基于嵌入式平台的人体特征识别项目成果简介

1. 项目背景

人体视频动作的识别是计算机视觉研究一个很重要的方向，目前大多数算法针对的是处理一个完整的动作视频，像训练的时候一样。然而很多实际应用场景比如自动驾驶，智能安防、家居都要求我们实时处理在线的视频流，并且能够实时给出预测结果。一些较大的模型就不能很好的发挥效果。需要一种新的方法来完成在嵌入式平台上的视频动作分类任务。

1. 研究目标

1. 在嵌入式平台部署一个人体行为识别模型。

2. 该模型尽可能小，在运行中使用尽可能少的资源。

3. 能够在线处理视频流，并达到低时延，高准确率的结果。

1. 项目创新与成果
2. 数据集处理

我们以jester手势视频动作数据集为例进行训练。

通过分析jester 数据集的特点，我们发现大多数动作是37帧左右，并且动作集中在中间12-26帧，于是我们把从中间隔选取8帧的算法变成： 把一个视频动作分成20%，60%，20%部分，然后分别从这三部分随机抽取1，6，1个帧。

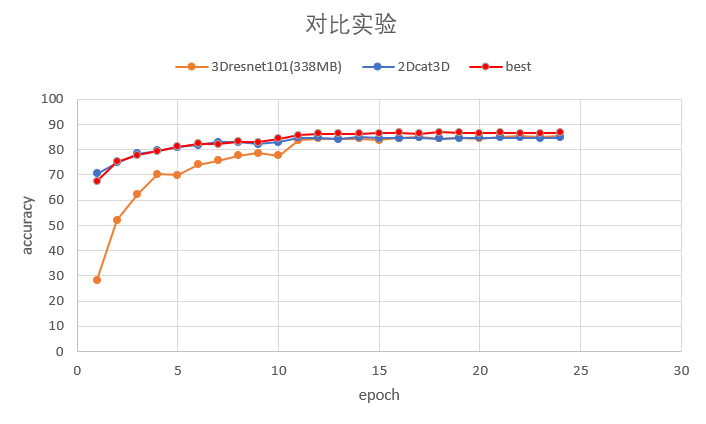
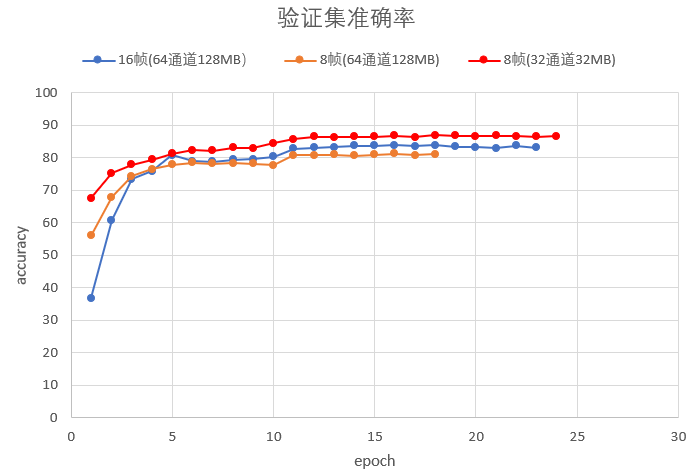
1. 模型

采用2D卷积+3D卷积的框架

1. 2D部分采用2D Resnet18，截取到layer2之前的部分用来提取图像的feature map。为减少嵌入式内存占用，这里删除了第三层下采样部分，直接输入3\*122\*122图像。输出大小128\*28\*28。
2. 连接部分用1\*1 的卷积核降维到32\*28\*28，以减少3D卷积核参数。
3. 采用ECO模型的思想，将N帧图片的feature map 按照时序叠加成32\*N\*28\*28的tensor 送入3D Resnet18 学习时序信息。
4. 我们更改了原论文中3D Resnet18 的结构，只截取layer1 到最后全连接的部分。并且layer1-layer4的3D卷积核的通道数减小一半到32，64，128，256。

这样我们的模型大小仅有32MB，预测准确率达86.7%。

下图展示了我们训练过程中最好的模型与研究过程中的其他模型的对比。



1. 部署
2. 取视频帧
3. 采用滑动窗口取样，窗口太长的会错过动作，或者造成延迟；太短的不能取得相对完整的动作时序信息。
4. 根据我们的模型输入，我们采用一个队列来缓存8帧动作图像。每次存入一个新的视频帧之前，先抛出时间最远的视频帧，使得队列始终保存最近的8帧。每当加入4帧后，进行一次预测。
5. 摄像头获取视频流的帧率是32fps，我们没隔4帧更新一次队列，这样队列8帧的跨度在1秒左右，能够较好地截取一个完整的手势动作。

