简介

Google出品，目前没有开源。作者提出一种轻量级的人脸检测模型，可以在旗舰手机芯片上跑到200-1000+FPS。其基本架构基于SSD和魔改的mobilenet。
作者对前置摄像头和后置摄像头构建了不同的模型。同时输出六个脸部关键点（分别是眼睛2，耳朵2，嘴巴中心1和鼻尖1）

模型设计思路

增大感受野：

作者首先注意到一个事实，在mobilenet当中1x1 conv也就是pointwise part占据了绝大部分的计算量。
一个形状为(s,s, c)的输入tensor在经过k x k的可分离卷积后，乘法和加法的总计算量为 s^2ck^2 。而后面的1x1卷积部分假设其channel为d,那么总的乘法和加法运算量为 s^2cd，因此两者相差系数 d/k^2 ，在原文中k=3。
对于一个形状为(56, 56 , 128)的tensor，在iphoneX Metal Performance Shaders框架下，16bit浮点运算，3x3可分离卷积核速度为0.087s，而1x1卷积（channel从128到128）速度慢了4.3x 0.3ms
作者认为增大卷积核运算量不会增大太多，所以将卷积核大小设置为5x5

调整res结构

MobilenetV2的residual block，中间的部分会扩大，而两边的部分会缩小，但是为了让depwith的部分channel小一些，降低运算量，作者又恢复到了原来resnet的residual结构。

double BlazeBlock

作者认为节省下来的计算量还可以够它再加一个5x5的blaze模块，这样感受野更大。

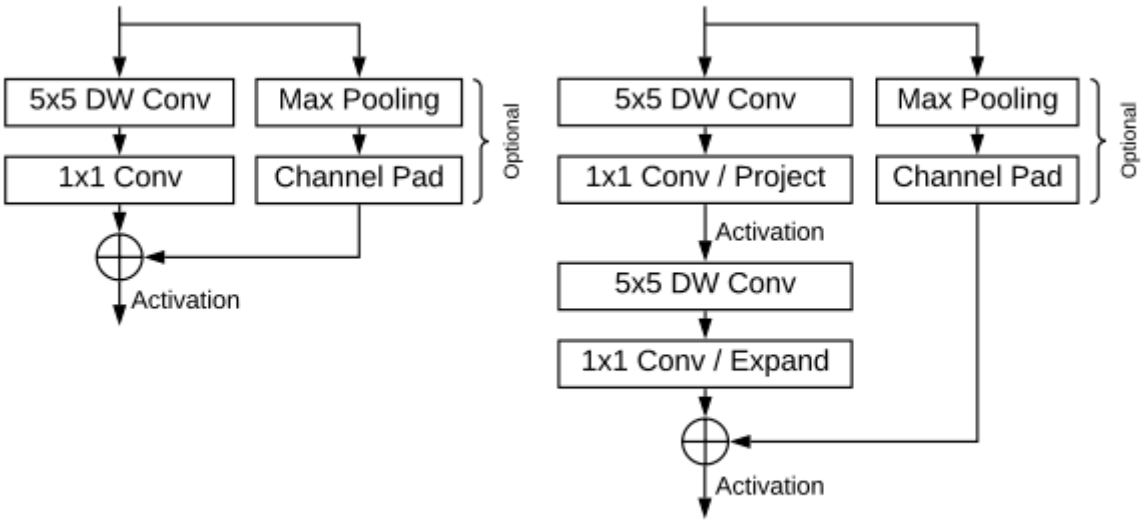

Figure 1. BlazeBlock (left) and double BlazeBlock

feature extractor

对于前置摄像头模型：

1. 输入尺寸128x128
2. 5 single blazeblock + 6 double blazeblocks
3. 最大channel depth=96
4. 最小spatial size 8x8

Appendix A. Feature extraction network architecture

Layer/block	Input size	Conv. kernel sizes
Convolution	128×128×3	128×128×3×24
Single BlazeBlock	64×64×24	5×5×24×1 1×1×24×24
Single BlazeBlock	64×64×24	5×5×24×1 1×1×24×24
Single BlazeBlock	64×64×24	5×5×24×1 (stride 2) 1×1×24×48
Single BlazeBlock	32×32×48	5×5×48×1 1×1×48×48
Single BlazeBlock	32×32×48	5×5×48×1 1×1×48×48
Double BlazeBlock	32×32×48	5×5×48×1 (stride 2) 1×1×48×24 5×5×24×1 1×1×24×96
Double BlazeBlock	16×16×96	5×5×96×1 1×1×96×24 5×5×24×1 1×1×24×96
Double BlazeBlock	16×16×96	5×5×96×1 1×1×96×24 5×5×24×1 1×1×24×96
Double BlazeBlock	16×16×96	5×5×96×1 (stride 2) 1×1×96×24 5×5×24×1 1×1×24×96
Double BlazeBlock	8×8×96	5×5×96×1 1×1×96×24 5×5×24×1 1×1×24×96
Double BlazeBlock	8×8×96	5×5×96×1 1×1×96×24 5×5×24×1 1×1×24×96

Table 4. BlazeFace feature extraction network architecture

其实不难看出作者的主要思想是降低了channel维度的大小，增加了卷积核的尺寸和个数，因为作者认为channel对运算量的影响更大，而卷积核的调整是比较efficient的。

anchor

经典的ssd架构其offer anchor的feature尺寸是1x1, 2x2, 4x4, 8x8和16x16。但是文章PPN证明了有些其实是多余的。
作者提出GPU和CPU的一个很大的区别在于GPU对于每层的计算成本是相对固定。（我的理解是，与其每个层都提anchor，不如一个层多提一些anchor） 作者再8x8之后就不再提anchor，而且aspect ration只有1，对人脸够用。作者会在8x8尺寸上提6个anchor，而不是在8x8,4x4，2x2的尺寸上各提2个anchor。

后处理

由于8x8的anchor个数比较多，所以bbox的重叠会比较严重，那么NMS在视频的检测当中会出现比较大的波动。
作者的想法是采用一种blending策略，将这些重叠的bbox做一个weighted mean来得到最后的回归参数。这种方法并不会增大NMS的额外运算量。

实验

Model	Average Precision	Inference Time, ms (iPhone XS)
MobileNetV2-SSD	97.95%	2.1
Ours	**98.61%**	**0.6**

Table 1. Frontal camera face detection performance

Table 2 gives a perspective on the GPU inference speed for the two network models across more flagship devices.

Device	MobileNetV2-SSD, ms	Ours, ms
Apple iPhone 7	4.2	**1.8**
Apple iPhone XS	2.1	**0.6**
Google Pixel 3	7.2	**3.4**
Huawei P20	21.3	**5.8**
Samsung Galaxy S9+ (SM-G965U1)	7.2	**3.7**

Table 2. Inference speed across several mobile devices