本周先对 R-C3D 进行了进一步的总结，然后看了 Rethinking the faster r-cnn architecture for temporal action localization 这篇论文。

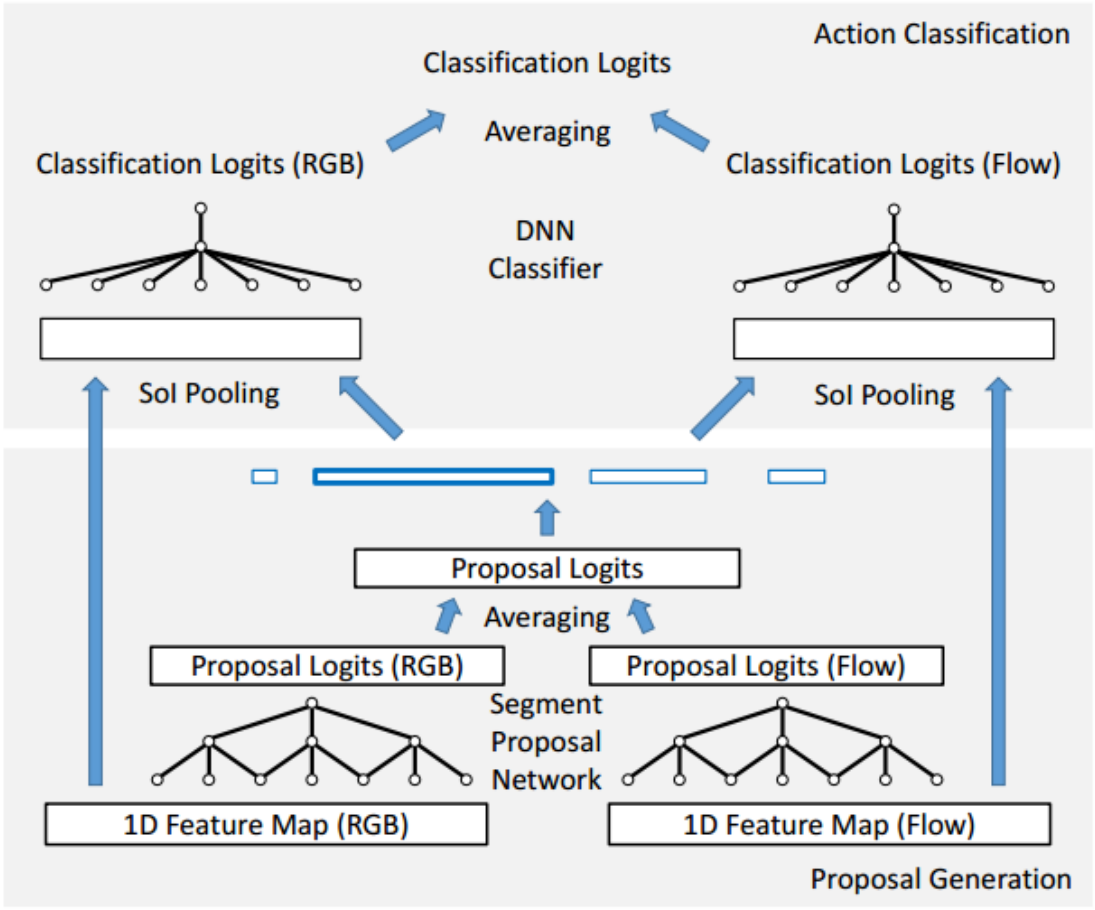
R-C3D 与 faster RCNN 区别与联系: R-C3D 是基于 faster RCNN 对目标检测的思想用来进行 temporal activity detection。和目标检测类似，R-C3D 将输入数据通过 C3D 网络生成需要的 feature map，通过两个阶段生成最终需要的图片序列以及对应的类型。第一个阶段产生 proposal 来确定 anchor 是否为正负样本，第二阶段对第一阶段得到的 proposal 进一步判断其类型并对 bounding 做回归处理，每阶段均有两组损失函数需要计算。R-C3D 是对一维的特征向量进行处理，faster RCNN 是对二维特征向量进行处理。因为有类似于 ROI pooling 的思路，R-C3D 可以对不同长度的视频序列进行处理。在训练网络时，faster RCNN 给出了详细的 4 步训练的方法达到 end to end 的目的，R-C3D 在 C3D 网络部分用 pretrain 的参数进行初始化，然后对整个网络所有参数进行训练。

TAL-Net (temporal action localization)，通过对处理二维图像的 faster rcnn 网络的变化来对图片序列进行处理，思路是将输入的三维图像序列通过 CNN 网络得到一维特征序列，通过 SPN(segment proposal network，和 RPN 类似结构)得到 proposal，通过 SoI pooling 后输入到 classifier 得到 localization 和 class 的结果。作者在此基础上提出了三点改进方案：

1.Receptive Field Alignment: 视频序列的长度会有很大的变化范围（几秒到几分钟），如果直接像 faster rcnn 一样使用给定的一些 anchor 去找 proposal 不一定能取得好的效果。文中提出了 multi-tower network，dilated temporal convolutions 的方法。针对不同的长度的使用不同的卷积进行处理，这些卷积均为类似的塔型结构，并使用空洞卷积（采样点之间有间隔，在保持参数个数不变的情况下增大了卷积核的感受野）进行处理。关于塔形结构，先将输入的一维 feature map 经过 maxpooling 平滑，再经过两层 kernel size 为 3 的卷积，每经过一个卷积会用一个 kernel size 为 2 的 pooling 层来使感受野随塔形结构层数增加呈指数级增加（若只用卷积层感受野随层数增加呈线性增加）。最后通过两个并行的 szie 为 1 的卷积层做 classification 和 boundary 的回归。

2.Context Feature Extraction: 文中认为正样本的前后部分会有一些对判断 proposal 有效的信息，因此把 SPN 部分的感受野扩大了一倍，头尾各加 proposal 的一半长度，经过一个 size 为 7 的 SoI pooling 和一个 FC 层，最后进行回归。

3.Late Feature Fusion:



大部分行为识别网络通过 two-stream 结构来得到更好的结构，文中假设 rgb 和光流得到的特征对 TAL 一样有效，于是设计了上图所示的结构，最后的 averaging 对两个 logits 进行元素值平均，这种方法比在生成一维向量序列时将二者进行 concatenate 得到的效果更好。