36 5

44!7322244 424 6374!5 6352244
3273432?2i26 !63 !333534!5i26

expasoft

74 5312

34X 234536 156174i6 _ 36i34 3233 75i43 44i334

3357434 163 627
175624 24 25 1 _435153 3 23i365i4i3 3755i315i2620 6 i6332 - 6
2i6634 24 32733 5i5i262 43 2. 100 43 2. IX

424X 436 233 3535 2 _ 36i34 53366i315 44i334

424 32224 15 2425 5336 76i4342i56
175624 24 20. 1 _435153 3 23i365i4i3 3755i315i2620 6 i6332 - 4.
175624 24 76i i73 7356232 424 522_4164 536224 3327322i5i262

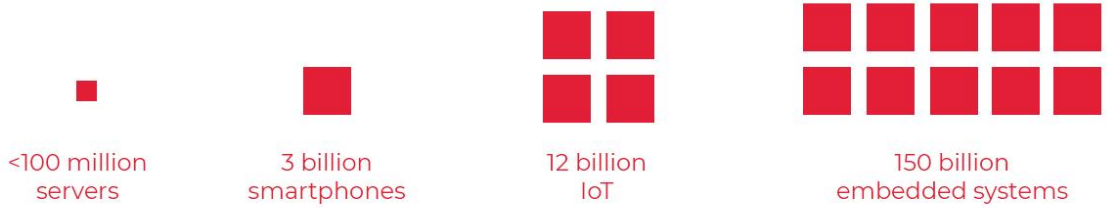
74X 451 3i7i4 3651624 _ 572i6322 3343523 7365

175624 24 3. 1 _435153 3 23i365i4i3 3755i315i262
32 4276334 2755i353 1 _435153 3 25145732

3353 1 33435237365 12 5265 163 323362143

Most of the AI will be on embedded systems

■ = 3 billion devices



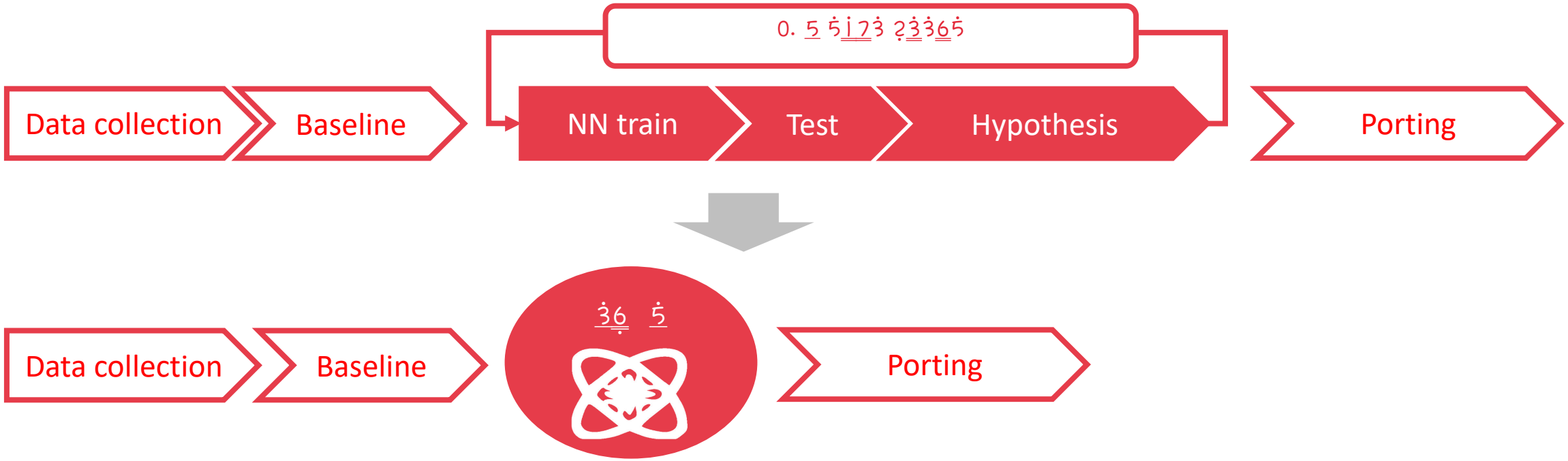
O'Reilly AI London 2019

5_7 3 _33435233342

X_12 72656?

b 3.._4..44

36 5 _ 5225 424 1 3343523342

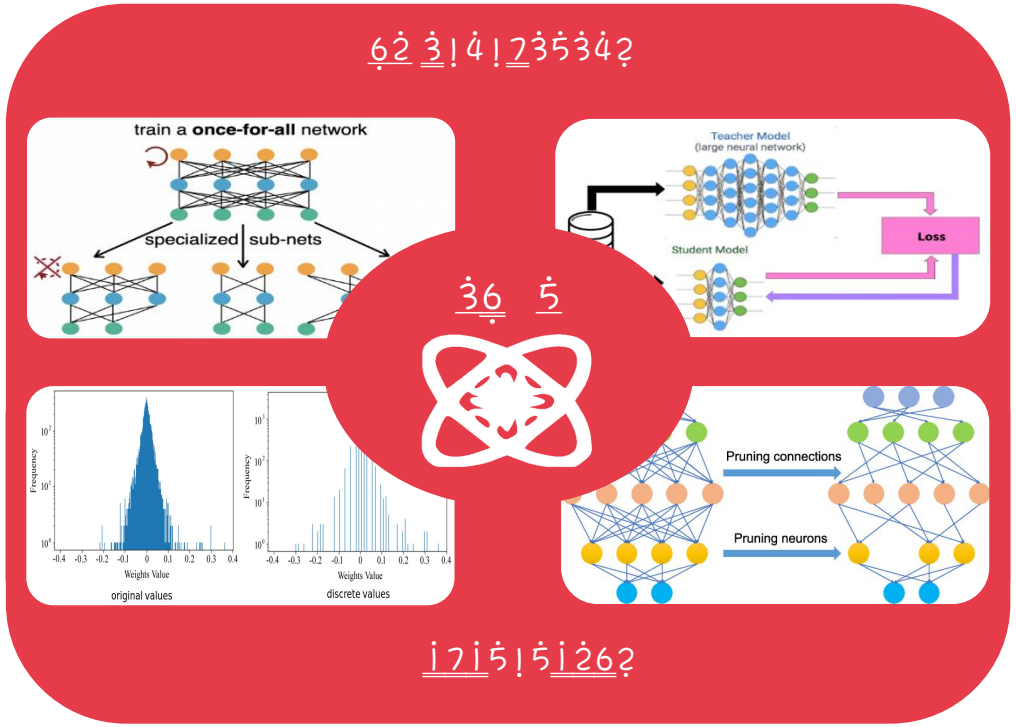


- ✓ 51765
- ✓ 57355
- ✓ 27332

36 5 _ 5225 424 | 3343523342

Data collection → Baseline → 36 5 → Porting

- ✓ 612
- ✓ 4765
- ✓ i7|65i1|5i26
- ✓ 3i25i55|5i26

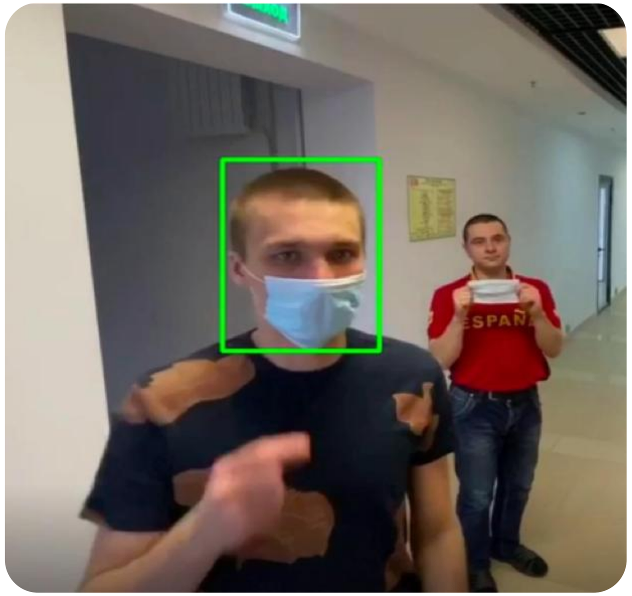


36 5 44!732244

32734322126. !333534!5126 73 52
25 5173? 2. 2 !3374!36 5222

433735126 ! 33435237365
5173
73 52 7. 5

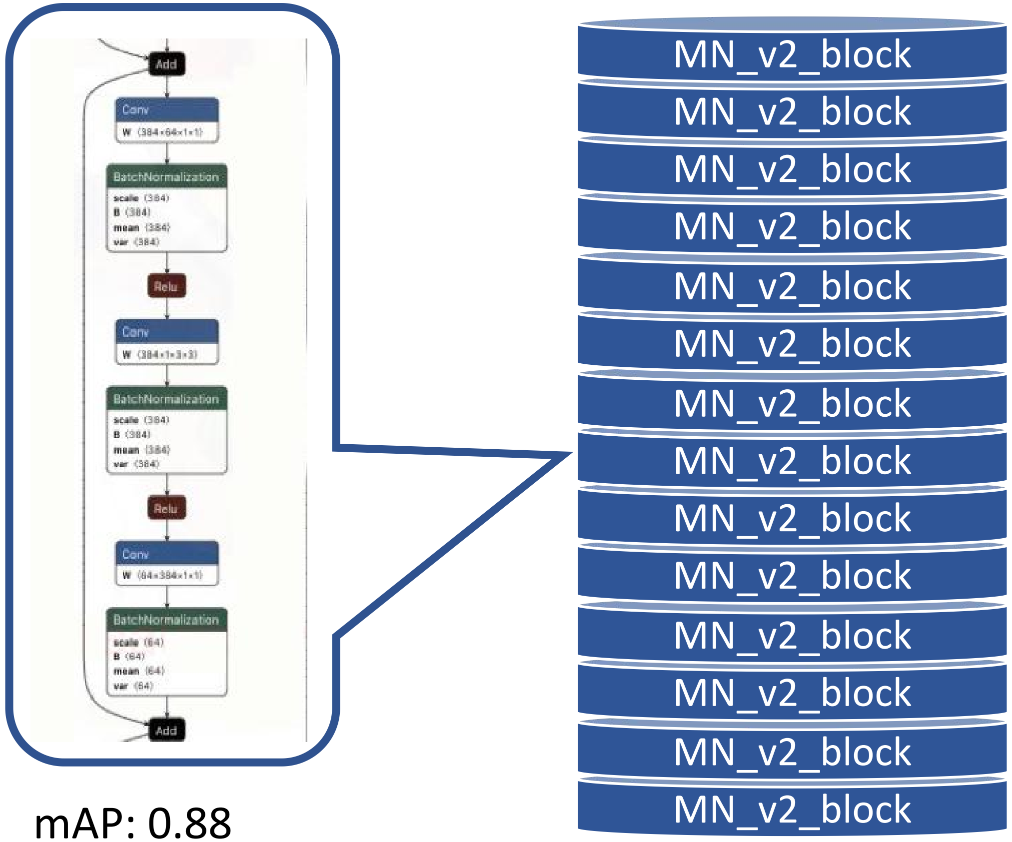
27151553 424 !66 62
35!54247



- ARM Cortex A7 1.2 GHz
- RAM: 50Mb

FPS: 1.7

MobileNet-v2 SSD



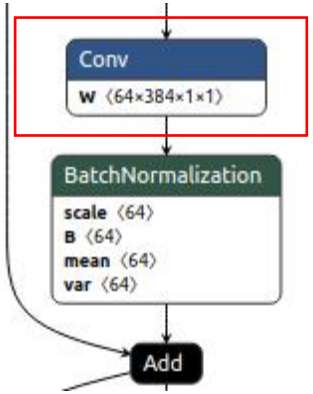
- mAP: 0.88
- MMAC: 522 mln.

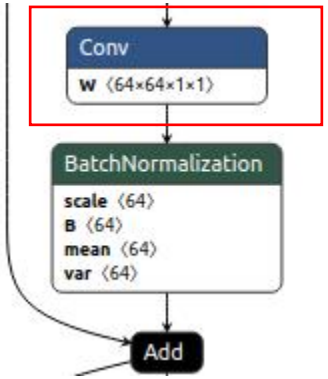
Results

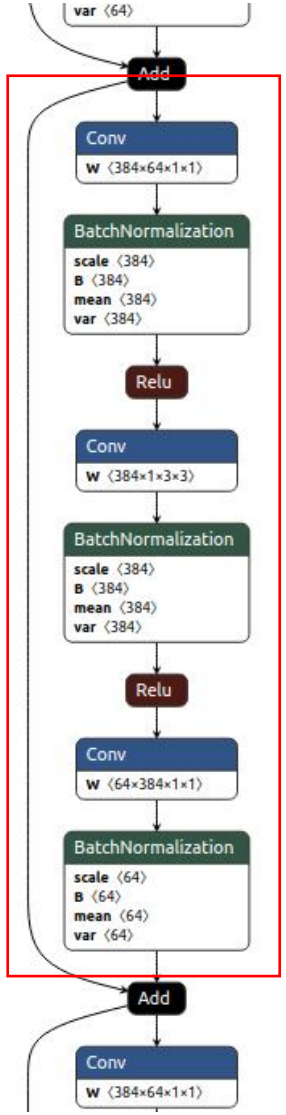
Neural Networks	Average Precision	MMAC	Layers	Model Size (Mb)	FPS
Baseline	0.88	522	16	15.3	1.7
Neural Architecture Search					
NAS (best accuracy)	0.89	501	15	10.6	1.8
NAS (best latency)	0.87	230	12	6.0	3.8
NAS + Pruning + Quantization					
NAS + prune	0.87	170	12	5.3	5.2
NAS + prune + quantizaton	0.87	170	12	1.4	7.2

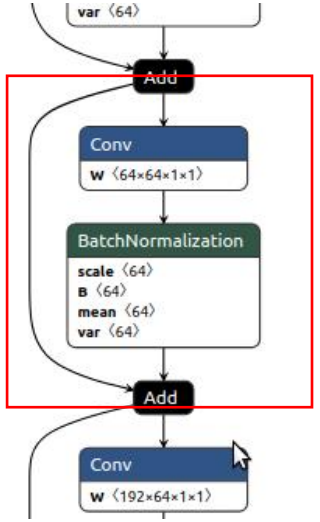
- ✓ Model size reduction - 10.9
- ✓ Latency reduction – 4.2

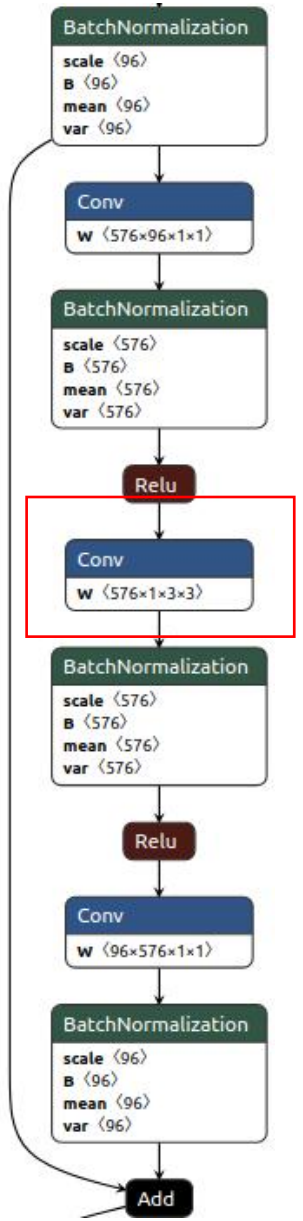
Results

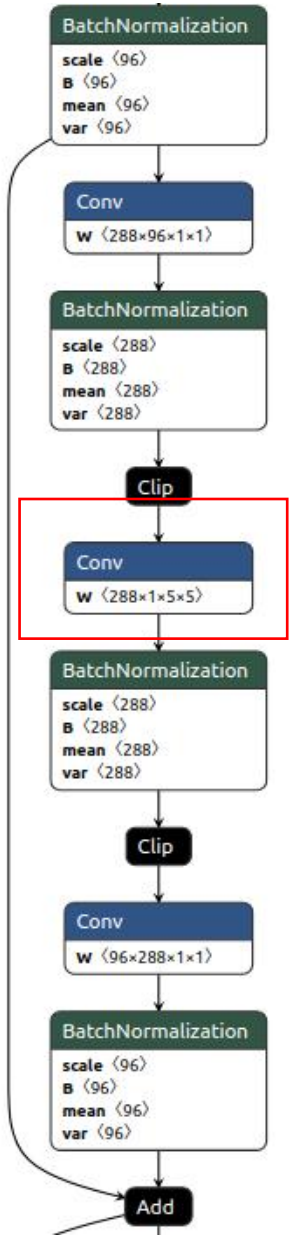

Conv
W 〈64×384×1×1〉
BatchNormalization
scale 〈64〉
B 〈64〉
mean 〈64〉
var 〈64〉
Add


Conv
W 〈64×64×1×1〉
BatchNormalization
scale 〈64〉
B 〈64〉
mean 〈64〉
var 〈64〉
Add


var 〈64〉
Add
Conv
W 〈384×64×1×1〉
BatchNormalization
scale 〈384〉
B 〈384〉
mean 〈384〉
var 〈384〉
Relu
Conv
W 〈384×1×3×3〉
BatchNormalization
scale 〈384〉
B 〈384〉
mean 〈384〉
var 〈384〉
Relu
Conv
W 〈64×384×1×1〉
BatchNormalization
scale 〈64〉
B 〈64〉
mean 〈64〉
var 〈64〉
Add
Conv
W 〈384×64×1×1〉


var 〈64〉
Add
Conv
W 〈64×64×1×1〉
BatchNormalization
scale 〈64〉
B 〈64〉
mean 〈64〉
var 〈64〉
Add
Conv
W 〈192×64×1×1〉


BatchNormalization
scale 〈96〉
B 〈96〉
mean 〈96〉
var 〈96〉
Conv
W 〈576×96×1×1〉
BatchNormalization
scale 〈576〉
B 〈576〉
mean 〈576〉
var 〈576〉
Relu
Conv
W 〈576×1×3×3〉
BatchNormalization
scale 〈576〉
B 〈576〉
mean 〈576〉
var 〈576〉
Relu
Conv
W 〈96×576×1×1〉
BatchNormalization
scale 〈96〉
B 〈96〉
mean 〈96〉
var 〈96〉
Add


BatchNormalization
scale 〈96〉
B 〈96〉
mean 〈96〉
var 〈96〉
Conv
W 〈288×96×1×1〉
BatchNormalization
scale 〈288〉
B 〈288〉
mean 〈288〉
var 〈288〉
Clip
Conv
W 〈288×1×5×5〉
BatchNormalization
scale 〈288〉
B 〈288〉
mean 〈288〉
var 〈288〉
Clip
Conv
W 〈96×288×1×1〉
BatchNormalization
scale 〈96〉
B 〈96〉
mean 〈96〉
var 〈96〉
Add

Benchmarks

Pascal VOC	mAP	MMAC	Acceleration
Retina-ResNet50 (baseline)	0.72	17086	1
ENOT_RetinaRN50-1	0.71	3458	4.9

CIFAR-100	Accuracy	MMAC	Acceleration
Resnet-50 (baseline)	0.76	1178	1
ENOT_RN-50_1	0.76	584	2
ENOT_RN-50_2	0.75	170	6.9

CIFAR-100	Accuracy	MMAC	Acceleration
Resnet-18 (baseline)	0.74	547	1
ENOT_RN-18_1	0.75	263	2
ENOT_MN_18_2	0.73	92	5.9

Mseg	mIoU	MMAC	Acceleration
HR-net (baseline)	0.44	46256	1
ENOT_HR-net	0.43	17731	2.6

4!33 433256i5i26 424 27!45 3!734!

3!23 4!33 433256i5i26

i7i5!5i262 !. МБ0 !47 _5. . 76!0 5. . 72

!3374!36 534243._!4534X5. XXX07 ._ XXX63

!333534!5i26. 32734322i26 !2 ._ !5 5i732

i33633 5i53 433256i5i26 424 4 i

3i23 335335i26 34

i7i5i5i262 i47 _ IX2 56I0 25. 72

I3374I36 534243. I4534X5. X7XI . X6X0

I33534I5i26. 327343 22i26 6 . I0 5i732

233|434 |3365|4|3|5|26 424 |6342|3 54 235

3|23 233|434 3365|4|3|5|26

|7|5|5|262 |47 _X. . 7610 4 45

334 534243._|45334X5._ |XX ._ 2X4

|3335334|5|26. 3273432 2|26 25 ._ 2. 5|732

74 136134373652

1 3231224 5 5312 226 125 34113 15 333 6534615126 15 22_ 2234 2153 4332561 5126 36155365 3 X2. 10 43_ _ 54134 1 2 54134 20 62437534 2. 10X

2 3231224 5 5312 525 263 34113 15 333 6534615126 15 22_ 2234 2153 4332561 5126 36155365 3 X2. 1X 43_ . 0 7763 2. 1XX


HUAWEI


LG


alhua
TECHNOLOGY


SberCloud


HISILICON


Яндекс Облако


PRISMA


rubetek


ЭЛЕКТРОННЫЙ
ГОРОД


форпост

56 1 64 6271

6313 15243 563 327335152421 1 1