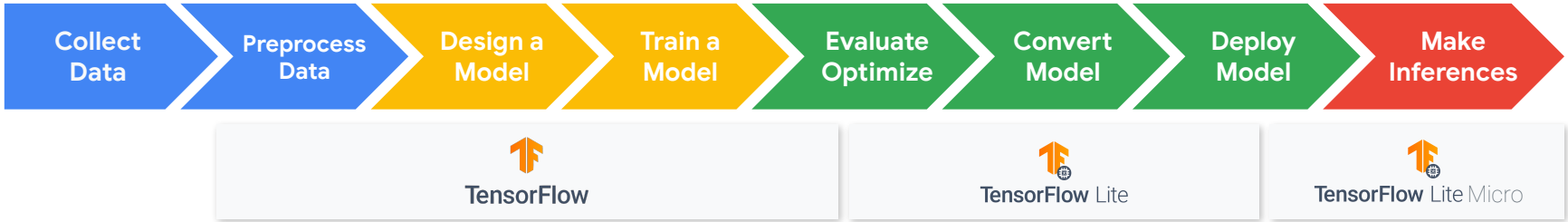
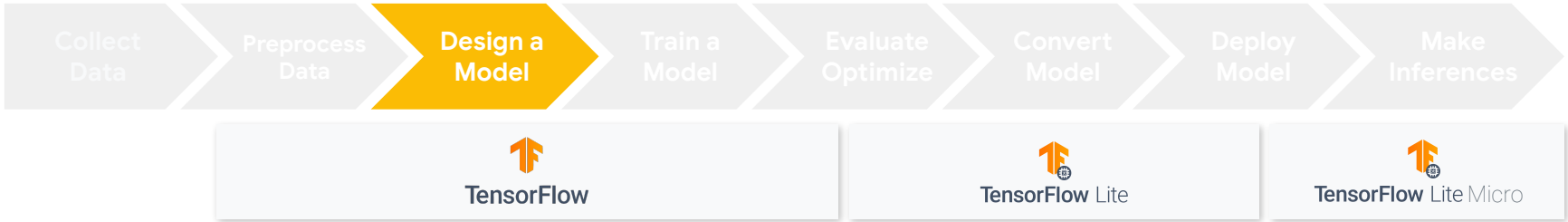
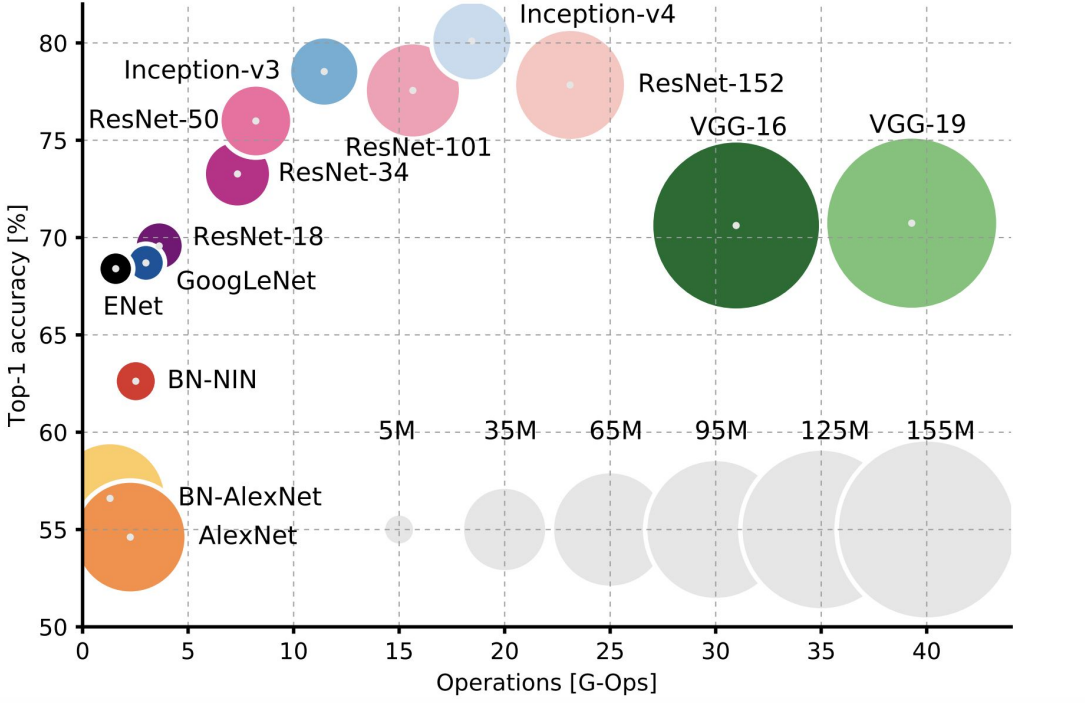
VWW: Model

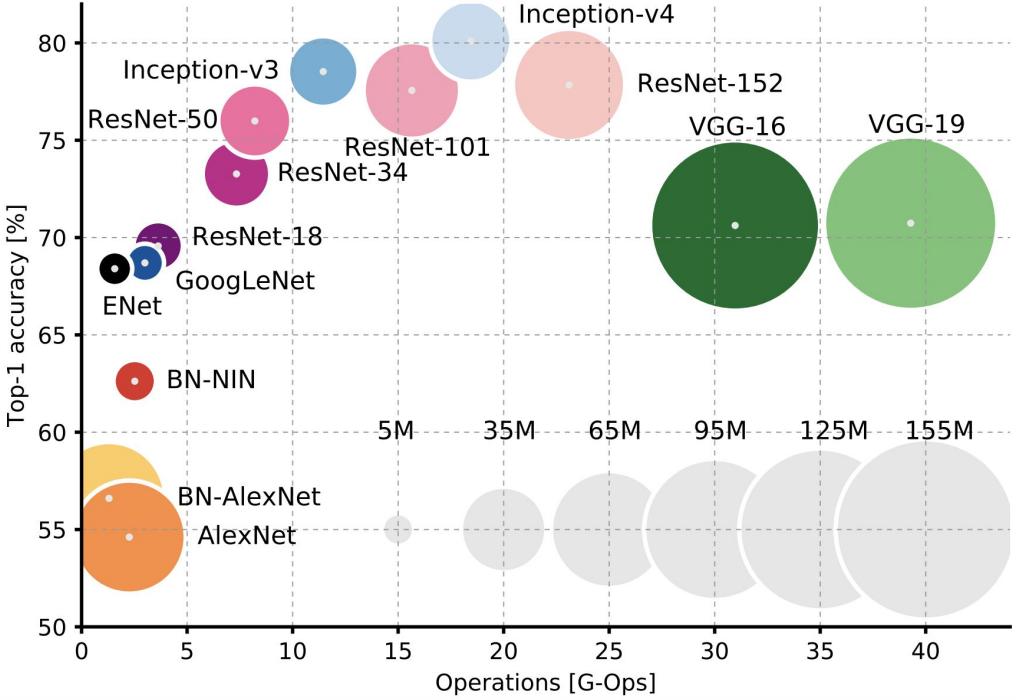

Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences
TensorFlow
TensorFlow Lite
TensorFlow Lite Micro

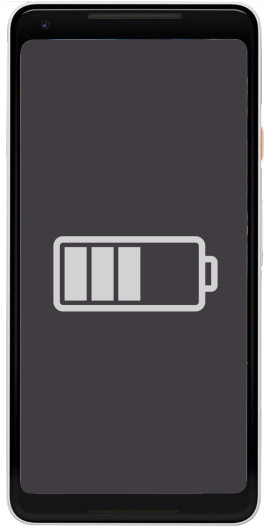

Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences
TensorFlow
TensorFlow Lite
TensorFlow Lite Micro

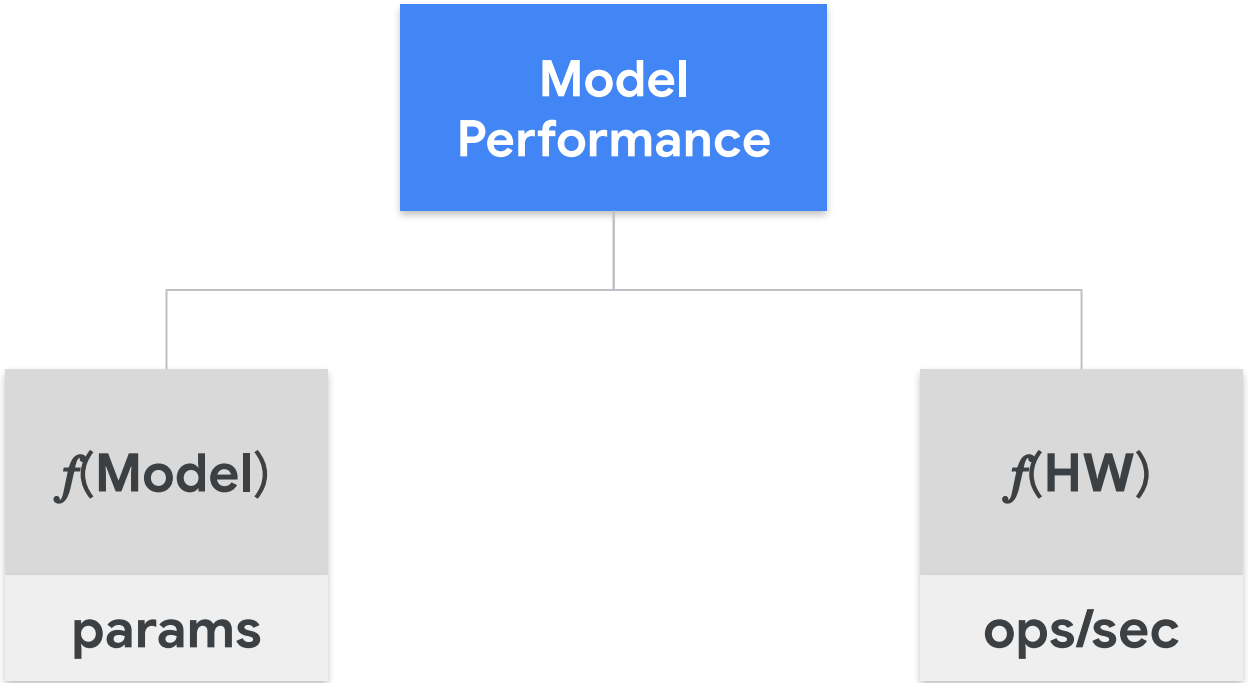
Model **Evolution**



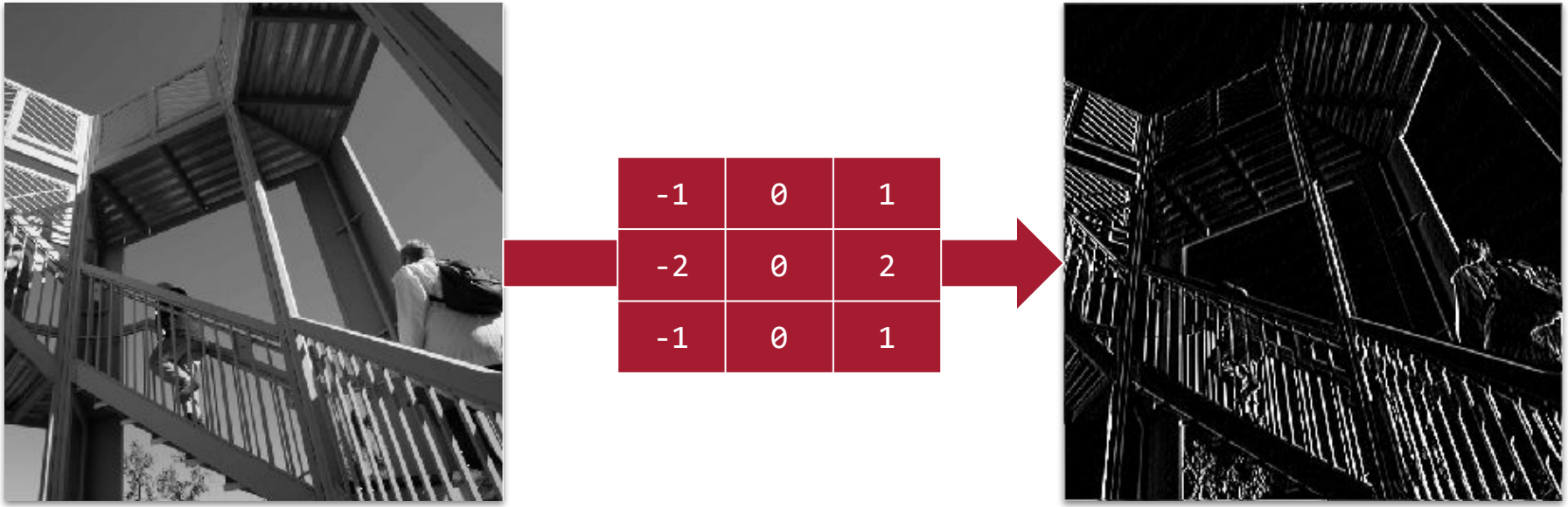
Model **Evolution**



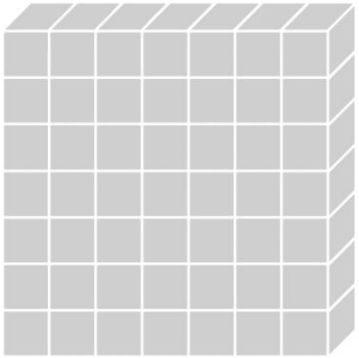



Model Performance
f(Model)
params
f(HW)
ops/sec

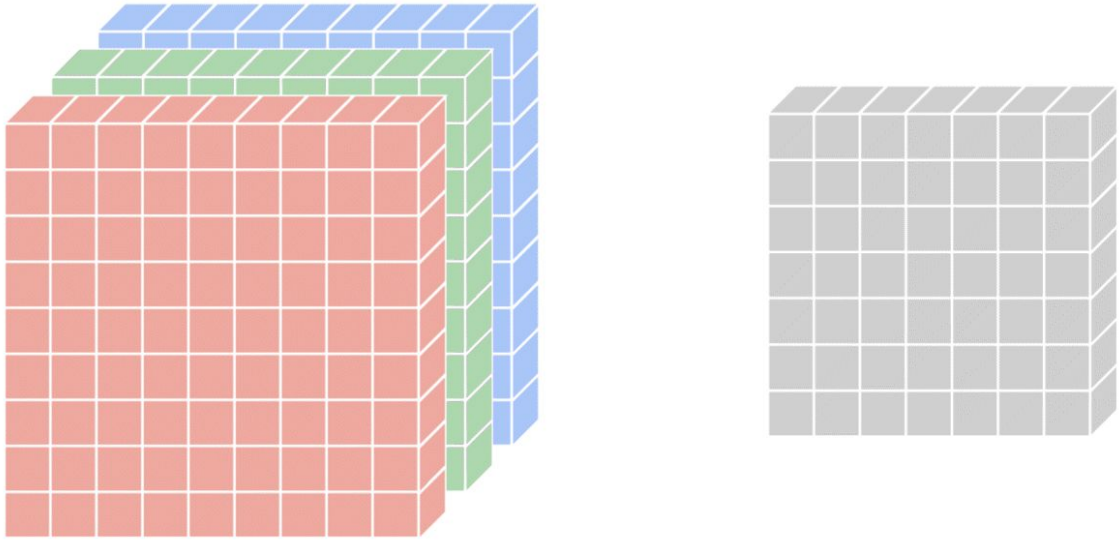
Recall: **Convolutions**



Standard Convolution ***(1 Channel)***

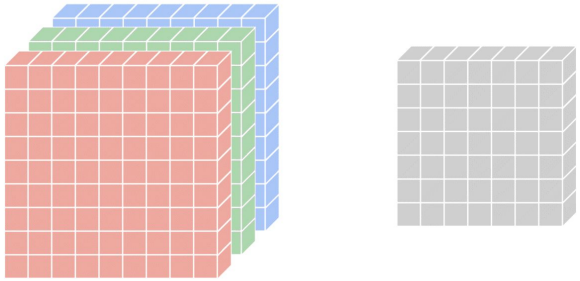


Standard Convolution ***(3 Channel**—e.g., RGB**)***

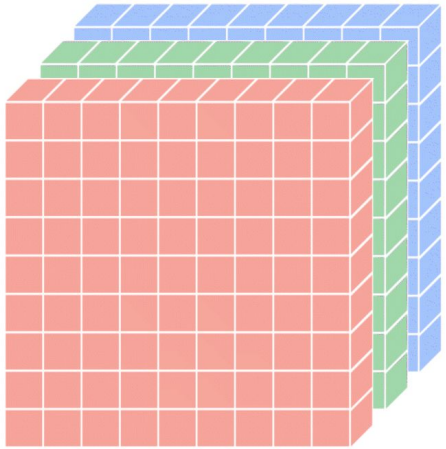


Standard Convolution ***(3 Channel**—e.g., RGB)*

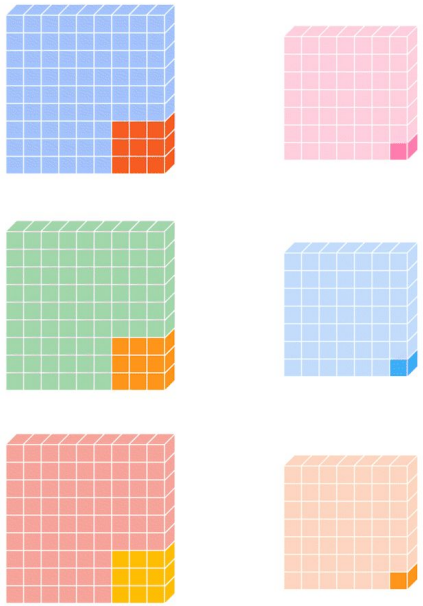
- Input Feature Map
  - $8 \times 8 \times 3$
  - Width $\times$ Height $\times$ Channels
- Kernel (*1 Filter*)
  - $3 \times 3 \times 3$



Depthwise Convolution ***(3 Channel**—e.g., RGB)*



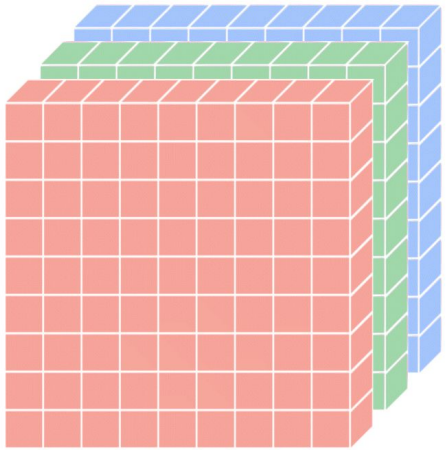
Pointwise Convolution



separable

Depthwise Convolution ***(3 Channel**—e.g., RGB)*

includes **pointwise** conv



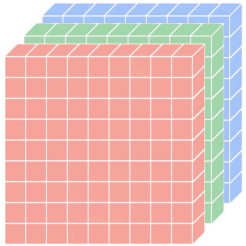
separable

Depthwise Convolution *(3 Channel—e.g., RGB)*

includes **pointwise** conv

Benefit?

Far fewer multiplications
than standard method
(especially when
using many filters)



separable Depthwise Convolution *(3 Channel—e.g., RGB)*

includes **pointwise** conv

Benefit?

Far fewer multiplications than standard method (especially when using many filters)

$$\frac{\textit{Depthwise Separable}}{\textit{Standard Conv}} = \frac{1}{N} + \frac{1}{D_K^2}$$

Filters

Kernel (filter) Dimensions

MobileNet v1

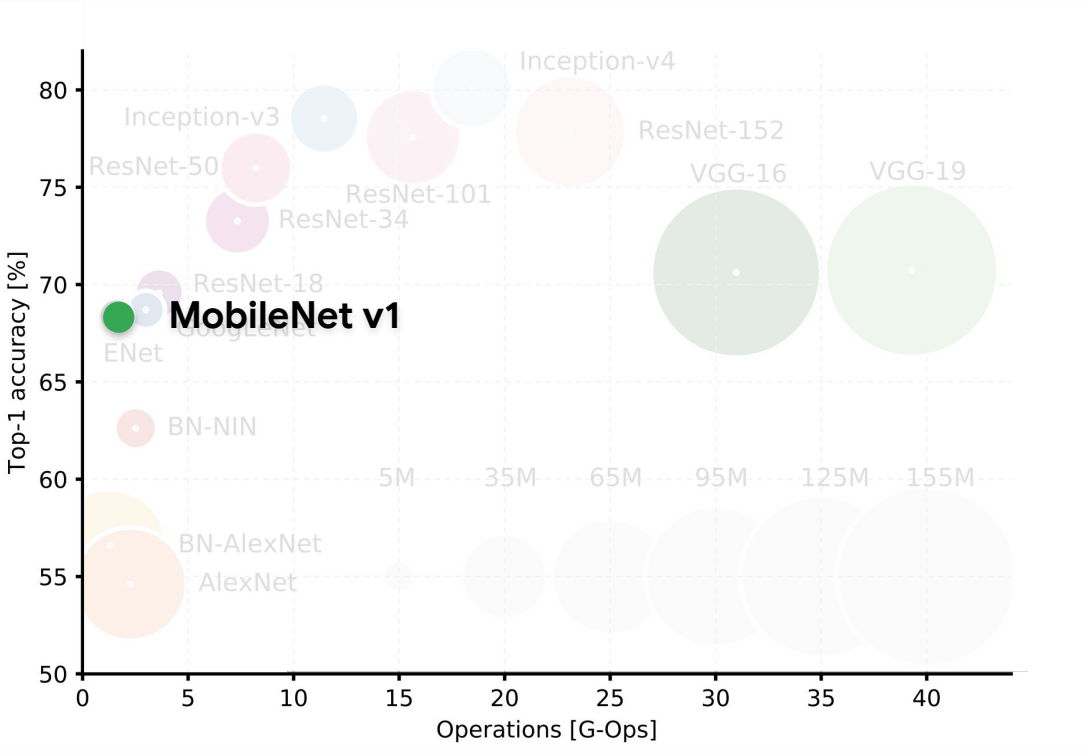
MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

Andrew G. Howard Menglong Zhu Bo Chen Dmitry Kalenichenko
Weijun Wang Tobias Weyand Marco Andreetto Hartwig Adam

Google Inc.

{howarda,menglong,bochen,dkalenichenko,weijunw,weyand,anm,hadam}@google.com

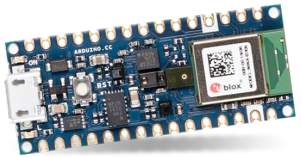
Model **Evolution**



MobileNet v1

Model	Size	Top-1 Accuracy
MobileNet v1	***16 MB***	0.713

Fine for mobile phones with GB of RAM, but 64X microcontroller RAM



Our board [Course 3 Kit] only has **256KB** of RAM (memory)

Further **Optimizations**

- Effect of **depth multiplier** on model size ⟶ top-1 accuracy
- The size of the model can be reduced further by parameter, ***α***
- ***α*** ⟶ (0, 1]

$$D_K \cdot D_K \cdot \alpha M \cdot D_F \cdot D_F + \alpha M \cdot \alpha N \cdot D_F \cdot D_F$$

Further **Optimizations**

Multiply-Accumulates

α	Image Size	MACs (millions)	Params (millions)	Top-1 Accuracy
1	224	569	4.24	70.7
1	128	186	4.14	64.1
0.75	224	317	2.59	68.4
0.75	128	104	2.59	61.8
0.5	224	150	1.34	64.0
0.5	128	49	1.34	56.2
0.25	224	41	0.47	50.6
0.25	128	14	0.47	41.2

Further **Optimizations**

Multiply-Accumulates

α	Image Size	MACs (millions)	Params (millions)	Top-1 Accuracy
1	224	569	**4.24**	**70.7**
1	128	186	4.14	64.1
0.75	224	317	**2.59**	**68.4**
0.75	128	104	2.59	61.8
0.5	224	150	1.34	64.0
0.5	128	49	1.34	56.2
0.25	224	41	0.47	50.6
0.25	128	14	0.47	41.2

Further **Optimizations**

Multiply-Accumulates

α	Image Size	MACs (millions)	Params (millions)	Top-1 Accuracy
1	**224**	**569**	**4.24**	**70.7**
1	128	186	4.14	64.1
0.75	**224**	**317**	**2.59**	**68.4**
0.75	128	104	2.59	61.8
0.5	**224**	**150**	**1.34**	**64.0**
0.5	128	49	1.34	56.2
0.25	**224**	**41**	**0.47**	**50.6**
0.25	128	14	0.47	41.2

Further **Optimizations**

Multiply-Accumulates

α	Image Size	MACs (millions)	Params (millions)	Top-1 Accuracy
1	224	569	**4.24**	**70.7**
1	128	186	4.14	64.1
0.75	224	317	**2.59**	**68.4**
0.75	128	104	2.59	61.8
0.5	224	150	**1.34**	**64.0**
0.5	128	49	1.34	56.2
0.25	224	41	**0.47**	**50.6**
0.25	128	14	0.47	41.2

Further **Optimizations**

Multiply-Accumulates

α	Image Size	MACs (millions)	Params (millions)	Top-1 Accuracy
1	224	**569**	**4.24**	**70.7**
1	128	186	4.14	64.1
0.75	224	**317**	**2.59**	**68.4**
0.75	128	104	2.59	61.8
0.5	224	**150**	**1.34**	**64.0**
0.5	128	49	1.34	56.2
0.25	224	**41**	**0.47**	**50.6**
0.25	128	14	0.47	41.2

Further **Optimizations**

Multiply-Accumulates

α	Image Size	MACs (millions)	Params (millions)	Top-1 Accuracy
1	224	569	4.24	70.7
1	128	186	4.14	64.1
0.75	224	317	2.59	68.4
0.75	128	104	2.59	61.8
0.5	224	150	1.34	64.0
0.5	128	49	1.34	56.2
0.25	224	41	0.47	50.6
0.25	128	14	0.47	41.2

Neural Architecture Search **(NAS)**