中国电信 CHINA TELECOM
世 界 触 手 可 及

中国电信移动终端研究测试中心
CHINA TELECOM MOBILE
DEVICE RESEARCH AND TESTING CENTER

KPI test for AI mobile device hardware requirements

FLOPS and FLOPs

- Floating Point Operations per Second (FLOPS) is a measure of hardware performance, used in fields of scientific computations that require floating-point calculations.
- Floating Point Operations (FLOPs) is a measure of computational complexity, and it can used for measuring algorithm or model complexity.
- When activation function is neglected, the computational complexity of convolution layer：

$$(2 \times C_i \times K^2 - 1) \times H \times W \times C_o$$

  - Where, C_i=input channel, k=kernel size, HW=output feature map size, C_o=output channel.

- When activation function is neglected, the computational complexity of fully connected layer：

$$(2 \times I - 1) \times O$$

  - where, I=input neuron numbers, O=output neuron numbers.
  - 2: 1 Mac equals to 2 operations
  - -1: when bias is neglected, it is -1, and when bias is taken into accounted, it without -1

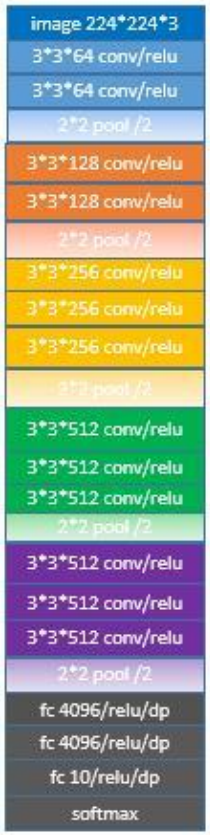
MAC and OPS definition

- MAC-Multiply-accumulate

- OPS - Operations of OPS only refers to multiply-accumulate(MAC) operations, not including input, output and other operations, and typically 1 MAC operation = 2 DL OPS; The number of MACs needed to compute an inference on a single image is a common metric to measure the efficiency of the model.

- OPS/W OPS per watt extend that measurement to describe performance efficiency.

The computational complexity of VGG16



MACC计算里总共 15.470GFlops 1.5470e+10 FLOPs

stage_1 224*224*3*3*3*64+224*224*64*3*3*64 =1,936,392,192 =**1.93*e+09**

stage2 112*112*64*3*3*128+112*112*128*3*3*128=2,774,532,096=**2.78*e+09**

stage3 56*56*128*3*3*256+56*56*256*3*3*256*2=4,624,220,160=**4.63*e+09**

stage4 28*28*256*3*3*512+28*28*512*3*3*512*2=4,624,220,160=**4.62*e+09**

stage5 14*14*512*3*3*512*3=1,387,266,048=**1.39*e+09**

fc6 7*7*512*4096=102,760,448=**1.03*e+08**
fc7 4096*4096=16,777,216=**1.68*e+07**
fc8 4096*1000=4,096,000=**4.10*e+6**

Three fully connected layers

参数138,38M
存储总共 527.79M = 138357544 *sizeof(float)/1024/1024=138357544*4/1024/1024=527.7921 M

stage1 3*3*3*64+64+64*3*3*64+64=38,720=**3.87*e+04**

stage2 64*3*3*128+128+128*3*3*128+128=221,440=**2.21*e+05**

stage3 128*3*3*256+256+256*3*3*256+256+256*3*3*256+256=1,475,328=**1.48*e+06**

stage4 256*3*3*512+512+(512*3*3*512+512)*2=5,899,776=**5.90*e+06**

stage5 (512*3*3*512+512)*3=7,079,424=**7.08*e+06**

fc6 512*4096*7*7+4096=102,764,544 =**1.03*e+08**
fc7 4096*4096+4096=16,781,312=**1.67*e+07**
fc8 4096*1000+1000=4,097,000=**4.10*e+06**

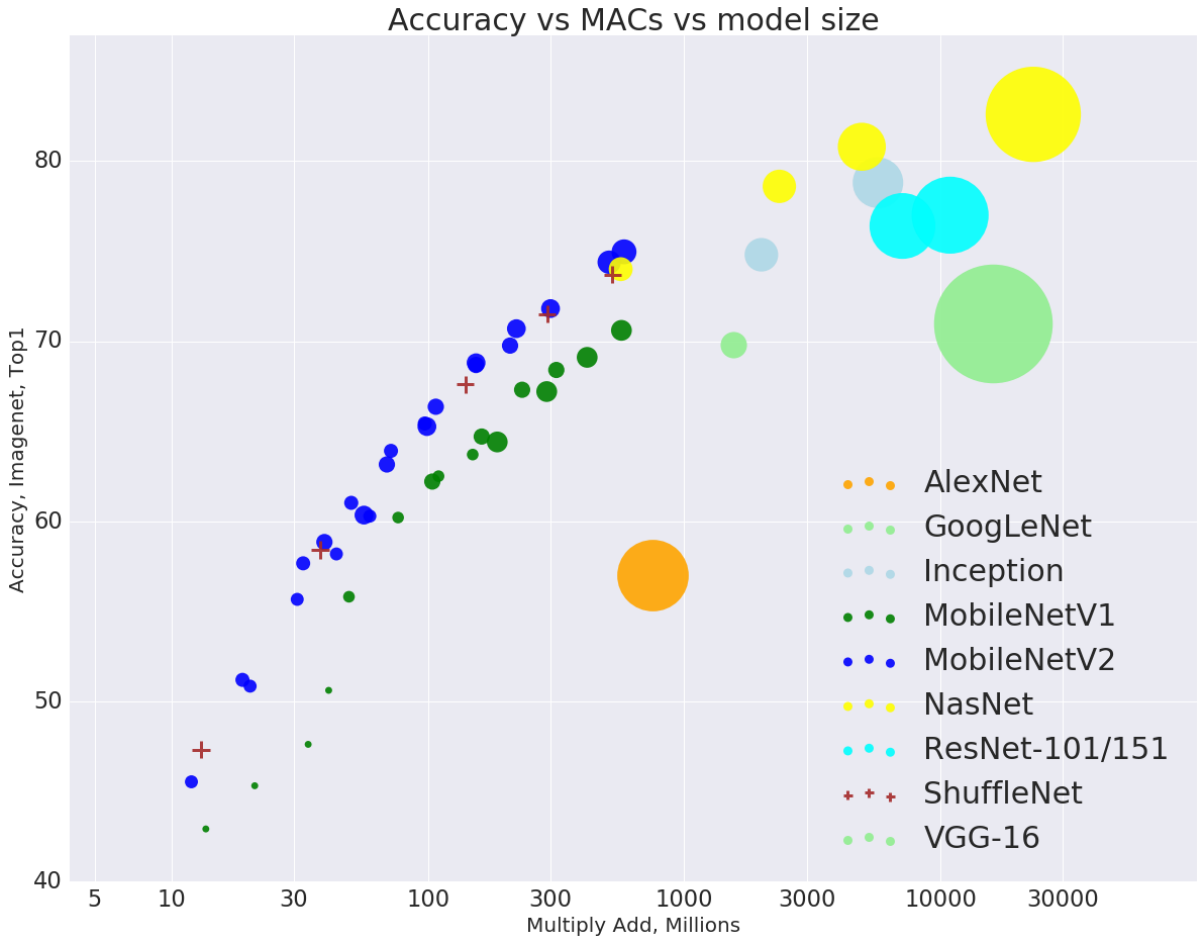
Note: The number shown in the figure is the MAC of VGG16, therefore its computational complexity is 2*15.47=30.94 Gflops = 0.0394 TFlops.

Neglect the last three fully connected layers, the computational complexity is 2*15.35=30.7GFlops=0.0307 TFlops

The more complex the model, the higher the accuracy

MACs, also sometimes known as MADDs - the number of multiply-accumulates needed to compute an inference on a single image is a common metric to measure the efficiency of the model.

Below is the graph comparing MobileNetV2 vs a few selected networks. The size of each blob represents the number of parameters. Note for ShuffleNet there are no published size numbers.



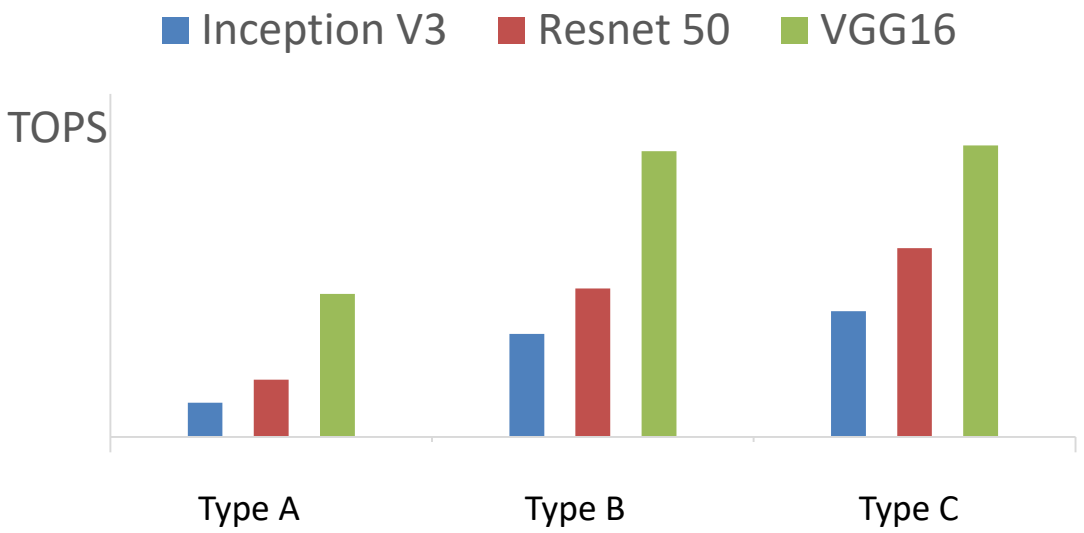
https://github.com/tensorflow/models/tree/master/research/slim/nets/mobilenet

Compared with other models, VGG test can better represent hardware performance

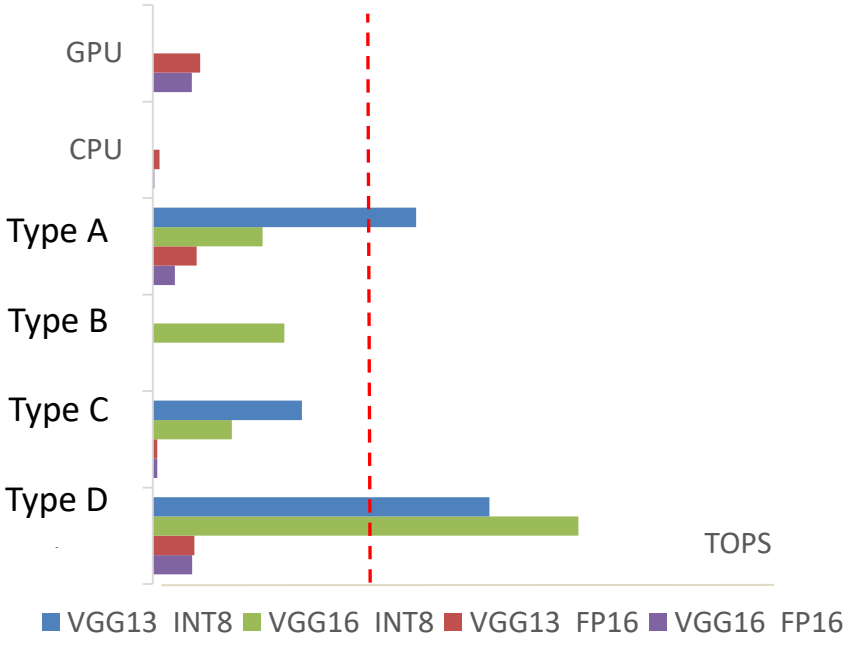
Model complexity in MAC

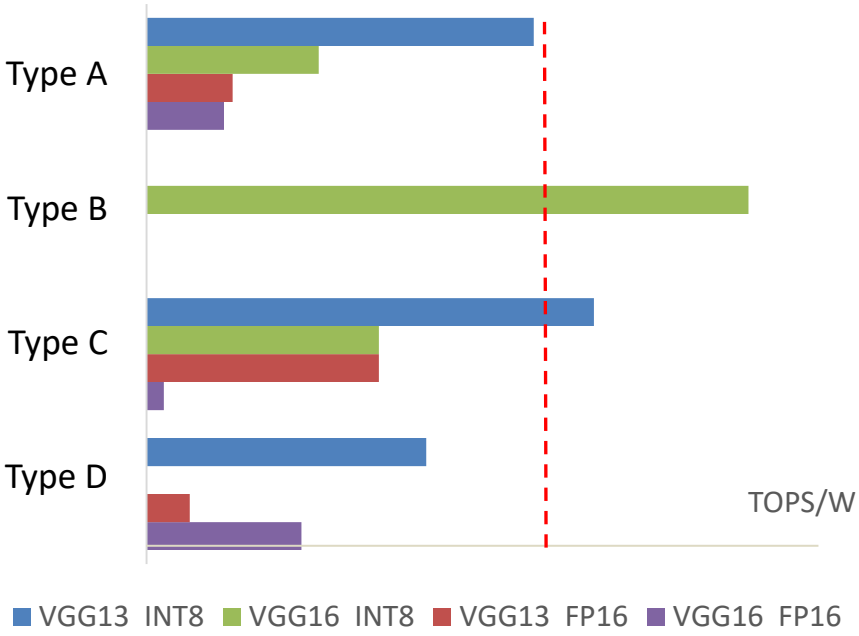
Incepion V3	2.85 GMAC
Resnet50	4.12 GMAC
VGG16 Proposed value	15.55 GMAC

OPS = (MACs of Model / inference time) x 2



Based on the following test data, VGG13 can better reflect hardware performance





Hardware requiements' KPIs

- Test method:
  - Remove the last three fully connected layers from VGG16 as a testing model.
  - Use ten different pictures of 224*224*3 as the input of the model.
  - Use power meter to measure AI mobile device's power.
- Calculation formula:
  - TOPS=0.0307*10/t, where t is the total amount of time to run 10 pictures in second.
  - TOPS/W=(0.0307*10/t)/P, where P is the average power to run 10 pictures in Watt.
- Proposed value：

	INT8	FP16
TOPS	1	0.5
TOPS/W	0.5	0.3