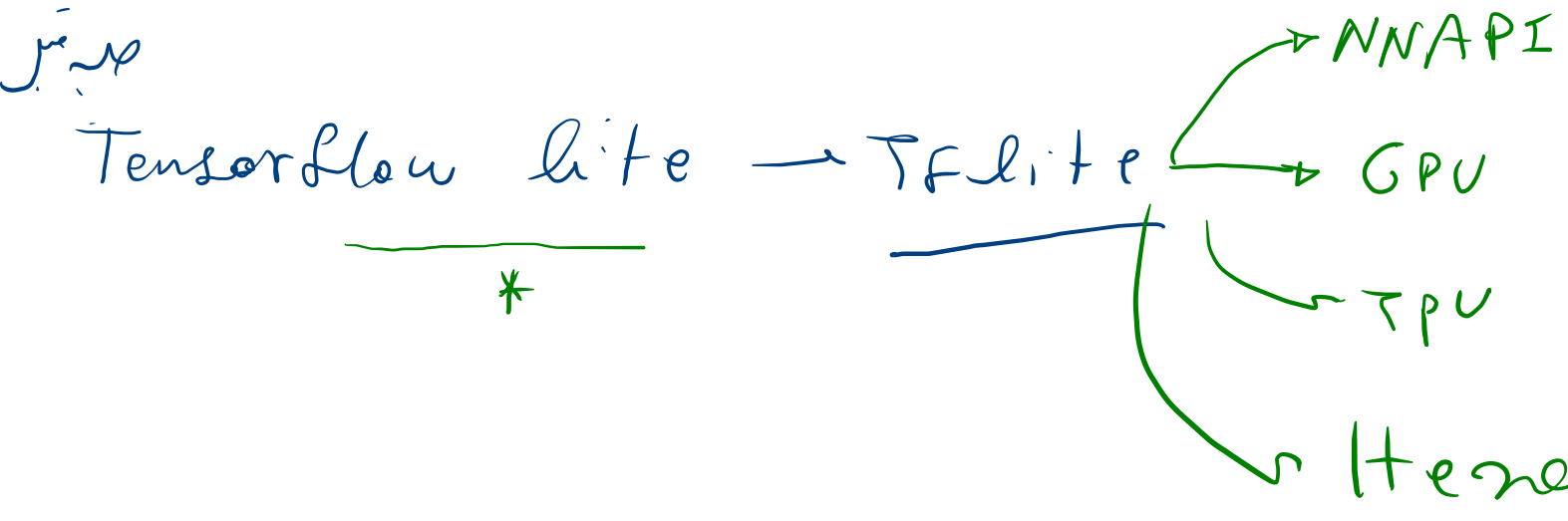
TRT.

↓

NVIDIA. / purpose: NVIDIA GPU

Tensor Run time (TRT)



JRT

optimization ⇝ real time applications

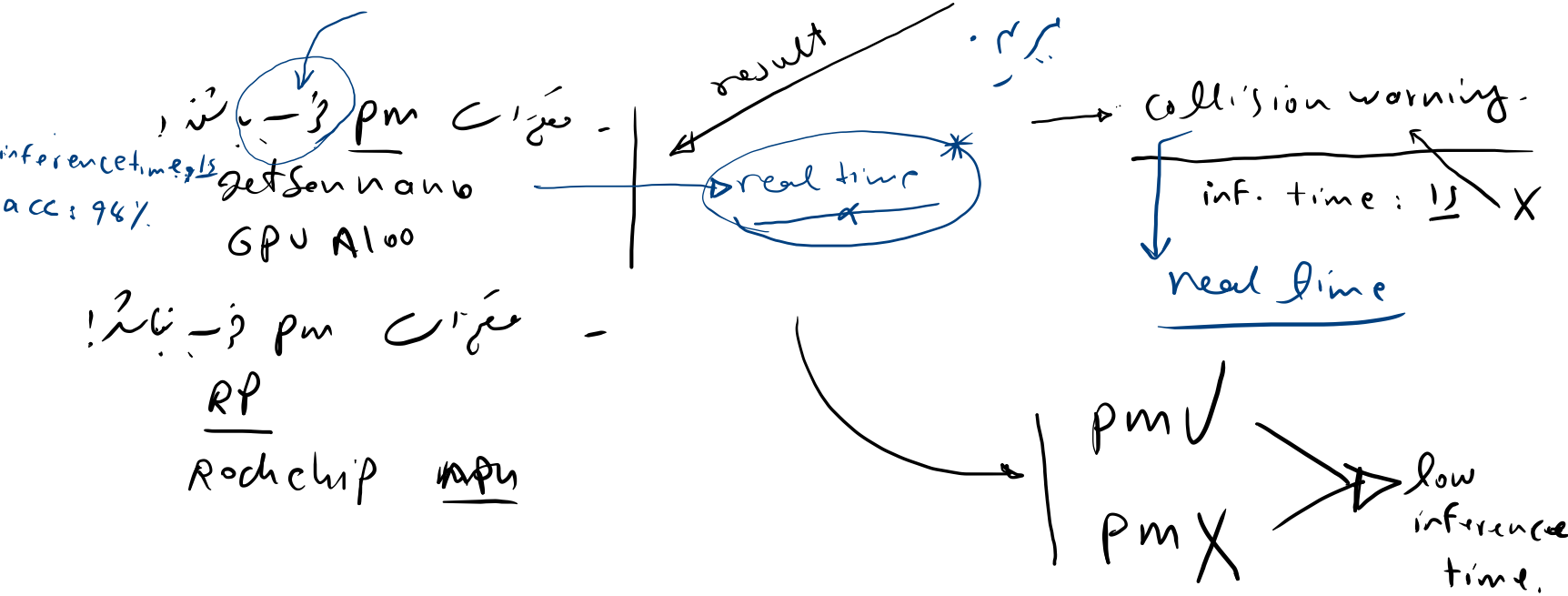
DL →

1. Train →
   - complex.
   - high computational cost.
   - long time process.

2. Inference →
   - predicting unseen data.
   - real application/environment.

   - time
   - accuracy.

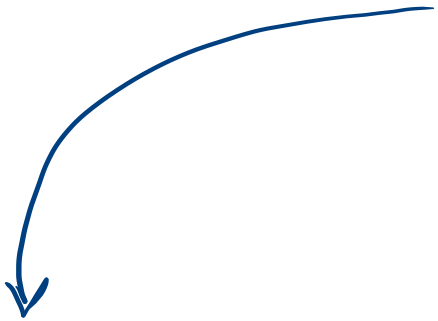
مدل Train شده را در فیزیک بدون تغییر در حین Inference



real time → response : < 40 ms

< 10 ms

< 5 ms



limited resource → must optimization.

imp. optimizati— libs —→ TFlite

→ TRT

X

Fp32

int8 Dynamic range.

-127 to +127

Fp

19,23487 → 4 X 19,23487 = 76,9 → cat

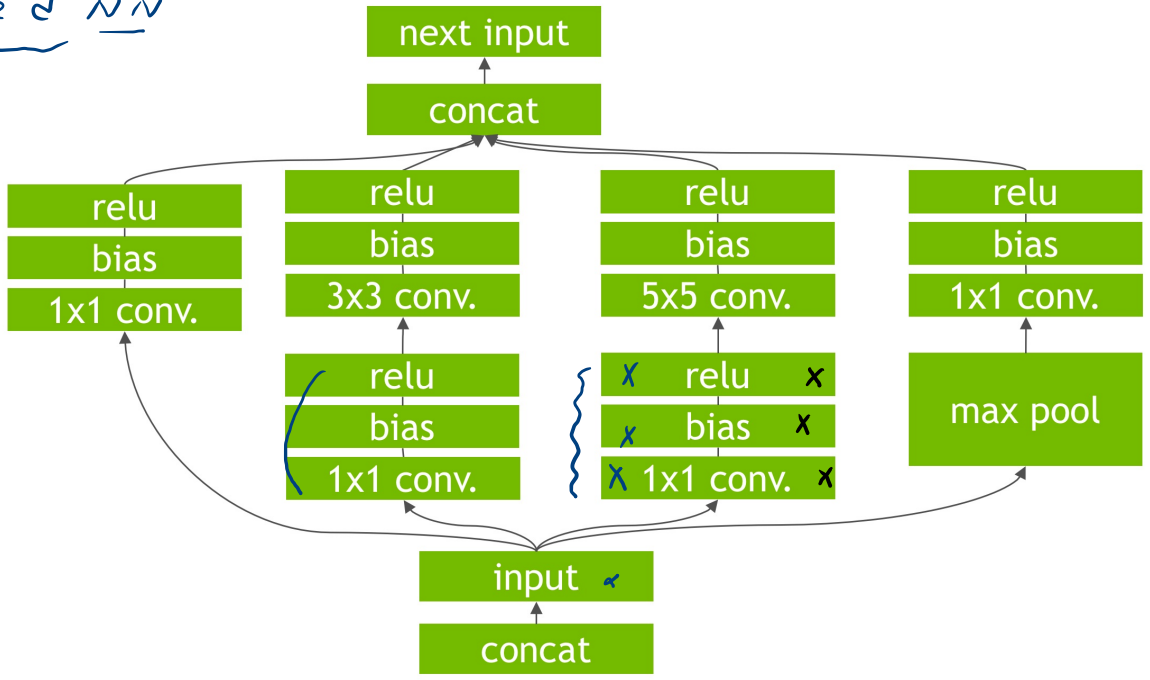
int

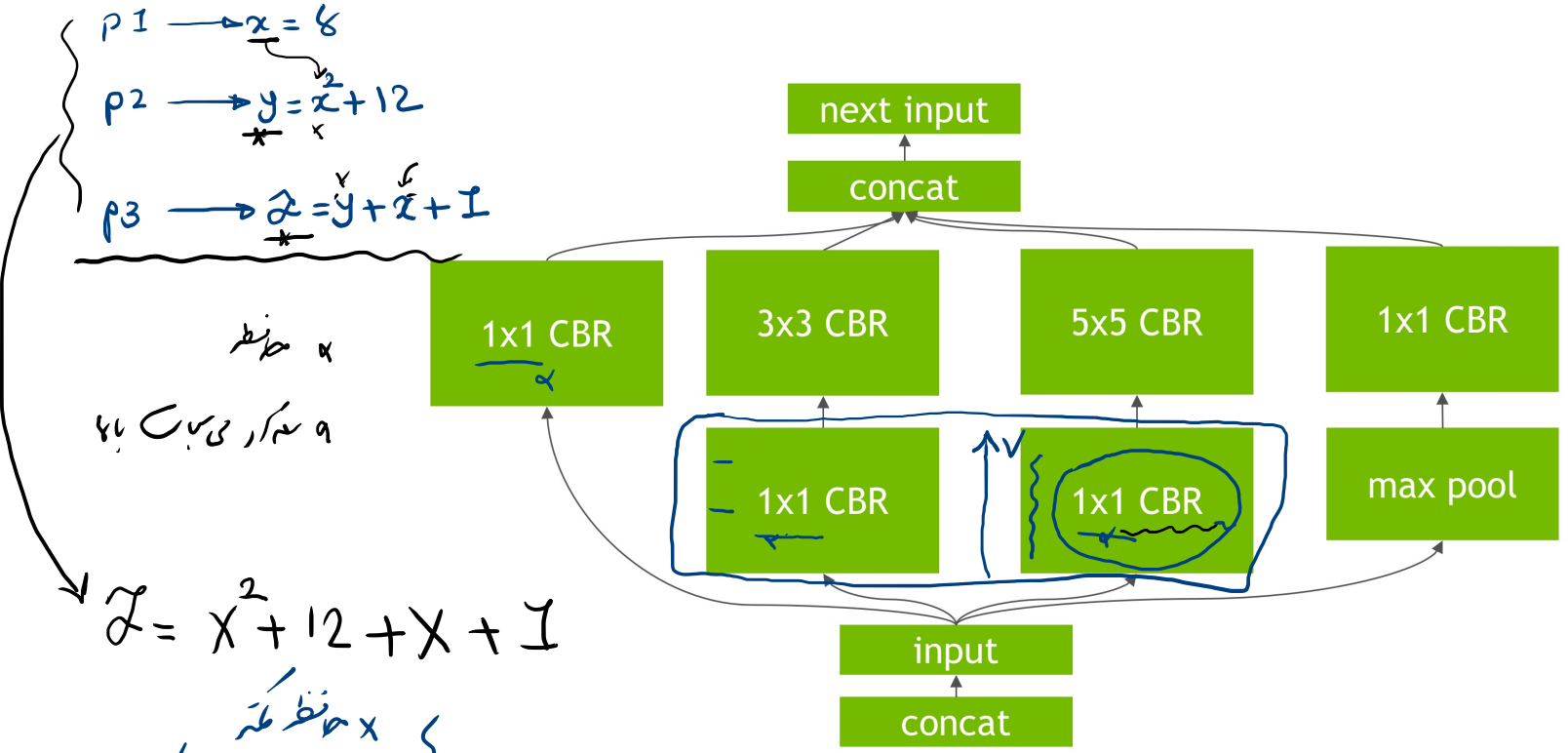
19 → 4 X 19 → 76 → dog

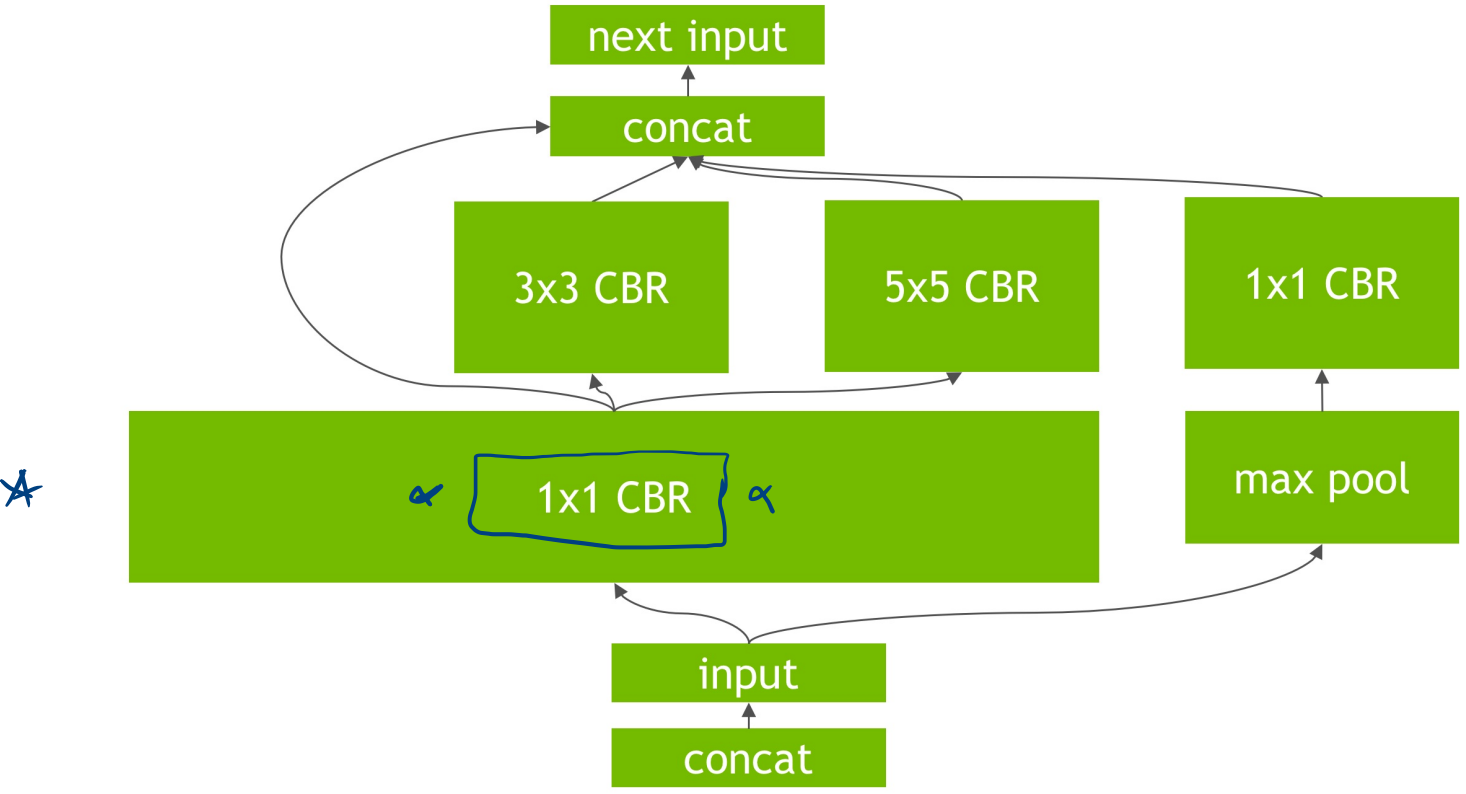
3. Layers Fusion.

TRT

unoptimized NN




next input
concat
1x1 CBR
3x3 CBR
5x5 CBR
1x1 CBR
1x1 CBR
1x1 CBR
max pool
input
concat


next input
concat
3x3 CBR
5x5 CBR
1x1 CBR
1x1 CBR
max pool
input
concat

layer fusion

network	before	after
VGG19	43	27
inceptionV3	309	113
Resnet 152	670	159

2. precision calibration

→ LPR : low precision Inference.

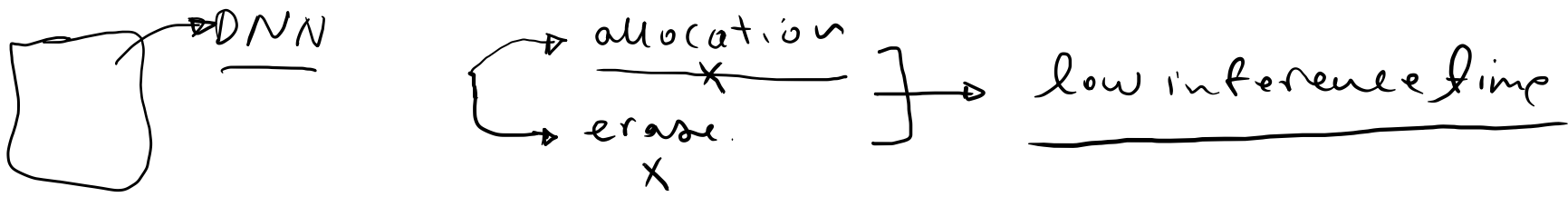
weights, bias, activat Func.

FP32 — DTC → FP16

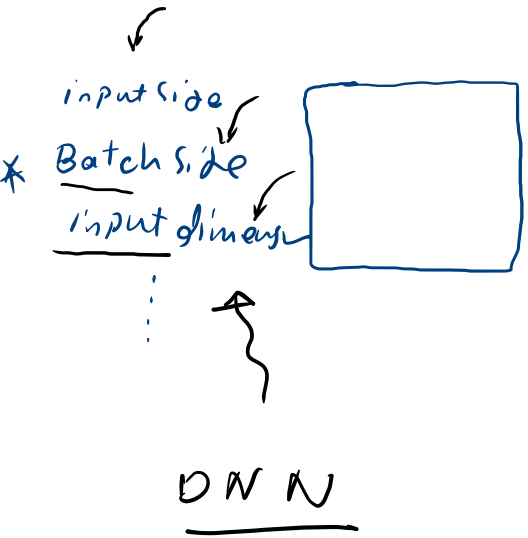
FP32 → INT8

mapping / FP32 → INT8 — KLD

3. Dynamic Tensor memory allocation!



4. Kernel auto tuning!



TRT ⇝ other parameters opt.

- ResNet 18

169

→

PT :	1X (Fp32)
TRT Fp32	→ 1.5X
TRT Fp16	→ 5.5X

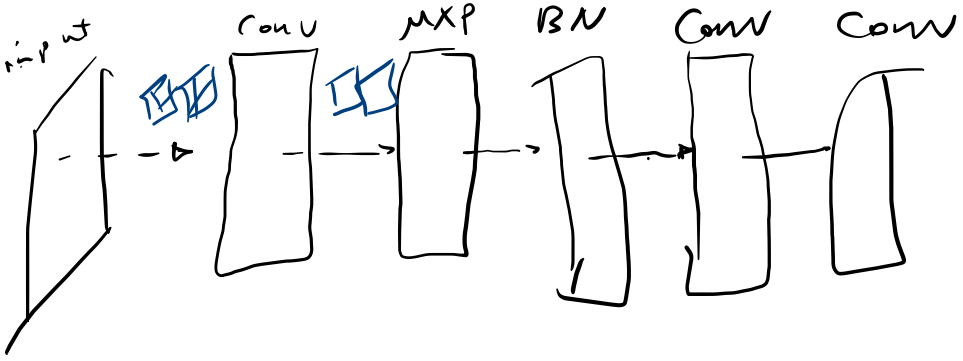
- Real Time

* TensorRT → C++ / python

pytorch ←

Tensorflow ←

* Resnet 50.



α

image Net