March 27, 2023

DSML: Computer Vision.

CNNs : Introduction to Transfer learning.

Class starts @ 9:05 pm.

Recap:

→ Classification problem.

→ Conv + Pool → ① Allows us to keep positional info intact.

② Reduce the number of parameters.

→ Dataset: Color images + class labels.

* Preprocessing: (a) Resize (padding) to make inputs uniform.

(b) Normalization: [0-255] → [0-1].

✓(c) Augmentation:

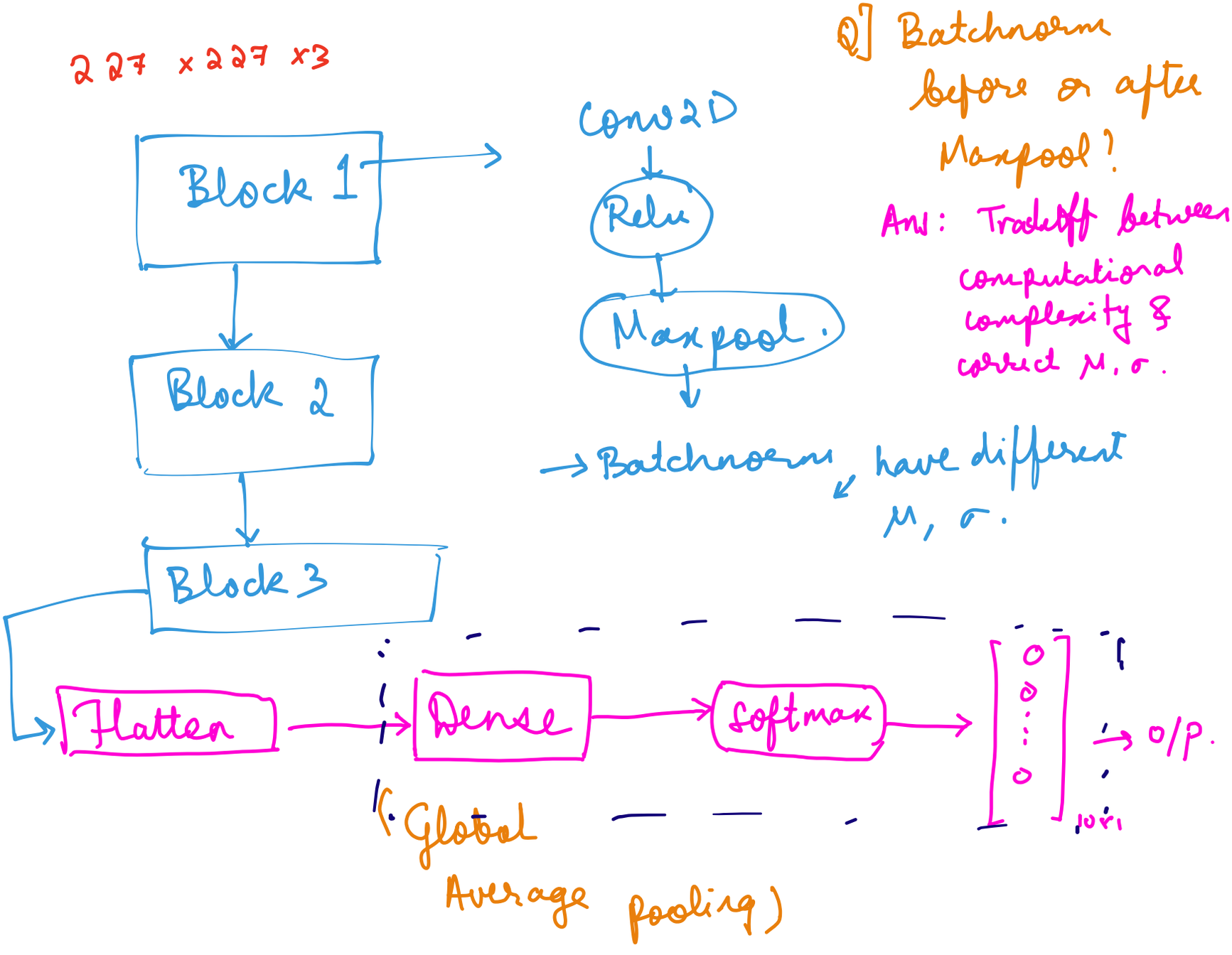
→ Design an architecture.

Same task different resolution.

Requires a lot of data. ← * Should I design from scratch everytime?

* Is it okay to just pick something that is known to work?

Saves time!! & computation.



Q] Batchnorm before or after Maxpool?

Ans: Tradeoff between computational complexity & correct μ, σ.

→ Batchnorm have different μ, σ.

Option 2: Pick something which works and use that.

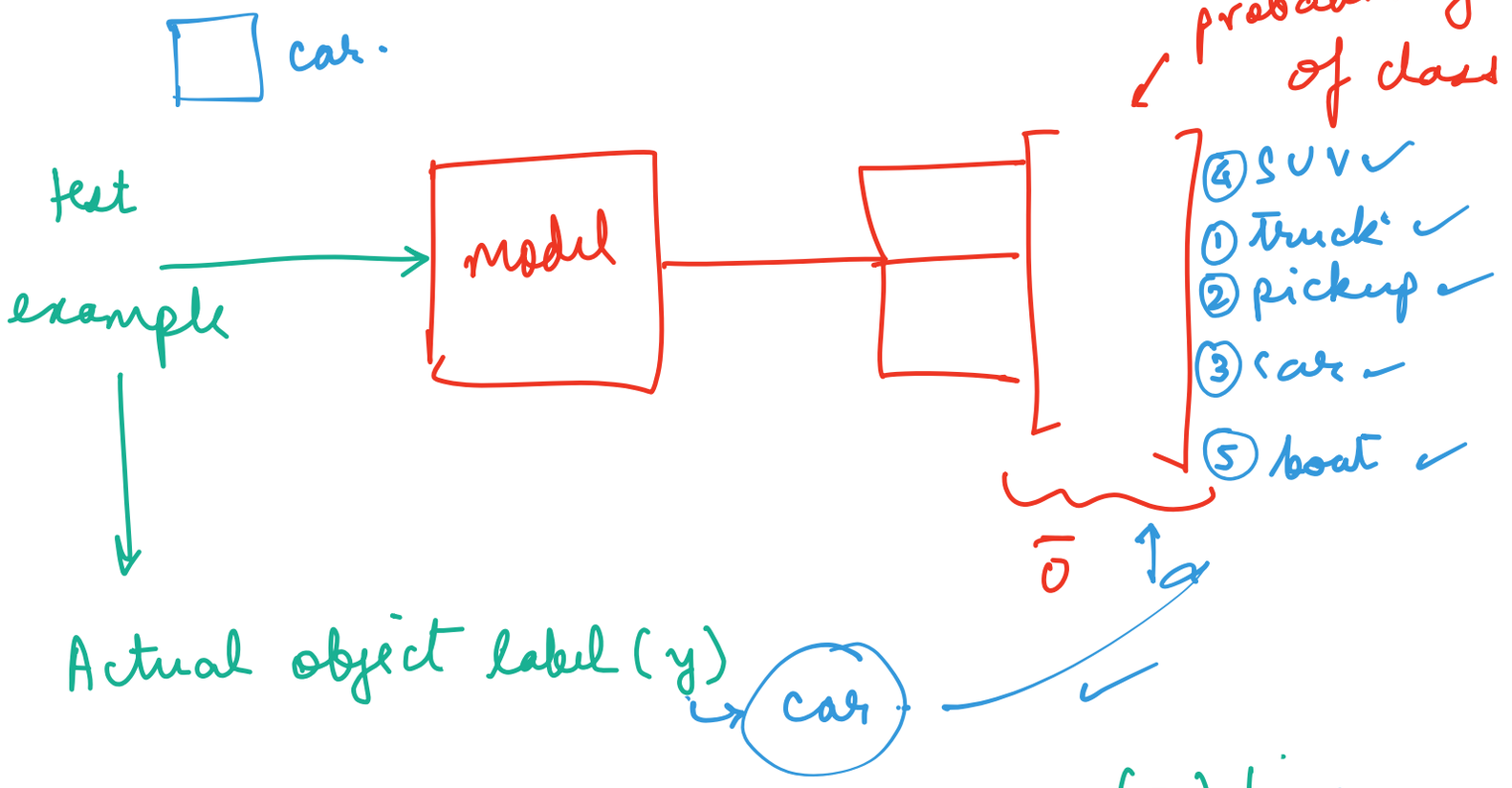
1] Should I use the architecture and train it from scratch?

✓2] Should I load the entire model which was trained, and use the pre-trained weights?

No DL vs. ✓ DL for cars.

→ Transfer learning.

How to measure accuracy of a model?



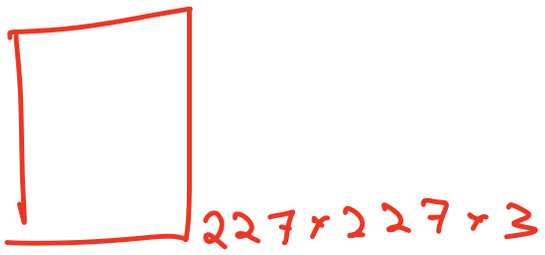
Top-1 accuracy: $|y - \text{argmax}(\bar{o})|$

Top-5 accuracy: $y_s = \{\text{top-5 labels by probability}\}$

if y in y_s:

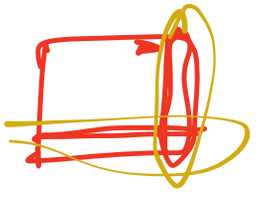
accuracy += 1

Input					227 x 227 x3
Conv1	96	11x11	4		55 x 55 x 96



11 x 11 x 96

stride : 4.



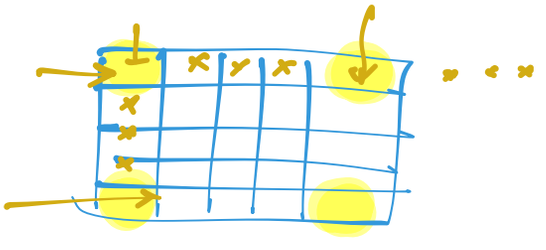
Valid : $(227 - 11 + 1) \times (227 - 11 + 1) \times 96$

$217 \times 217 \times 96$

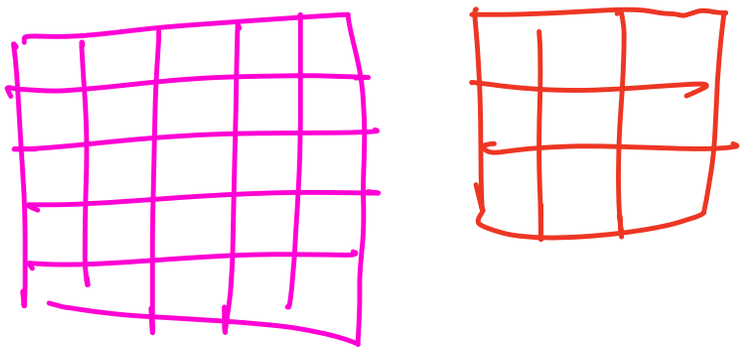
54 54

216×216

4 4

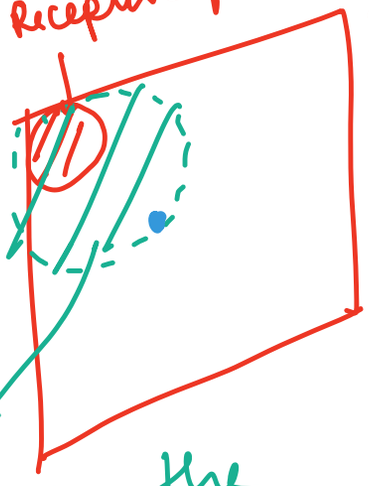


$55 \times 55 \times 96$.

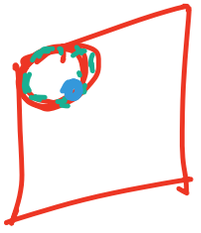


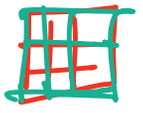
Why is 3×3 good enough?

Receptive field size for first 3×3 kernel.



Max pool





R.F size for the second 3×3 kernel.

Block: Conv + Max pool.

Layer	Filter/neurons	Filter size	Stride	Padding	Size of feature map
Input					224x224x3
Conv	64	3x3	1	same	224x224x64
Conv	64	3x3	1	same	224x224x64
Max Pool		2x2	2		112x112x64
Conv	128	3x3	1	same	112x112x128
Conv	128	3x3	1	same	112x112x128
Max Pool		2x2	2		56x56x128
Conv	256	3x3	1	same	56x56x256
Conv	256	3x3	1	same	56x56x256
Conv	256	3x3	1	same	56x56x256
Max Pool		2x2	2		28x28x256
Conv	512	3x3	1	same	28x28x512
Conv	512	3x3	1	same	28x28x512
Conv	512	3x3	1	same	28x28x512
Max Pool		2x2	2		14x14x512
Conv	512	3x3	1	same	14x14x512
Conv	512	3x3	1	same	14x14x512
Conv	512	3x3	1	same	14x14x512
Max Pool		2x2	2		7x7x512
FC					4096
FC					4096
FC					1000

1st

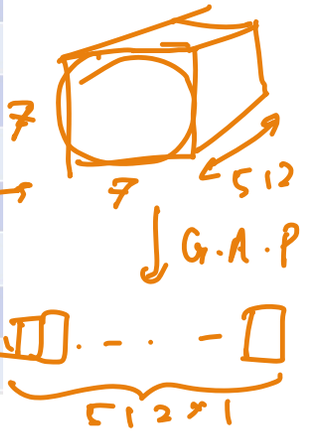
2nd

3rd

4th

5th

Dense



Q] If we have VGG trained on:

(a) Imagenet → 1000 classes, 10,000 images per class

(b) Pascal VOC → ≈20 classes, 1000 images

(c) CIFAR → Subset of Imagenet with 10 classes, 5000 images per class

(d) Medical data → < 10000

(e) MS-COCO → < 100,000

Which one to choose?

→ We want to do object detection for a small dataset.

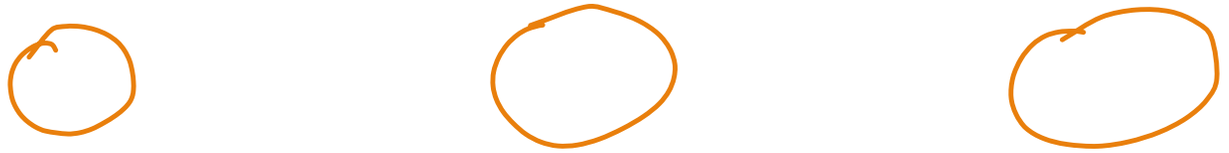
classes.

Why to keep only conv layers & get rid of dense layers?

✓ (1) Dense layers → more params.
(This results in reduced model size).

— * Spatial features in conv, not dense.
* Translation invariance in conv layers.

Visual cortex:



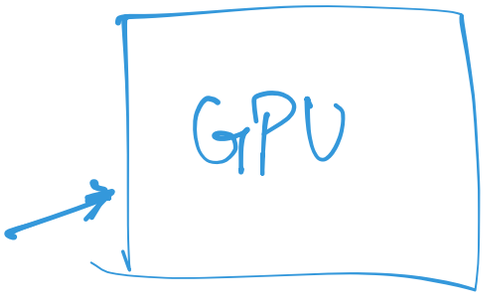
V1
lines

V2
Edges

V4
objects.

Takeaway: The first few layers of CNNs store very general features. We can use these for

almost any object detection task.



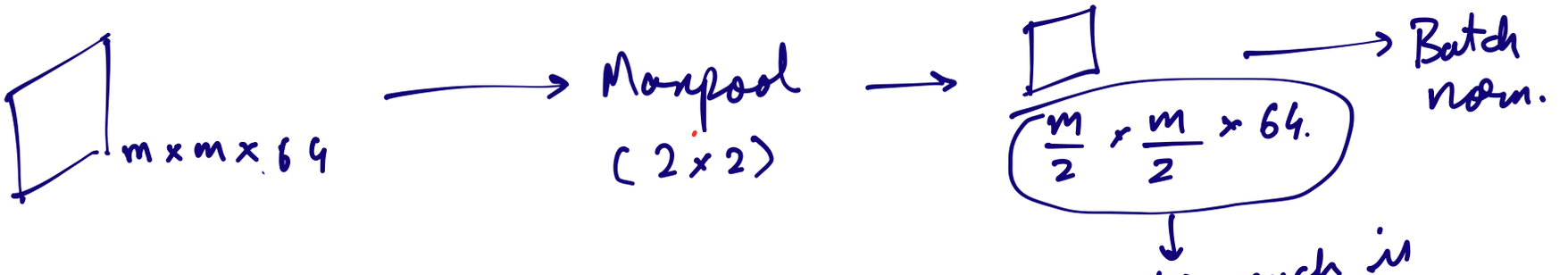
Rule of thumb:

→ Try to train / test architecture which completely occupy the GPU.

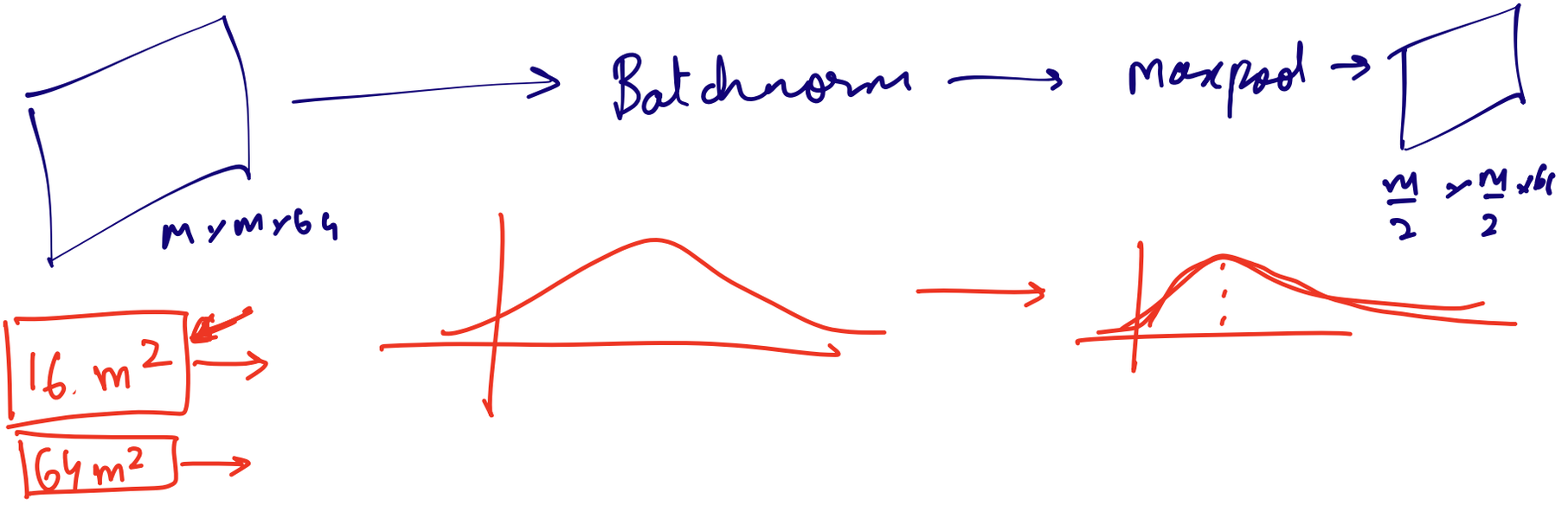
16 GB

3 GB

1] Maxpool → Batchnorm.



2] Batchnorm → Maxpool.



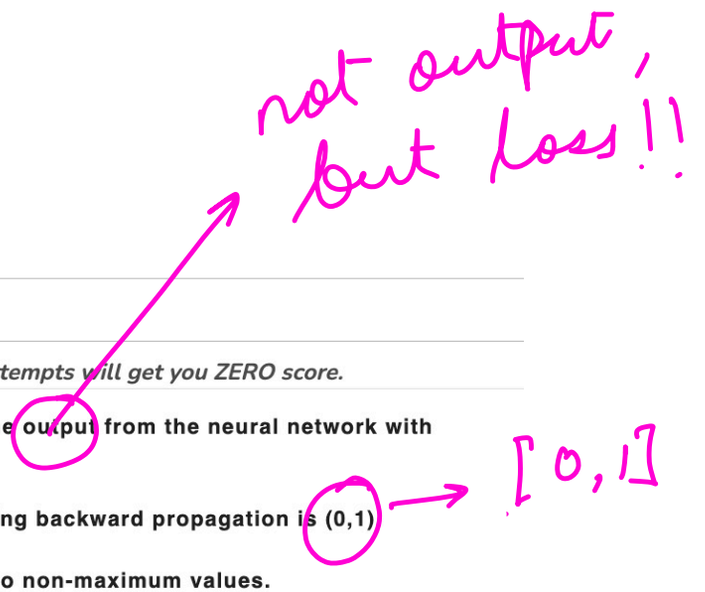
All About Backpropagation

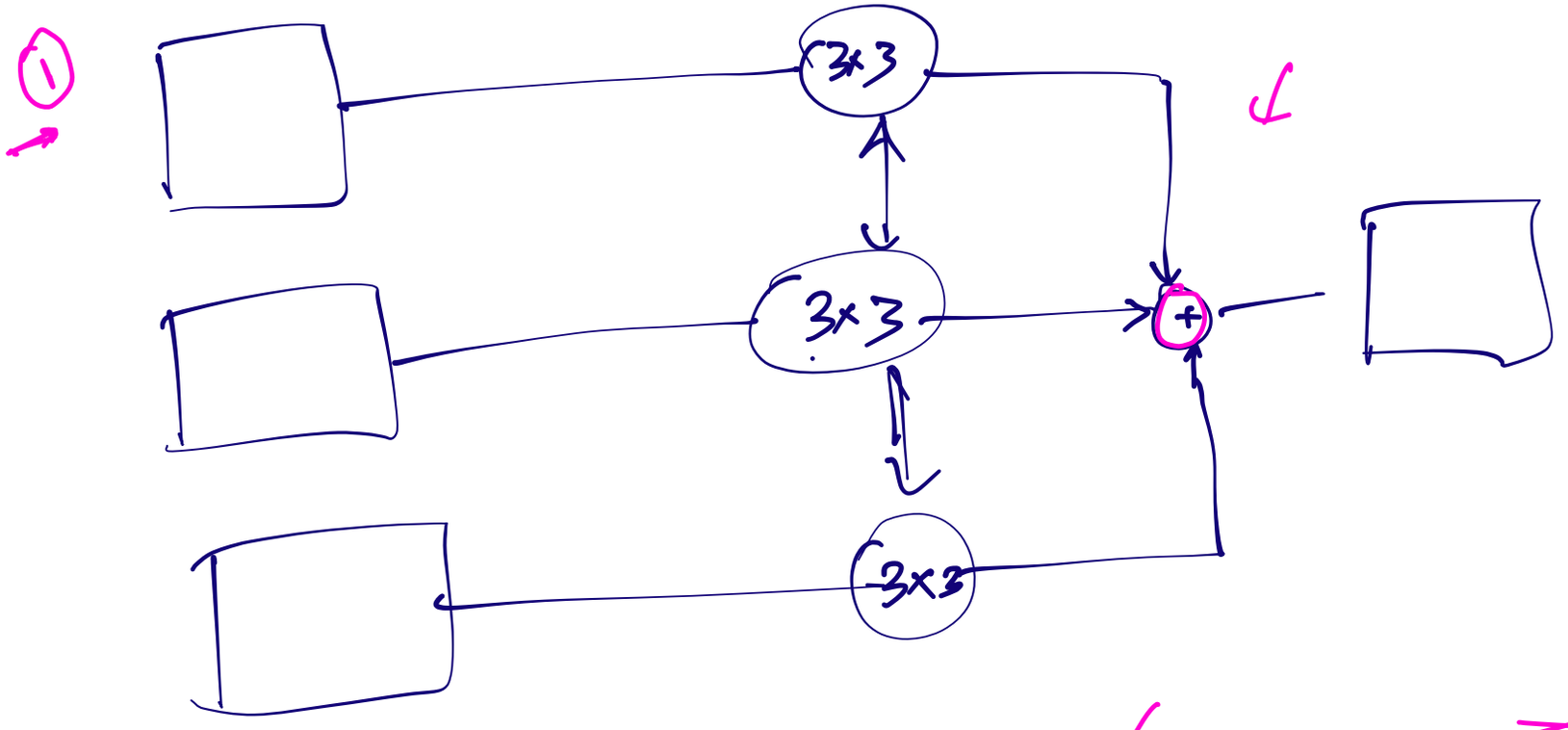
Which of the options is/are false?

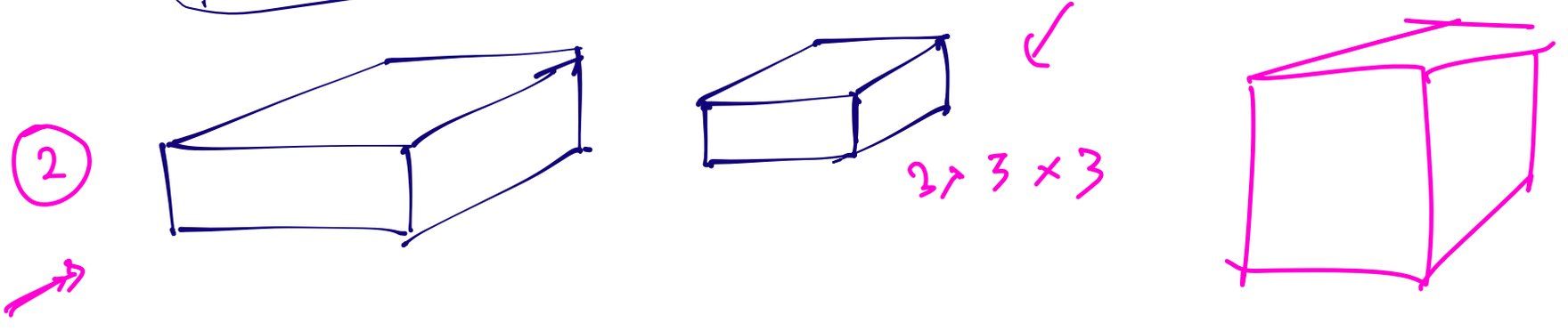
HINTS | Complete Solution

You will get full points if and only if you give CORRECT ANSWER in first attempt. All later attempts will get you ZERO score.

- [x] **In backward propagation, we get the gradients by calculating the derivative of the output from the neural network with respect to the trainable parameters.**
- [ ] **The range of values provided by the gradient of the tanh activation function during backward propagation is (0,1)**
- [ ] **In maxpool during the backward propagation, There is no gradient with respect to non-maximum values.**
- [x] **The trainable parameters in a max pooling layer are calculated using : (((shape of the width of the filter * the shape of height of the filter * number of channels in input image) + 1) * number of filters)**







3 - D conv.

128 × 128 × 5 .