Intro to Neural Networks

BA865 – Mohannad Elhamod

BOSTON UNIVERSITY

Boston University Questrom School of Business

CNNs

Convolutional Networks

A Problem of Scalability

- How many parameters in this network?
- Do we really need to learn all these parameters?

3 x 224 x 224 = ~150k pixels

Figure courtesy of Robert Bond

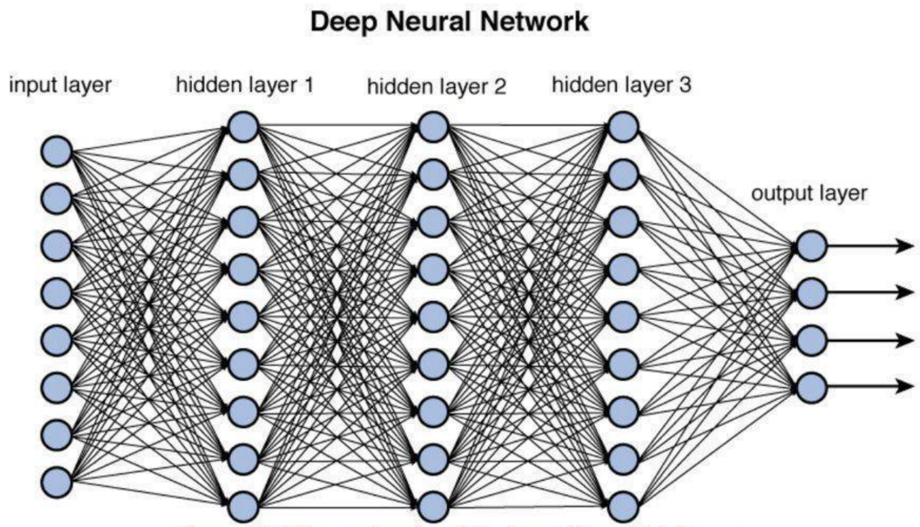


Figure 12.2 Deep network architecture with multiple layers.

Figure courtesy of Ravindra Parmar

Structure in Images

- Interesting images have:
  - Locality of information.
  - Spatial invariance.

Figure courtesy of Robert Bond

vs.

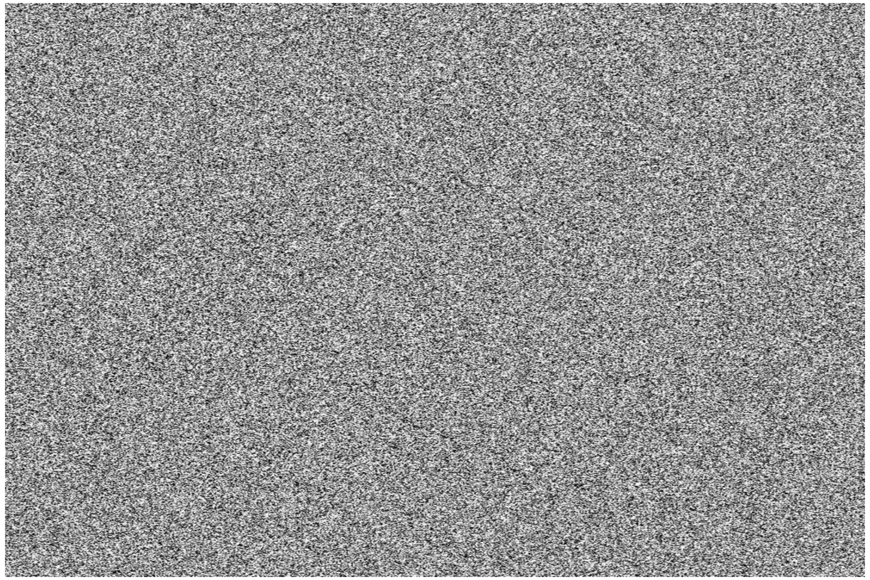


Figure courtesy of Jorge Stolfi

Convolutional Filters

- Instead of learning a mesh of all possible parameters, let's learn local descriptors ***(kernels or filters)*** that can be reused across the image!

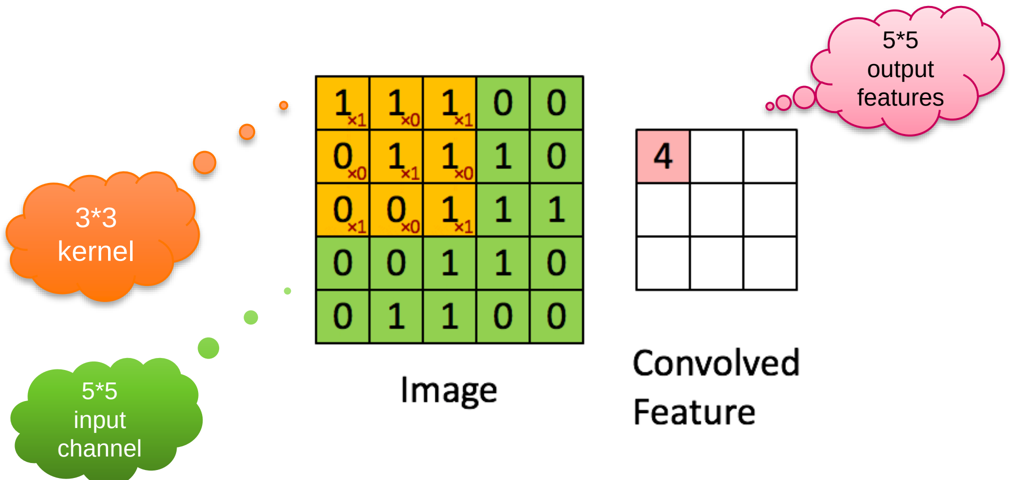


Figure courtesy of Daniel Nouri

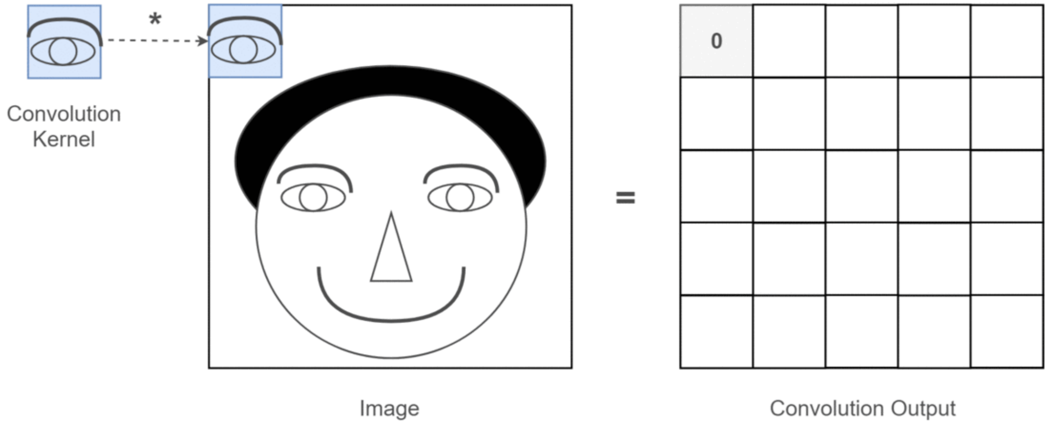
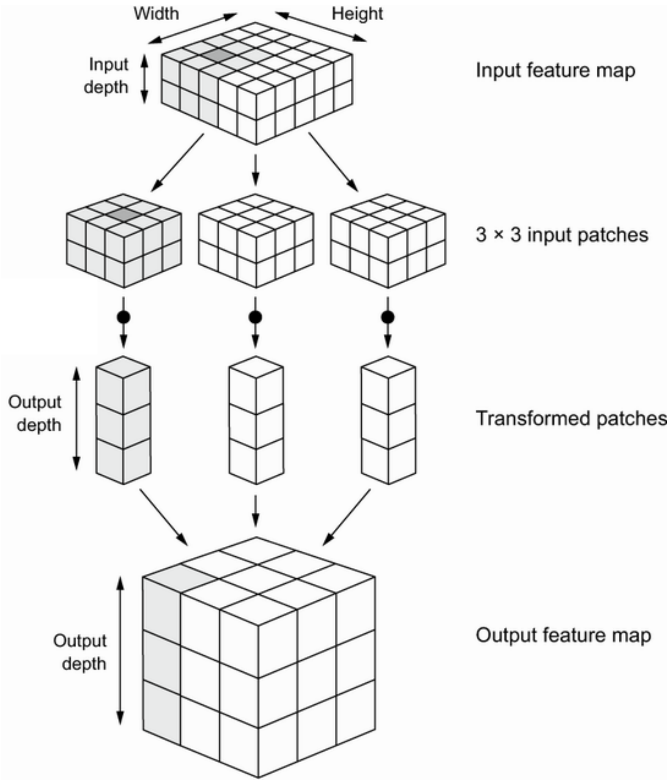


Figure courtesy of Thushan Ganegedara

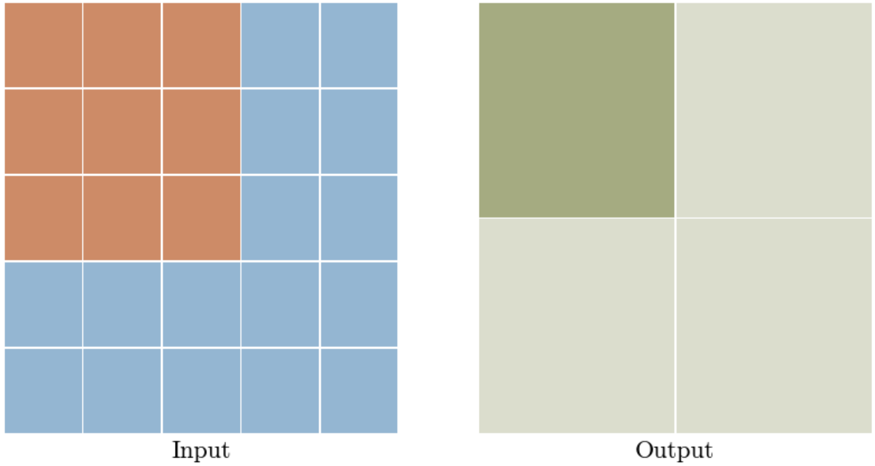
Mathematically Speaking…



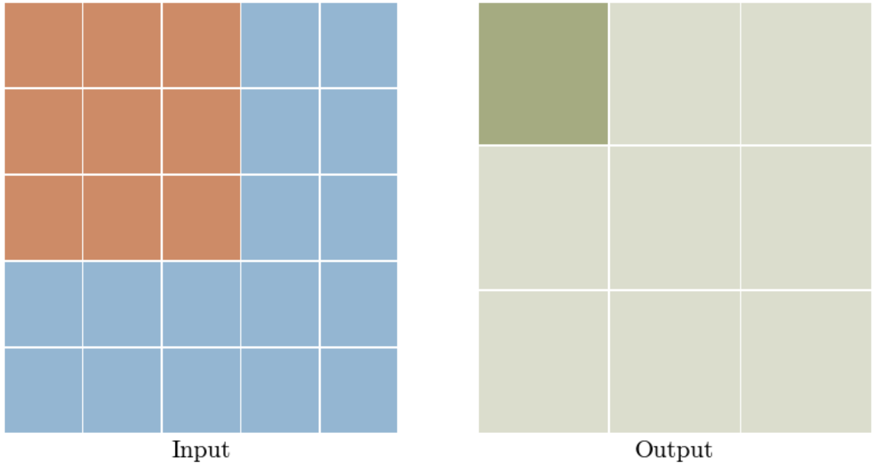
Stride and Padding

- Stride controls feature field overlap.
- Padding controls the down-sampling.

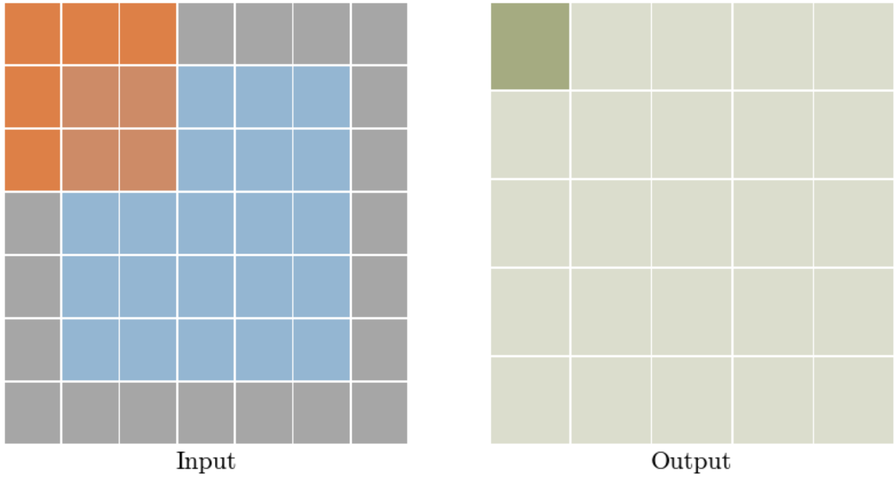
Type: conv - Stride: 2 Padding: 0



Type: conv - Stride: 1 Padding: 0



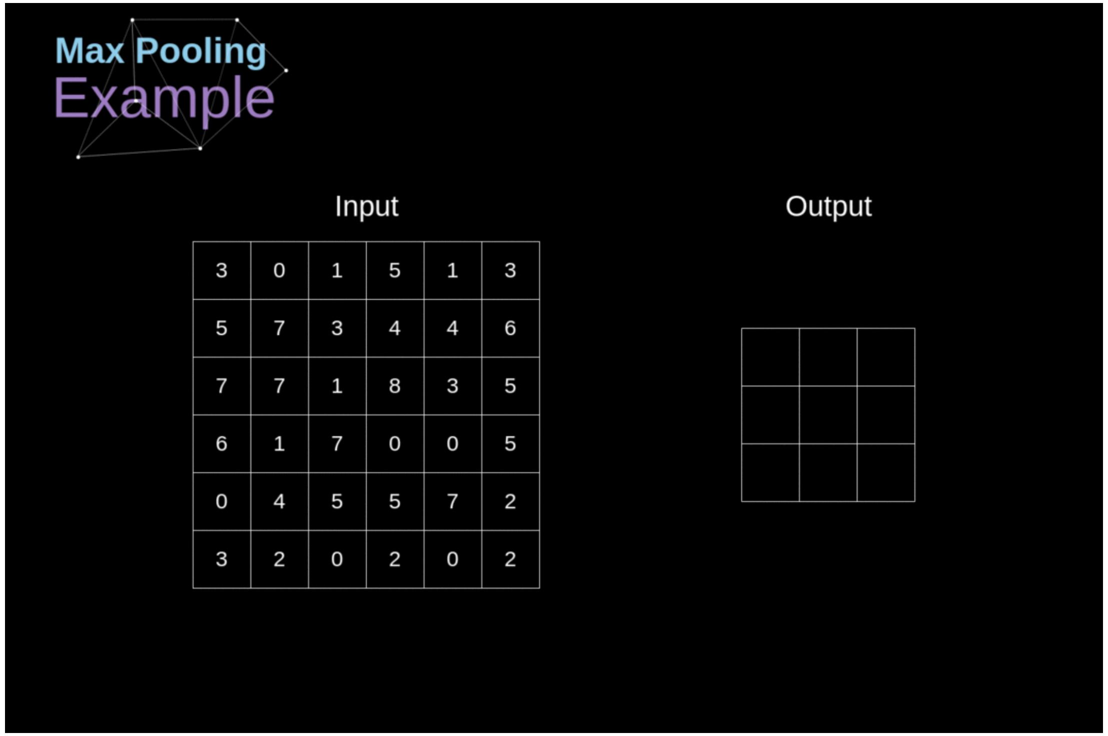
Type: conv - Stride: 1 Padding: 1



Aqeel Anwar

Non-Linearity: MaxPooling

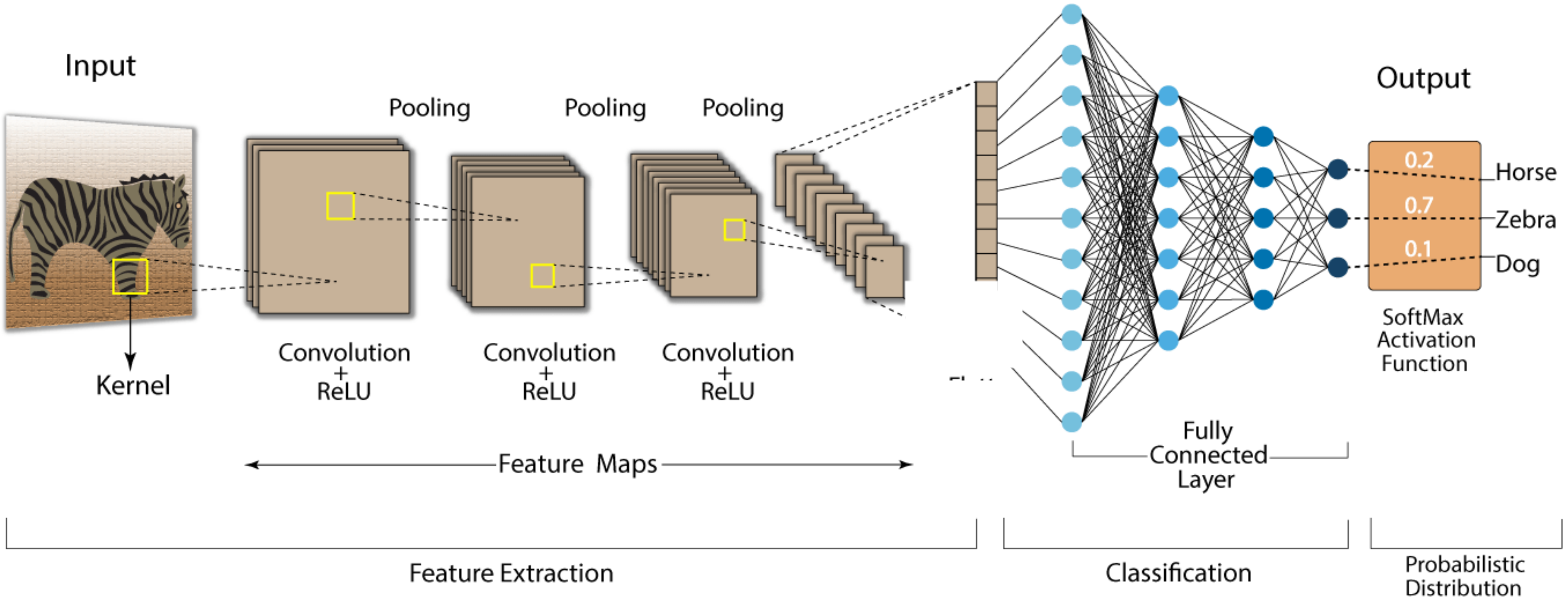
- In addition to being a non-linearity…
  - it helps down-sample the image.
  - It helps summarize information in terms of larger blocks.



3	0	1	5	1	3
5	7	3	4	4	6
7	7	1	8	3	5
6	1	7	0	0	5
0	4	5	5	7	2
3	2	0	2	0	2

Putting It All Together

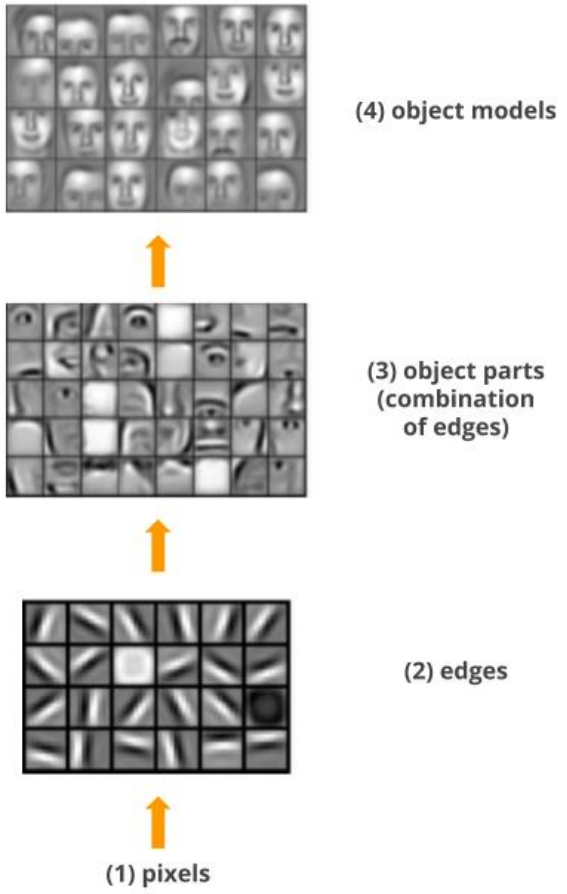
- Deeper layers generally have more kernels that are smaller.



Afaque Umer

Learned Features

- Early layers learn low-level features.
  - spots, edges, etc.
- Later layers learn to detect high-level features as a combination of low-level features.
  - Eyes, ears, hair, etc.
- [Demo]()



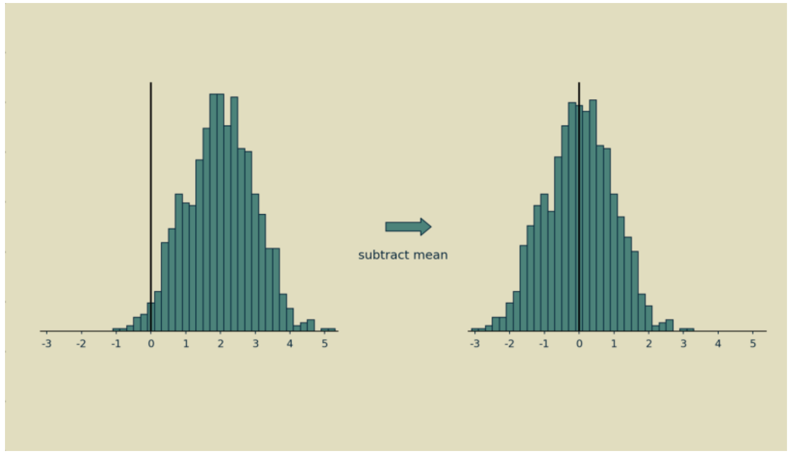
https://micro-dimensions.com

Hyper-Parameters

Continued…

Batch Normalization

- Even if input data is properly normalized, the gradient in subsequent layers may vanish or explode.
- For that, you may add <u>*"batch normalization"*</u> at every layer.



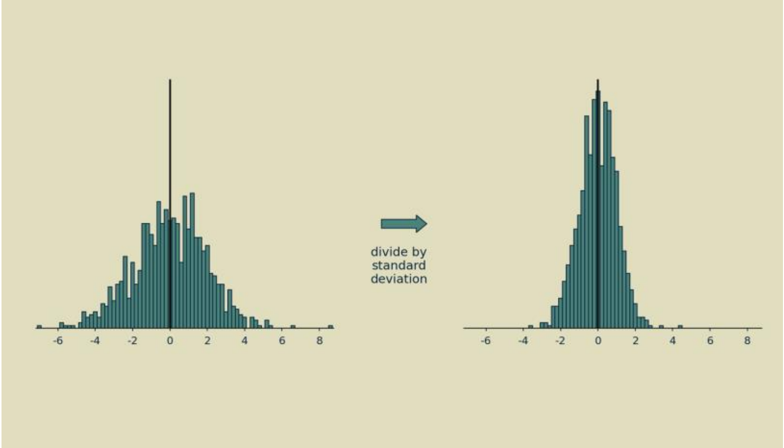


Figure courtesy of Brandon Rohrer

Learning Rate: Schedulers

- Since larger learning rates may converge faster but smaller ones are more stable, you could adjust the learning rate in phases to get the best of both worlds!
  - This way, you still converge but faster.

- Using a scheduler is a common practice.

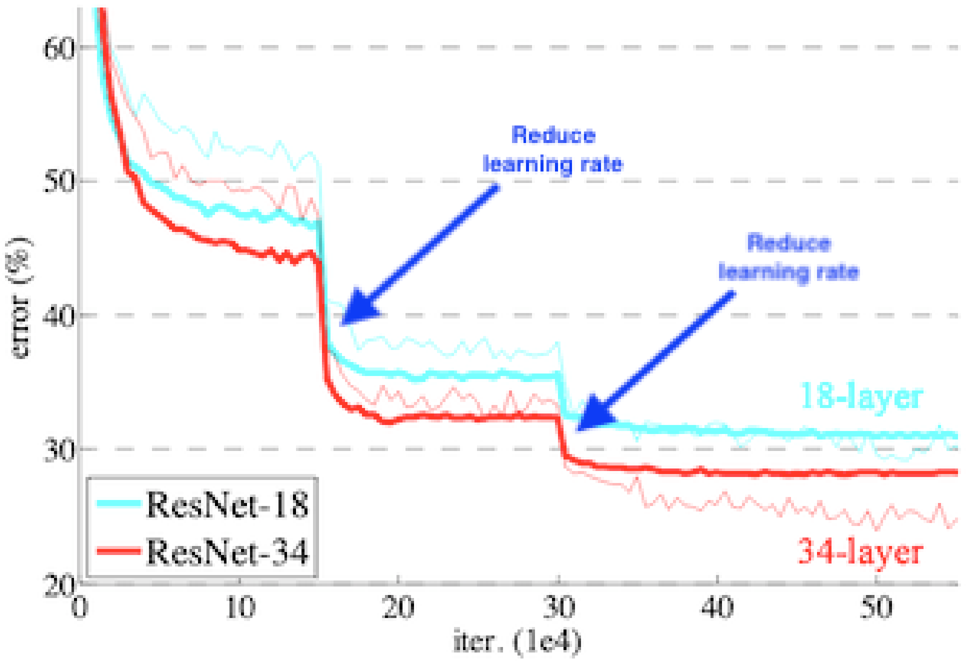


Figure courtesy of B. D. Hammel

Hyper-Parameter Tuning

BOSTON UNIVERSITY

Be Smart About It

- It is expensive!
  - 1 hyper-parameter with 3 values → 3 experiments
  - 2 hyper-parameter with 3 values each → 9 experiments
  - 3 hyper-parameter with 3 values each → 27 experiments
  - … exponential growth!

Be Smart About It

- It is expensive!
- Start with generally accepted wisdom:
  - Start with good initial guesses.
  - Different settings work better for different models/problems (e.g., SGD + momentum for computer vision vs. Adam otherwise)
- Be picky about what to fine-tune.
  - Use early stopping.
  - Learning rate is the most important parameter!

Hyper-Parameter Tuning Methods

- Generally, use log-scale for numerical hyper-parameters.
- Random and Adaptive searches generally find optimal values faster than grid searches.

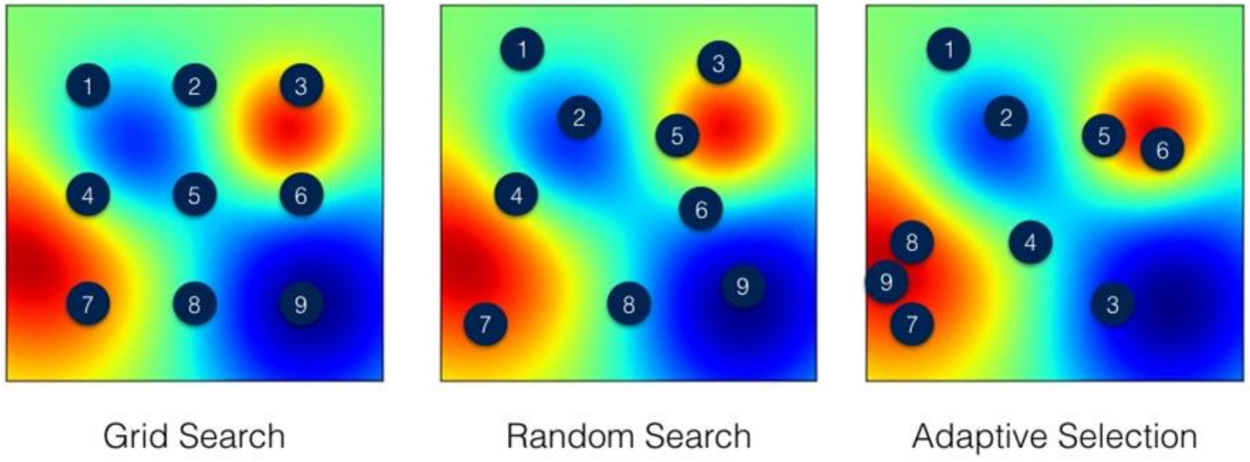


Figure courtesy of Liam Li