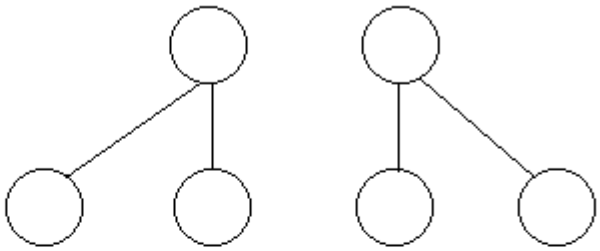
本周看了 CDC(Convolutional-De-Convolutional)网络的论文，因为有些课开始准备期末作业和考试了，所以进度这方面进度比较慢。

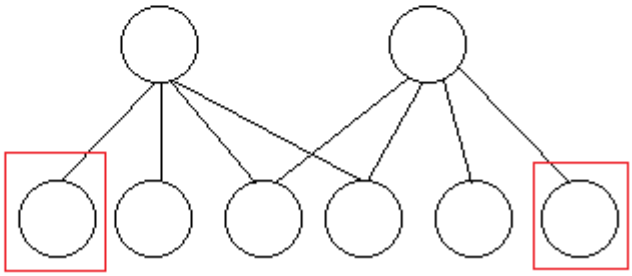
本文主要提供了一种思路，在 frame-level 进行 spatial action recognition，即输出有输入视频帧数个特征向量，每个特征向量分别表示对应帧所对应的类别。与单独训练每张图不同的是采用了 C3D 网络来对一个视频段进行训练，再通过反卷积操作使输出特征量符合输入视频的尺寸。

1D 反卷积: 反卷积里面的 size, stride, padding 均指的卷积方法中的各个变量，也就是说反卷积得到的 feature map 通过相同 size, stride, padding 的卷积能得到反卷积前的 feature map。

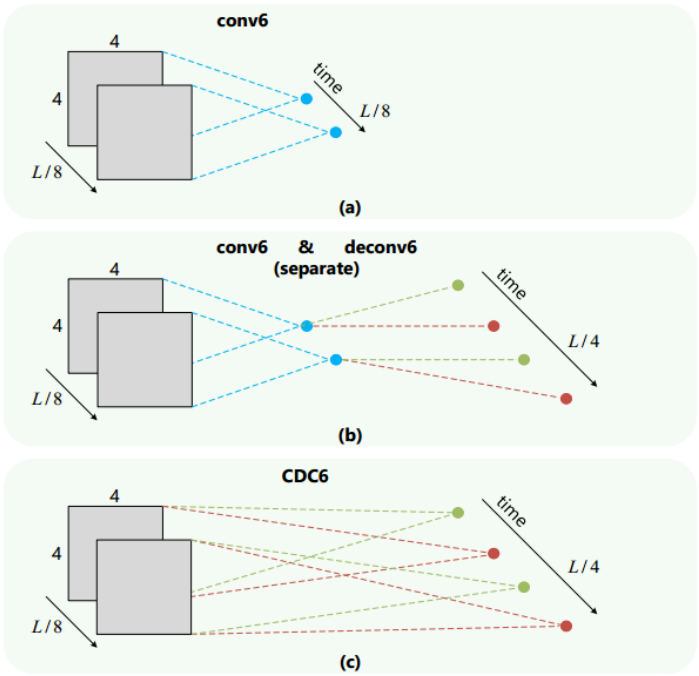
size=2, stride=2,
padding=0



size=4, stride=2
padding=1



如图所示上图为最基本的 2d 反卷积，下图红框为 padding 的特征，在进行完反卷积操作后会消除，两者最终都将使原来深度为 L 的 feature maps 扩大到 2L 长，后者中每个 feature map 都会由两个原 feature map 共同得到，文中采取的网络使用的是后者的反卷积算法。关于卷积和反卷积同时进行，则是用下图(C)的模式，每个 4*4 的 feature map 通过 4*4*2 个权重得到两个输出的点，若用上图 padding=1 的模式每个输入的 feature 会连接 4*4*4 个权。



损失函数:

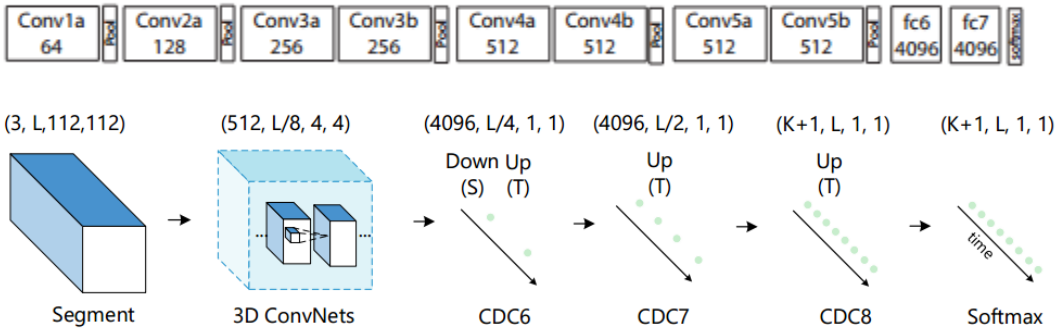
$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{L}\left(-\log\left(P_n^{(z_n)}[t]\right)\right), \qquad P_n^{(i)}[t] = \frac{e^{O_n^{(i)}[t]}}{\sum_{j=1}^{K+1} e^{O_n^{(j)}[t]}}$$

i 表示第 i 类视频，n 表示 minibatch 中第 n 个 segment，t 表示 segment 中第 t 张 frame，此处没有采用交叉熵损失函数而是采用指数损失函数，当得到的正确类的置信度越接近 1 时 $P_n^{(i)}[t]$越接近 1，从而其 log 值越接近于 0，其损失函数也趋向于 0。

CDC6,CDC7 用 FC6,FC7 初始化: 在训练权重初始化时，3D ConvNets 部分直接采用 C3D 部分的预训练权值，CDC6 和 CDC7 采用 FC6 和 FC7 的权值进行初始化，如上图(a)(b)所示，蓝点的数量与 C3D 网络的 FC6 层一样均为 4096 个，因此每个绿点和红点用对应的蓝点做权值相同的初始化。C3C7 层同理。

训练: 通过一个 32frames 的 sliding window 来框出输入视频序列，且确保每个 sliding window 中至少有一帧为正样本（class≠0），CDC8 以前的网络用 0.00001 的 learning rate，CDC8 本身用 0.0001 的 learning rate 进行训练，CDC8 使用随机初始化值，整个网络收敛速度会比较快。

测试: 可以通过得到每个 frame 的 score 设置一个门限将满足条件的 frame 合并成同一类，但这种方法鲁棒性差，文中采取了用其他论文中的方法来得到大致的 proposal，再通过 CDC 网络进行精确定位。将得到的 proposal 首尾各延长原长度的 1/8 倍，通过 slide windows 得到每张图的 score，且 slide windows 之间没有重叠，每个 slide window 都与至少一个 proposal segment 相交。之后通过取每一个 proposal 的平均置信度的最大值来确定这个 proposal 的类别，若这一类不归为背景，则进一步对其 boundary 进行回归。通过对 proposal segment 上预测 class 的置信度进行高斯核密度估计(Gaussian kernel density estimation)得到均值和标准差，从扩大后的 proposal segment 两边向中间依次计算，找到置信度不低于（均值-方差）的置信度的 frame 作为最终有标签视频的头尾。



Pipeline: CDC 网络使用 C3D 网络的模板进行修改，3D ConvNets 部分直接使用 C3D 的 Conv1a 到 Conv5b 提取视频特征，输入视频的长度只受显卡内存影响，size 为 112*112，经过 3D ConvNets 得到 512*(L/8)*4*4 的 feature maps，为了得到最后(K+1)*L*1*1 ((类别标签+背景标签)*原视频长度*H*W)，对 4*4feature maps 做卷积得到 1*1 的 feature maps，将 L/8 通过三次反卷积得到长度为 L 的 feature maps。CDC6 步骤同时进行 2d 卷积和 1d 反卷积，而不是先卷积再反卷积，CDC7 再进行一次 1d 反卷积，CDC8 做 1d 反卷积的同时做类似 FC 的操作，将 4096 维变换为 K+1 维，最终在每一个 frame 上进行 softmax 得到最后输出。