Google Blaze Face:

* Lightweight feature extraction network inspired by Mobilenet V1/V2.

* Produces 6 facial Keypoint coordinates (eye centers, ear fragions, mouth center and nose tip).

* On an $S \times S \times C$ input a $K \times K$ depthwise convolution involes $s^2 c K^2$ multiply-add operations.

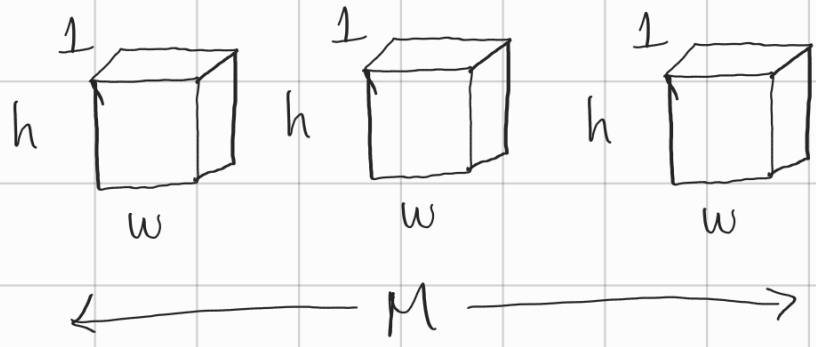
Feature Extractor :

Input: (128, 128, 3)

MobileNet Architecture:

* Depthwise Convolution:

Applies a single filter to each channel.



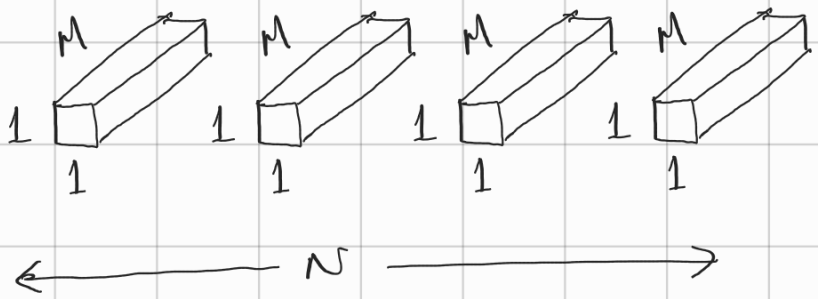
∴ h = height

w = width

M = # channels

* Pointwise Convolution:

Applies a 1 × 1 convolution among the input channels



∴ w = h = 1

M = # Input channels

N = # Output Channels

Model:

$F: \mathbb{X} \rightarrow \mathbb{Y}$

$\mathbb{X} \in \mathbb{R}^{h \times w \times ch}$, $\mathbb{Y} \subset \mathbb{Z}^{c}$

∴ h: height
w: width
ch: channels
c: classes

Input (224 x 224 x 3)

Architecture:

Layers:

- Depthwise Conv
- Pointwise Conv
- Avg Pooling (7x7)
- Fully Connected
- Softmax

- Depthwise:

$$\hat{G}_{k,l,m} = \sum_{i,j} \hat{K}_{i,j,m} \cdot F_{k+i-1,\, l+j-1,\, m}$$

where $\hat{K}$ → Depthwise Conv Kernel $(D_K \times D_K \times M)$

· Softmax:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

$$\sigma : \mathbb{R}^K \longrightarrow [0,1]^K$$

$\therefore \vec{z}$ = Input Vector

K = # classes

e^{z_i} = Exp Input

e^{z_j} = Exp Output

Note: Google's Blaze Face implements in Depthwise convolution a Kernel Size of 5x5