1 302 个注释帧的测试集。

二、生成 tube proposals（两种方法）：

第一种：Tube Proposal Network（TPN）

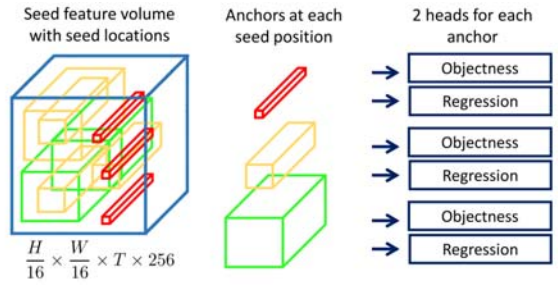

Figure 3: Architecture of Tube Proposal Network (TPN). Tube anchors are plotted in red, yellow and green.

管提案的定义：连续视频帧上的区域建议序列，只考虑与均匀线性运动相对应的提议。

结构：

将一大块连续帧作为输入并产生一组管道建议，每个管道建议都有一个对象得分和一个精致的区域位置。

1、CNN 特征提取：对输入的每一帧提取特征，最后沿时间堆叠

2、再送入 3 × 3 × 3 体积循环层（conv3D）来形成特征卷

3、种子位置：特征卷的所有空间位置；种子特征：空间位置相应的特征，与 k 个参考 tube（tube anchors）相关联。在特征体积中共享相同中心轴的 tube anchors 对应于相同的种子位置以及相同的种子特征向量。

4、每个种子特征向量通过 anchor network（两个全连接层），对 K 个 tube anchors 每个产生两个输出：对象性得分和管回归的参数

5、收集所有 tube anchors 的回归位置作为 tube proposals，每个 proposal 有一个 anchor 对象性得分。

6、修剪 proposals：基于对象性得分应用 NMS（使用 tube 重叠率）

训练：

1、 在 mini-batches 上训练

2、 tube anchors 空间位置固定，具有不同比例和纵横比

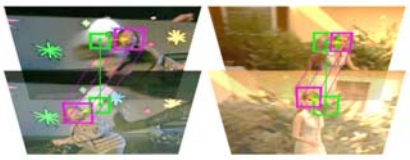

Figure 4: Examples of TPN proposal. Examples of TPN proposals. Tube anchors and final proposal are plotted by green and magenta respectively.

这个图表明 TPN 可以在没有运动的情况下将 tube anchors 回归到有动作的 tube proposals，tube anchors 和最终 proposals 分别用绿色和品红色绘制。

3、 tube anchors 的标记（原理与下面的 Tube-CNN 相同）：

$\theta_{GT} \geq 0.5$：正（区别在于此处为任意类）

$\theta_{GT} \leq 0.3$：负（背景）

第二种：Tube proposals by tracking box proposals

1、输入含 T 连续帧的块，用 Selective Search（作者实验中用 faster rcnn）在起始帧生成 box proposals

2、使用 Kanade-Lucas-Tomasi (KLT)跟踪器来获得块的起始帧和结束帧之间的点轨迹

3、每个 box proposals 与一组内部点轨迹相关联，对应于一组移动方向。然后将这些方向聚集成 N_e 个方向箱

4、在方向箱里面找 $N_{k\#}$ 个最密集的箱并构建 $N_{k\#}$ 个候选时间路径。

5、沿着假设的时间路径线性地移动箱来构建管提议。

三、Tube-CNN for object detection

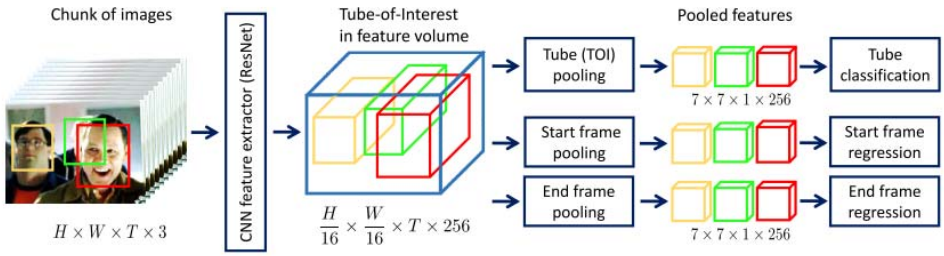

Figure 2: Architecture of Tube-CNN for object detection. The input is a chunk of images and a set of tube proposals. The model starts with the CNN feature extractor to get a spatio-temporal feature volume. Afterwards the model splits into the three branches: tube classification (starts with Tube-of-Interest, TOI, pooling), regression on the start and end frames, which begins with the frame-level Region-of-Interest pooling, ROI.

1、 架构

由三个主要块组成：CNN 特征提取，管分类和管回归。

CNN 特征提取：对输入的每一帧提取特征，最后沿时间堆叠

不懂之处：To achieve good performance, we reuse the idea of sharing computations from first CNN layers among all proposals [13, 10]. For the architecture on this block, we can

管分类：

把 tube proposals 的坐标映射到 Tube-of-Interest in feature volume 上面的一个子卷；

对每一帧进行最大池化到一个固定尺寸；
最后把这些 feature map 连接；
将连接后的 feature map 块送入分类网络的最后一部分（如 resnet101 的第四部分）；
分类采用交叉熵损失

管回归：

由两个网络组成，分别用于在 tube 的两端进行边框回归
对起始帧的 feature map 池化，然后采用 L1 损失回归
结束帧同上

2、 监督

Tube-CNN 输入：一段 T 连续帧和一组管提案

标签分配：tube proposal 和 groundtruth tubes 之间的管道重叠

Overlap 定义：分别求两个 tube 对应的两端（起始、结束）的 IOU，取最小，记为 θ_{GT}

那么：

$\theta_{GT} \geq 0.5$：正（类标签）

$0.1 \leq \theta_{GT} < 0.5$：负（背景）

附录

深度学习词汇：

1、batchsize：批大小在深度学习中，一般采用 SGD 训练，即每次训练在训练集中取 batchsize 个样本训练；

2、iteration：1 个 iteration 等于使用 batchsize 个样本训练一次；

3、epoch：1 个 epoch 等于使用训练集中的全部样本训练一次；

4、mini batch：把 100 万样本分成 1000 份， 每份 1000 个样本， 这些子集就称为 mini batch。在 mini batch 我们在一个 epoch 中就能进行 1000 次的梯度下降，而在 full batch 中只有一次。