Space & Compute Complexity of MLP :-

Computations / operations.

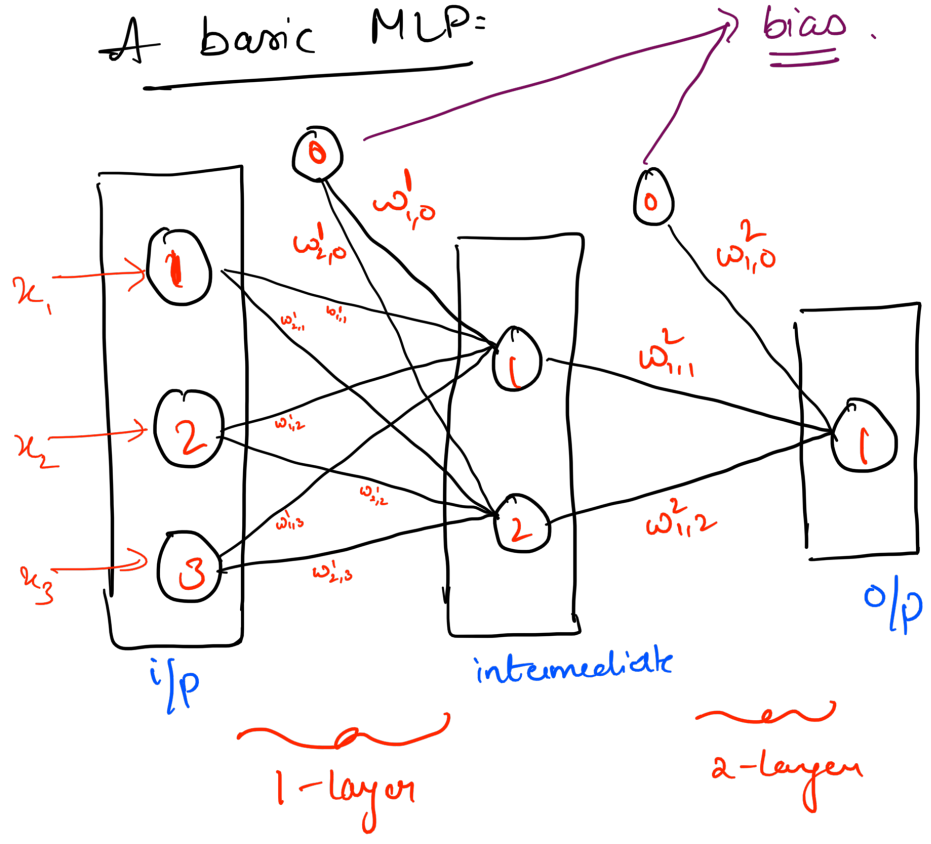
a ✓
b ✓
c ✓

$c = a + b$

1 addition operation
assigment of (a+b) into c

Computations → Energy ↓
↓
Space (may)

A basic MLP:



w^l ⇒ layer

$w_{0,i}$ → i/p neuron
→ o/p neuron

w ⇒ scalar (small, no bold)

$\underline{w}$ ⇒ vector / tensor (small, bold)

$\underline{\underline{W}}$ ⇒ matrix / tensor (caps, bold)

W ⇒ constant (caps, constant)

Q1) What amount of space does such a NN take in RAM to process an i/p ?

Q2) How many computation does such a NN take ?

↳ Eqn. of a perceptron:

$$y_0^l = w_{0,0}^l + \sum_{j=1}^{i} w_{0,i}^l \, x_j^l$$

$i \Rightarrow$ # i/p neurons

$0 \Rightarrow$ o/p neurons.

bias input is always 1 i.e., $x_0 = 1$

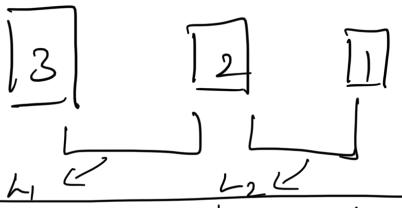
↳ assumptions $\Rightarrow$ $f(\cdot)$ in $y = f(mx+c)$ is Identity fn.

additions = i

multiplications = i

} total computations = $2i$

Eg: # neurons = 6



	L1	L2	T
A	6	2	8
M	6	2	8
T	12	4	16

⇒ operations are measured in ⇒ MAC units

↳ non-pipelined arch / implementation

$= 2 \times O \times i$

total no of o/p neurons per layer.

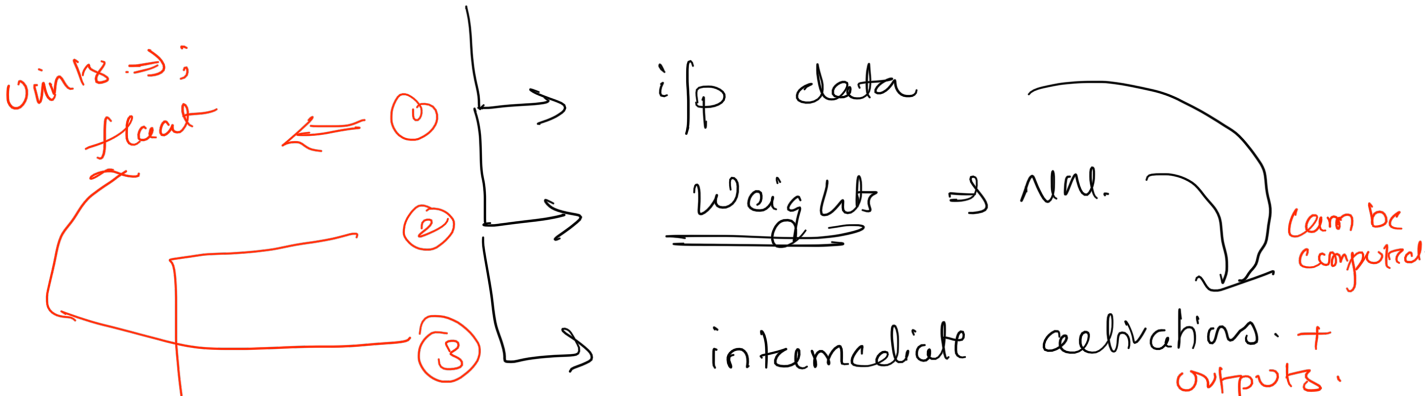
↳ pipelined arch.

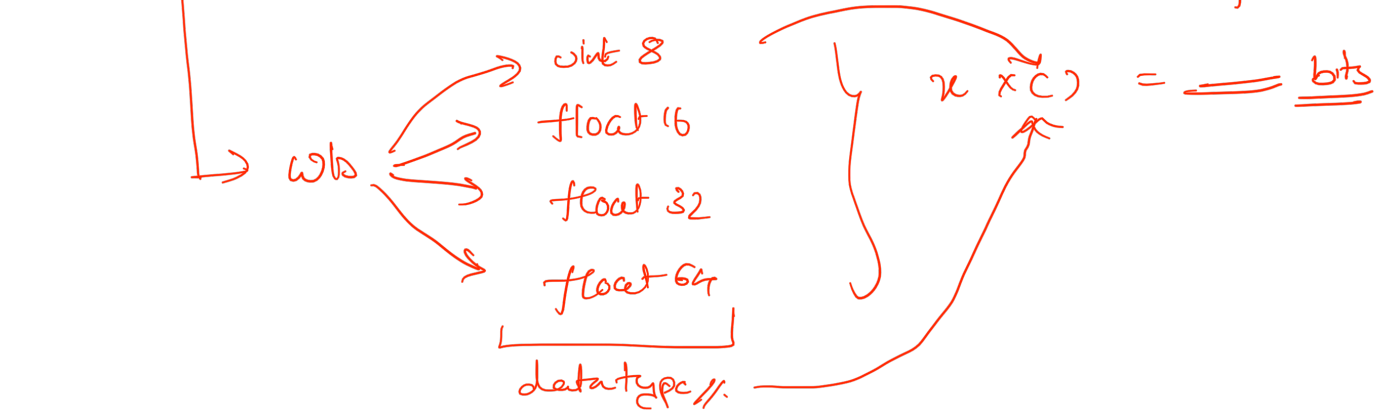
$= (i+1)\, O$ MACs

Here.

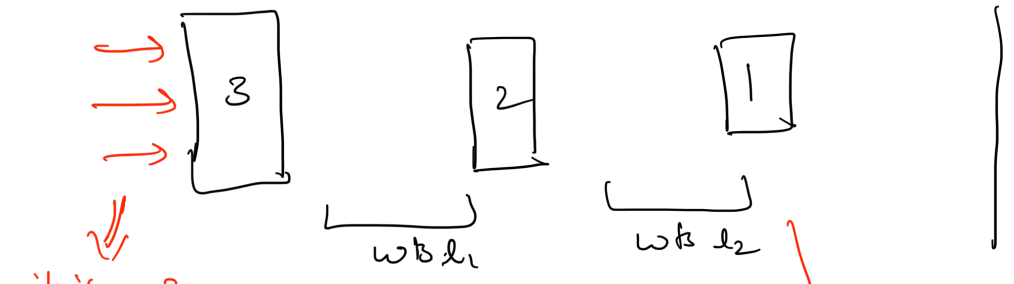
Space Complexity:

↳ Q) What occupies space in a NN? ⇒ RAM





Eg:



wts in l_1 = 8

wts in l_2 = 3

Total wts = 11

ips = 3 → i/p

ips = 2 → intermediate activations

ops = 1 → o/p

total data = 6

assuming data type ⇒ float 32 bit

Space for wts ⇒ 11 × 32 bit

" " data/activations ⇒ 6 × 32 bit

17 × 32 bit

= 544 bits

↳ min. RAM required to process '1' sample of data using given NN

When we have a batch of data flow?

↳ Space for wts is same

" " data/activations

↳ × B

batch size.

Formulas for SC:

→ # of neurons (i/p or o/p) × (datatype) × Batch size (or) product of preceding dimensions

$$SC = S_w + S_{i/p} + S_{o/p} + S_{IA}$$

$$S_w = (\#\ wts) \times \text{datatype}$$

$$\Downarrow$$

only for MLP / FC Nw

$$\#\ wts = \sum_{l=1}^{L} O^l \times (i^l + 1)$$

i^l → # of i/p neurons of layer l.

O^l → # of o/p neurons of layer l

L ⇒ total no. of layers.

$$S_{IA} = \sum_{l=1}^{L-1} \left(\#\ \text{of o/p neurons of layer } l \right) \times (\text{datatype}) \times \text{Batch size}$$

$L-1$ ⇒ Since we already considered output activations separately.

(or can be product of preceding dimensions)

$$\Downarrow$$

→ each o/p neuron has an o/p activation

amount of data / intermediate activations present.