主要内容

- 量化方法MIN-MAX、ADMM量化
- TensorRT 的组网及使用

output
Dataset
weight
input
scale
dequant
min_max
conv
fp32
int8
Max
127
-127

$fp32\ V = int8\ V \cdot S$

$int8\ V = \dfrac{fp32\ V}{S}$

$f(x) = \dfrac{x}{S}$

$f^{-1}(x) = S \cdot X$

量化过程

float32

Weight of fp32 / S → int8

Tensor

$L = \left(S \cdot E\left(\dfrac{X}{S}\right) - X\right)^2$

$\dfrac{\partial L}{\partial S} = E\left(\dfrac{X}{S}\right) \cdot \left[S \cdot E\left(\dfrac{X}{S}\right) - X\right]$

$E\left(\dfrac{X}{S}\right) \cdot \left(S_{k+1} \cdot E\left(\dfrac{X}{S_k}\right) - X\right) = 0$

ADMM

(cont) C

lct inside input

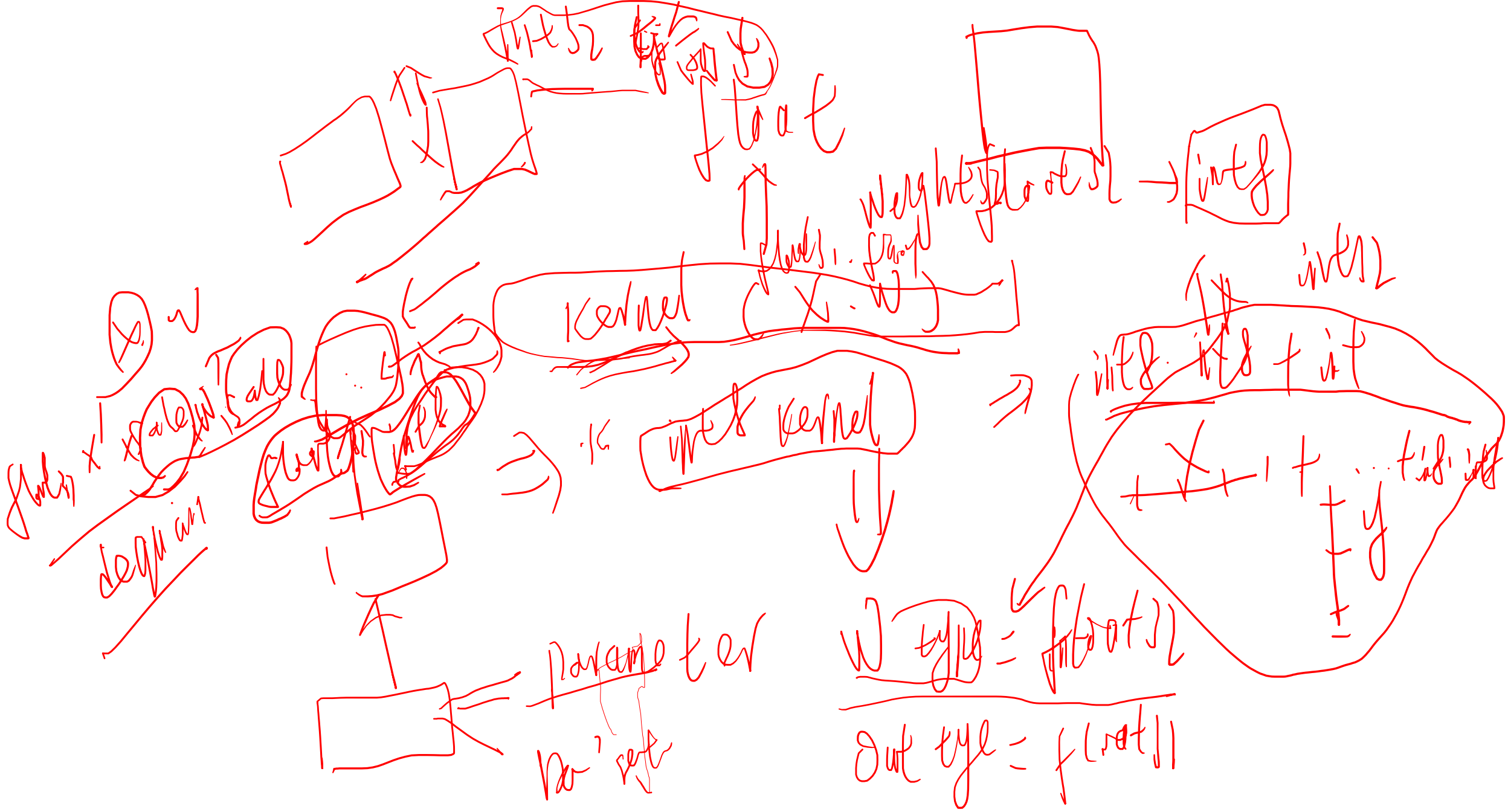
权重引化：

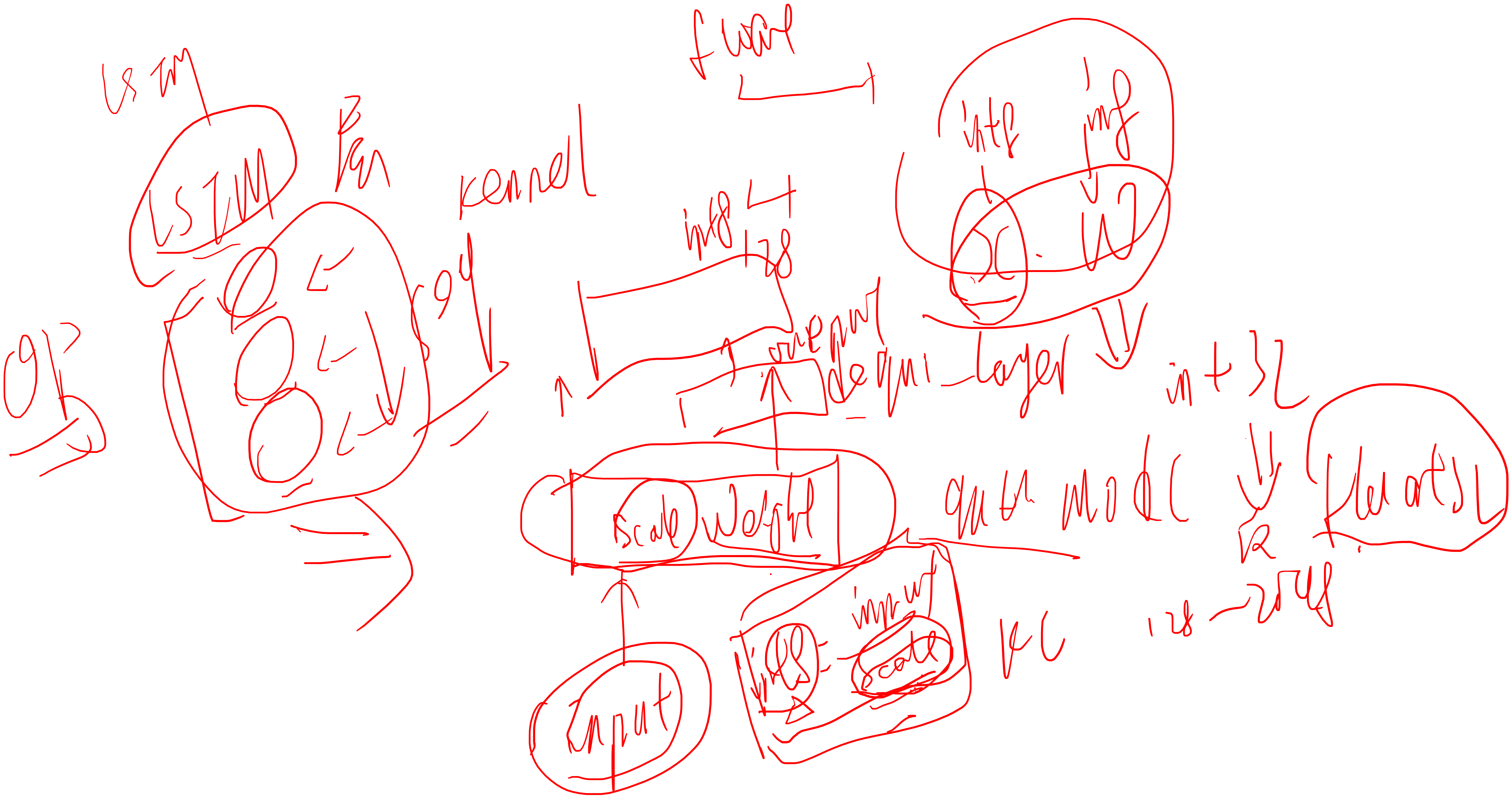
$$S_{k+1} = E\left(\frac{x}{S_k}\right) \cdot x$$

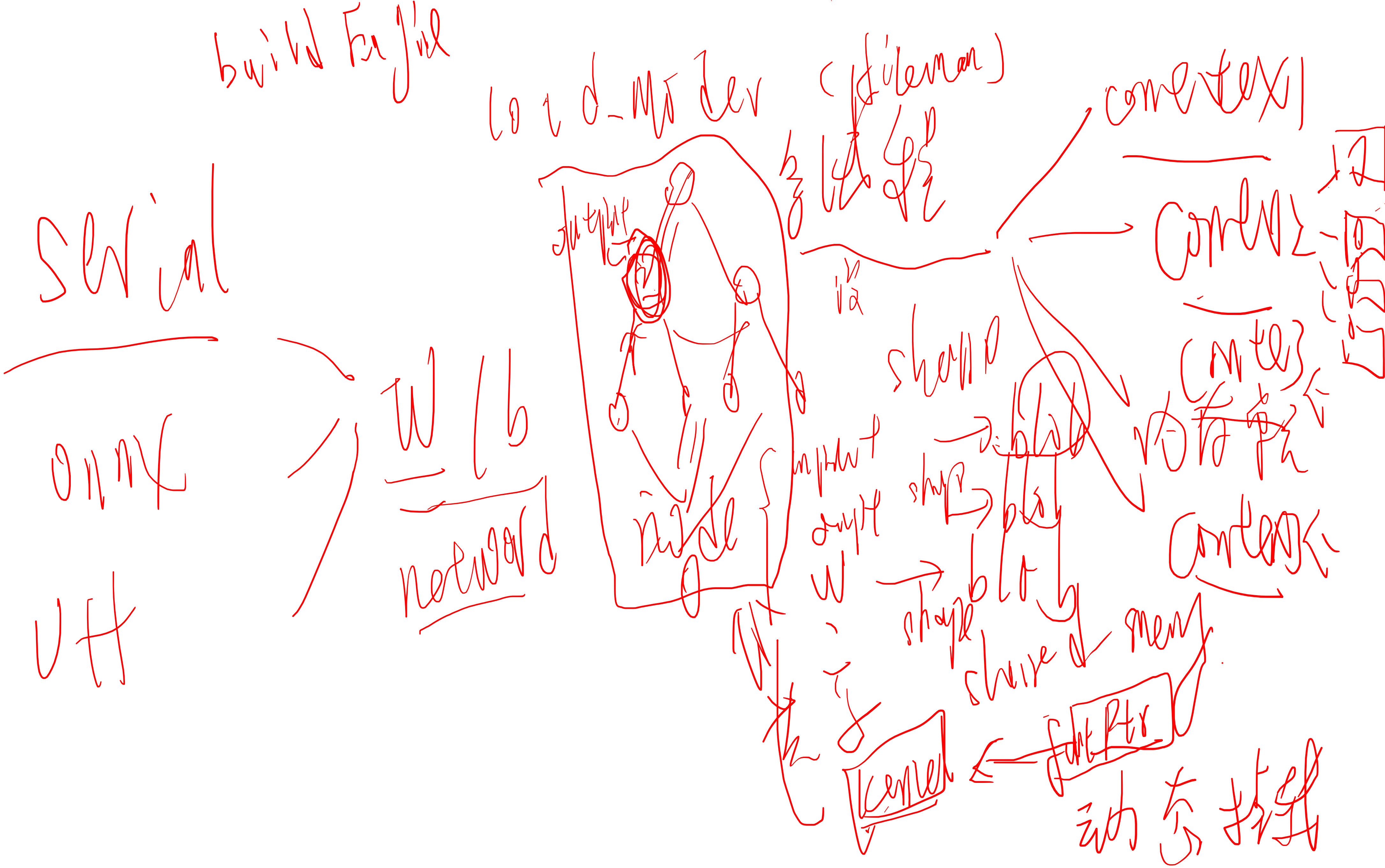
$$E\left(\frac{x}{S_k}\right) \cdot E\left(\frac{x}{S_k}\right)$$

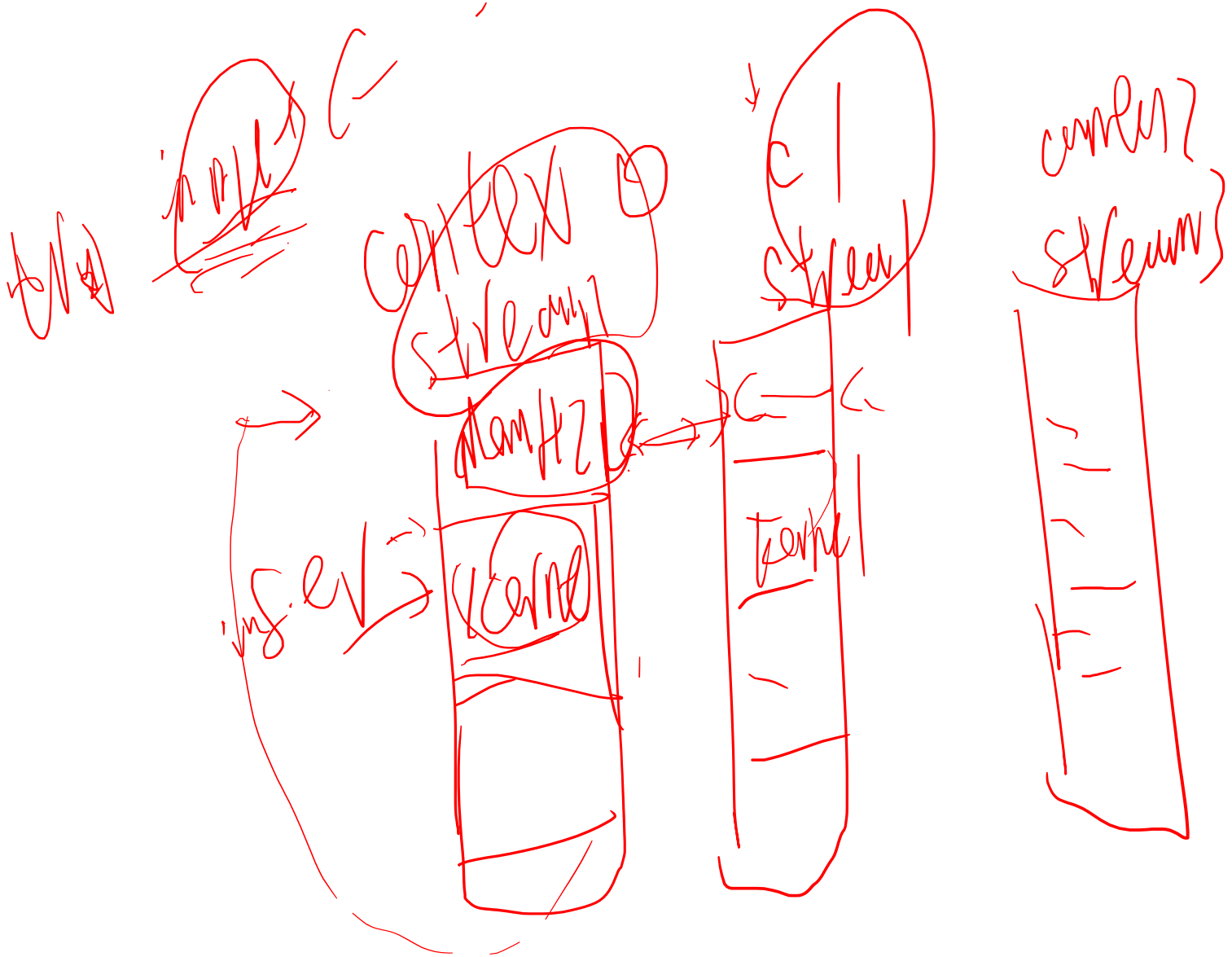
$$E\left(\frac{x}{S_k}\right)$$

S_k


kernel
weight
parameter
out type = float32


LSTM
kernel
int8
128
Scale
Weight
input
int32
Scale


build Engine
load_model (filename)
Serial
onnx
network
node
context
shared mem
kernel



考勤&反馈