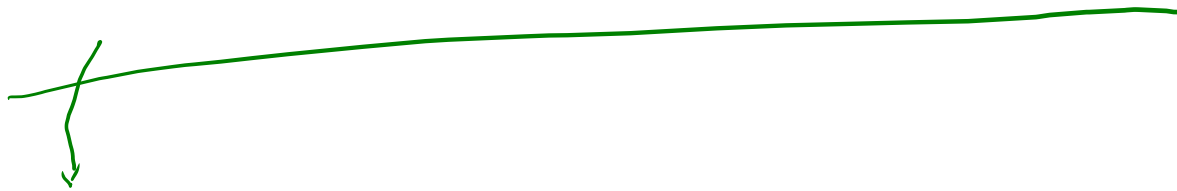


TensorFlow lite



 .h5

.tf

inference time ← resource

h5

resource

- mobiles
- micr. controllers
- Raspberry pie.
-

edge device
↓
ex mobiles

مجازاً ہمارے اکوڑز میں ہیں! موبائل صحیح نظریہ ہوتا ہے

main models

میں deployment کے
↓
problems

- latency
- privacy ~ on device programming
- power consumption.

post-training

Rel

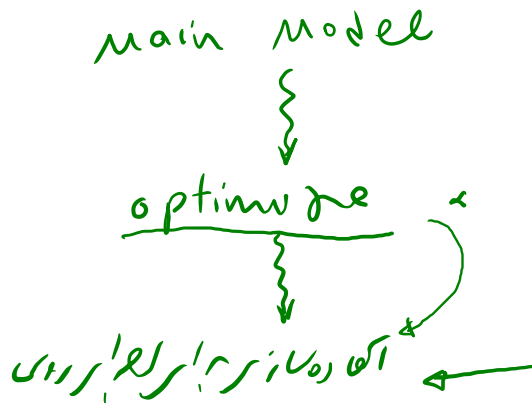
↓
Optimization.

.h5 → .tfh5
.tf

Delegate

α

→ CPU
→ GPU
→ TPU
i



edge TPU

NN API
GPU

Hexagon

CPU

GPU

TPU

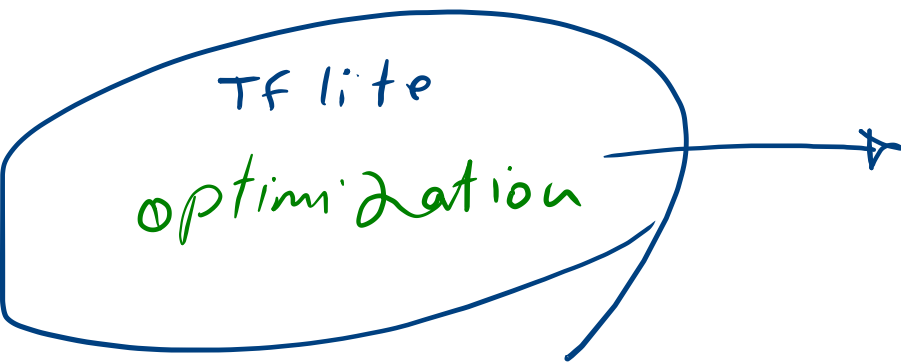
Tensor flow
lite

TF lite

ok

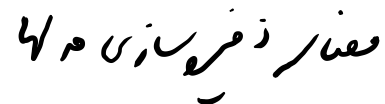
- mobile → Android
→ ios
- embedded linux
- microcontroller ✓
- java / c++ / c / python

TF lite
optimization



برای optimize کردن چکار کنند؟

P.



سہار (نوع المام + اسوسن) مانکن



α weight, data types → float64

α float16

α int16

α int8

α optimization

activation تغییرات در خروجی

Relu → 0,513691432

int8(Relu)

دستور
← optimization

prediction تأثیری ندارد،
بعضی از نودها و درختها
حذف می کنند!

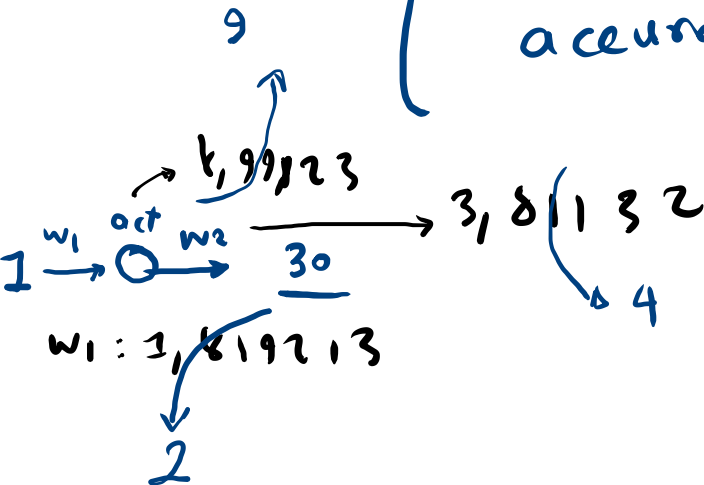
موضوع دوم

pruning
clustering

۲

optimization عملية

[accuracy دقة]



model type	<u>delegate</u> GPU	ARM <u>NNAPE</u>	<u>Hexagon</u>	<u>ios</u> <u>CoreML</u>	A17
float 32	✓	✓	X	✓	
float 16	✓	X	X	✓	
integer	<div data-bbox="319 461 586 699"> <div> <div>int</div> <div> w & a: 16 bit w & a: int k bit </div> </div> <div> w: 16 bit a: k bit </div> </div>	✓	✓	X	
dynamic range	✓	✓	X	X	

4 mnist



main model acc : 1/95



TF lite



acc : 1/80

float 32

float 16

int 16

int 8



14mb



4mb

Experience!

int8

float16



Gpu

low speed
70ms

high speed

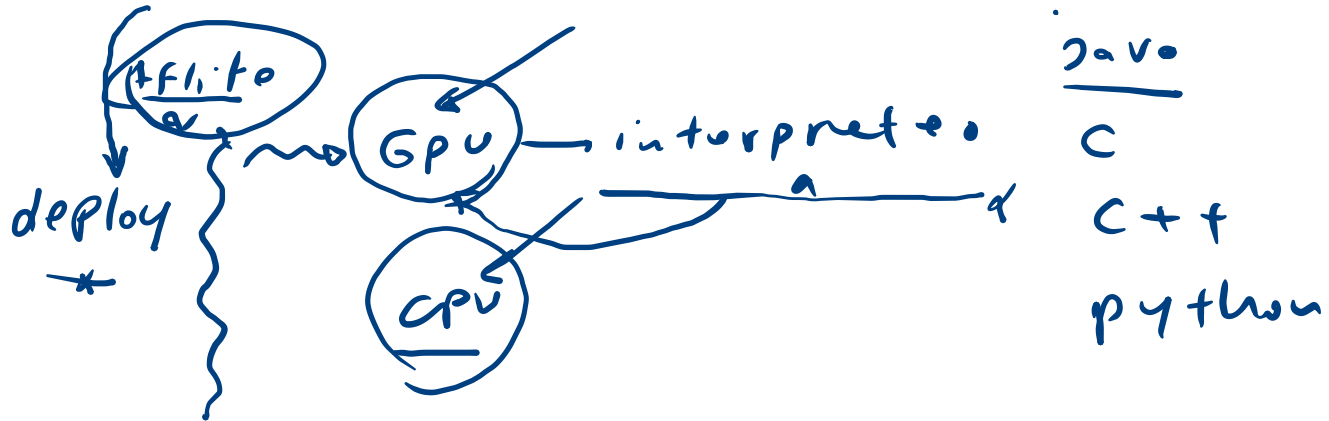
CPU (NNAPI)

high speed
40ms

low speed.

main model.

optimization

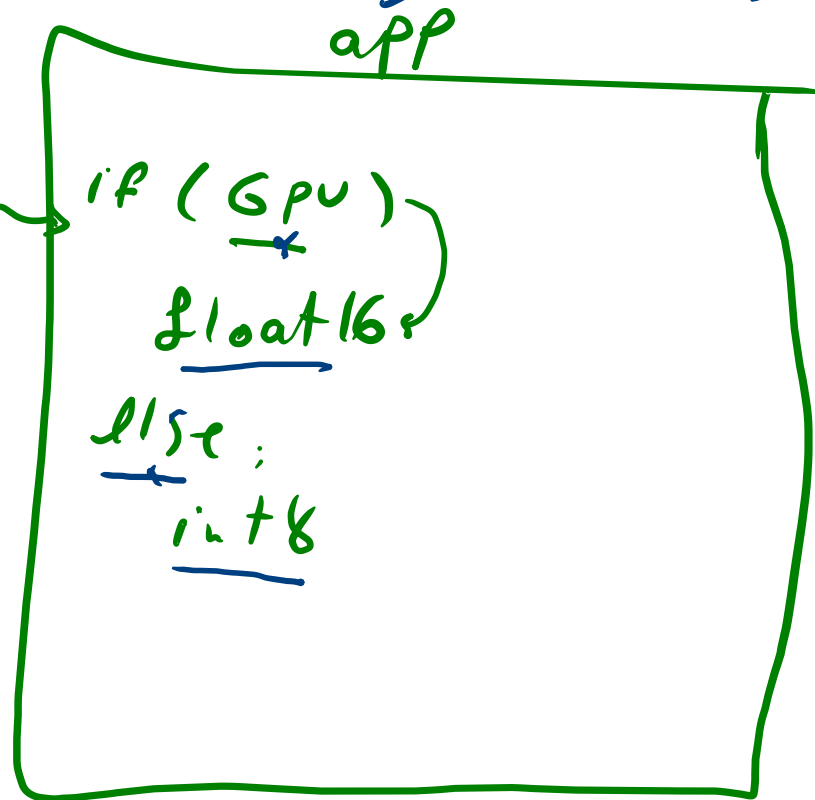


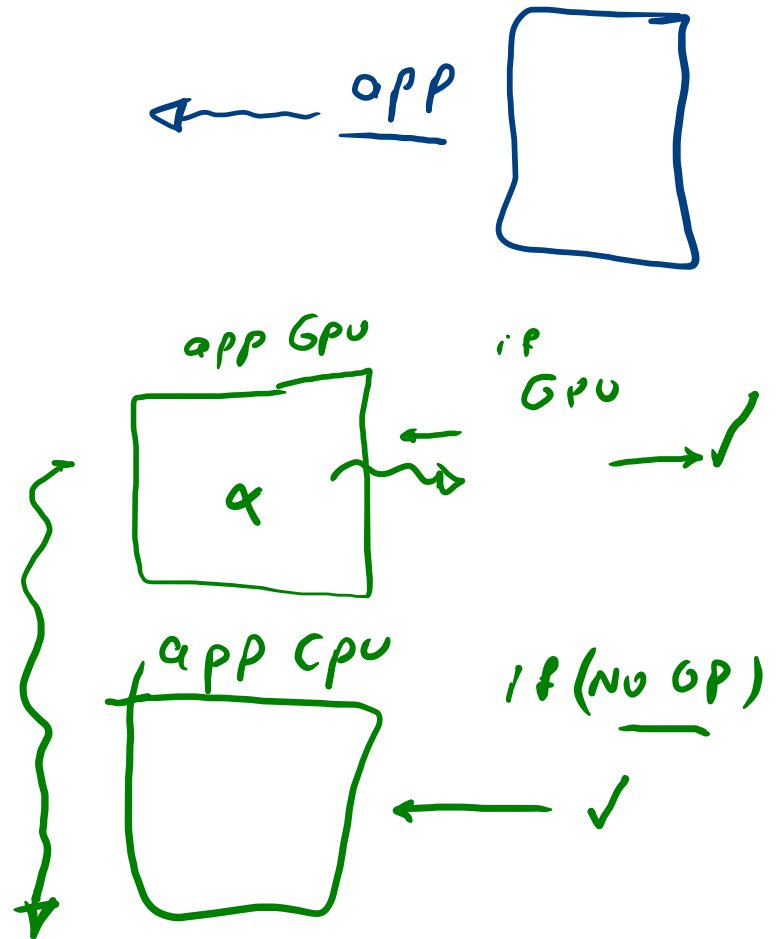
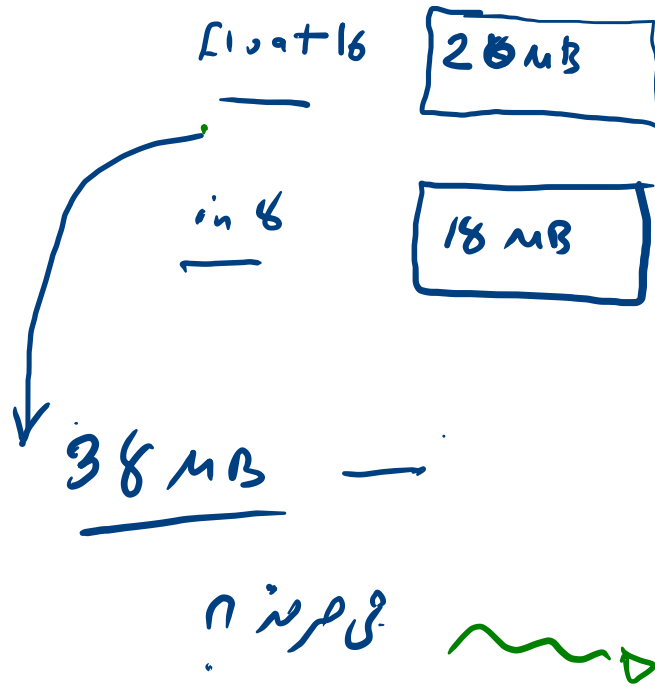
سوال: وقتی می‌دانیم که user از accelerator استفاده می‌کند



→ int8 → CPU

→ float16 → GPU





app

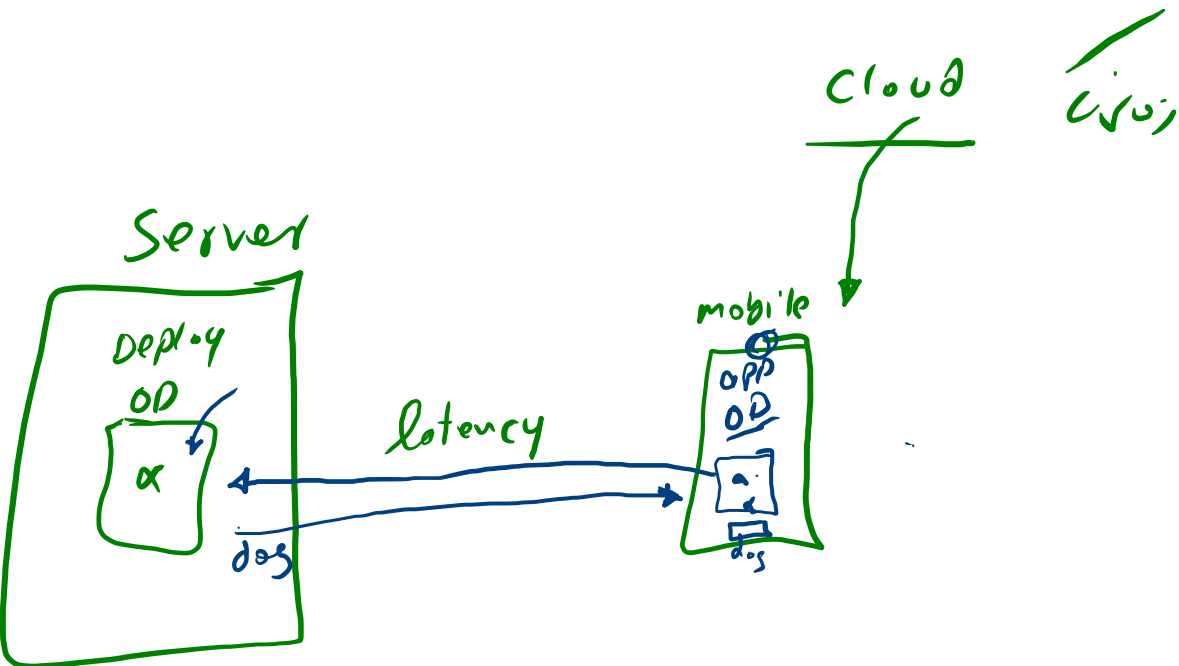


a GPU, a var

download → GPU

No GPU

download CPU



End