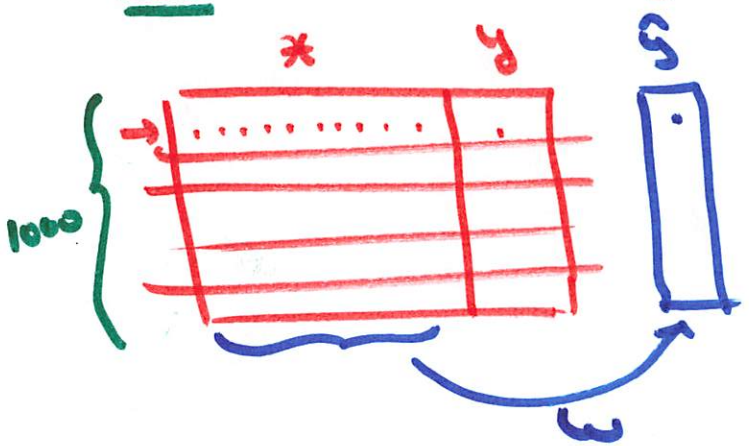
Agenda

- The Challanges: Over fitting & local optima
- The Training
  - Epochs, Batch Size, Iterations
  - Gradient Descent (GD) Vs Stochastic GD (SGD) Vs Mini-Batch GD
  - SGD with momentum
  - Learning rates and adaptive learning rates
  - Weight Initialization
  - Batch Normalization
- Guarding against over-fitting
  - L1/L2 Regularization
  - Data Augumentation
  - Drop outs
- Neural Network Architectures

GD

$y = f(x)$?



$$L(\omega) = \frac{1}{N}\sum_{i=1\ldots\ldots1000}\left(\hat{y}_i - y_i\right)^2$$

Slope of L w.r.t ω

∇L

$$\omega^{n+1} = \omega^n - \eta \nabla L^n$$

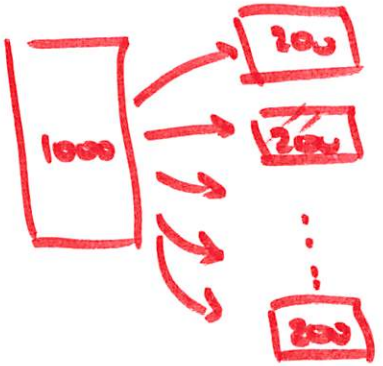
$\omega^0 = ?$

SGD

(GD) $L(w) = \frac{1}{N} \sum_{i=1 \ldots 1000} (\hat{y}_i - y_i)^2$

(SGD) $L(w) = (\hat{y}_i - y)^2$ ← randomly chosen

Mini Batch SGD



$L(w) = \frac{1}{N_b} \sum_{i \in B} (\hat{y} - y)^2$

$N = 1000$

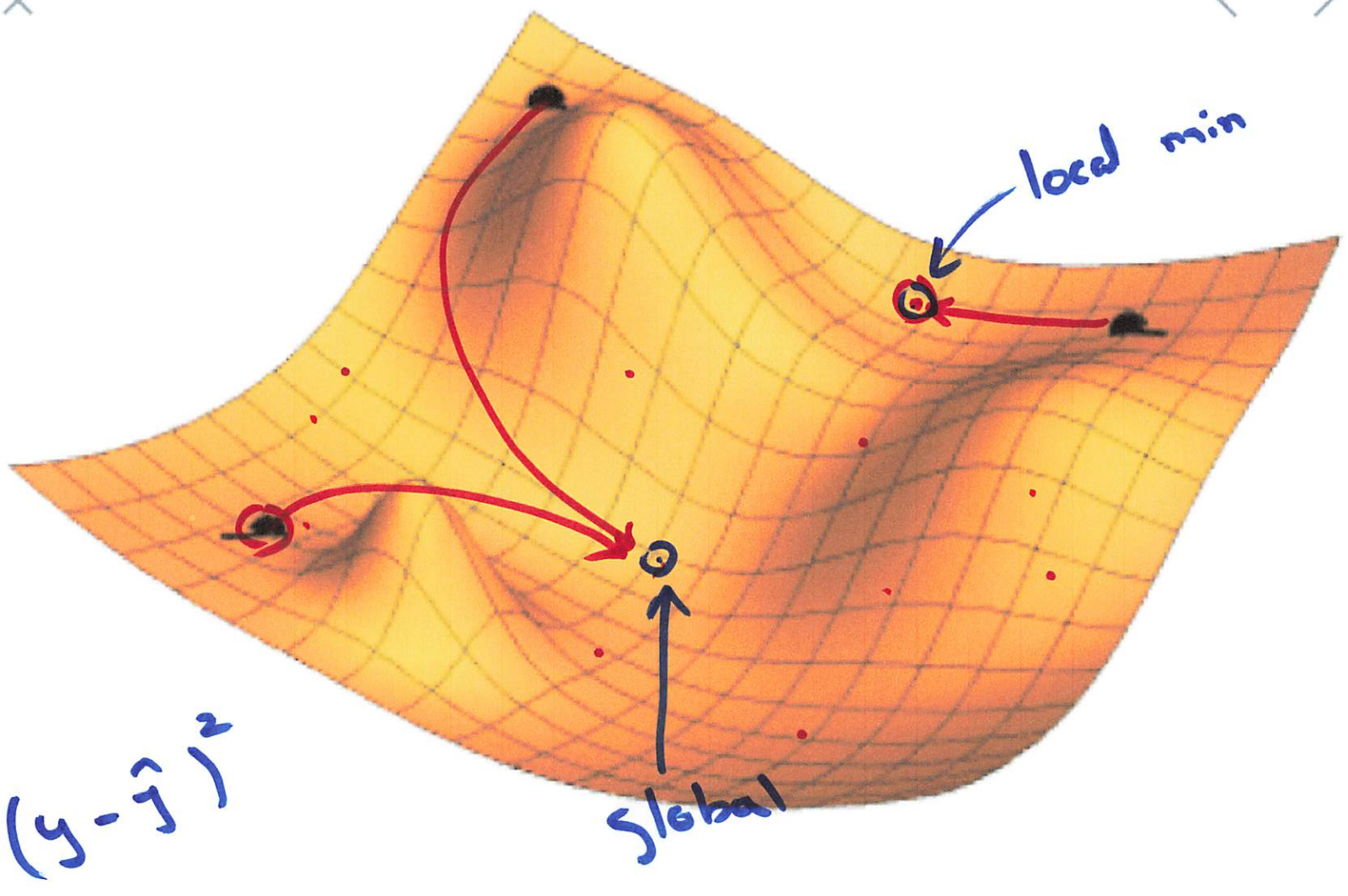
$N_b = 200$

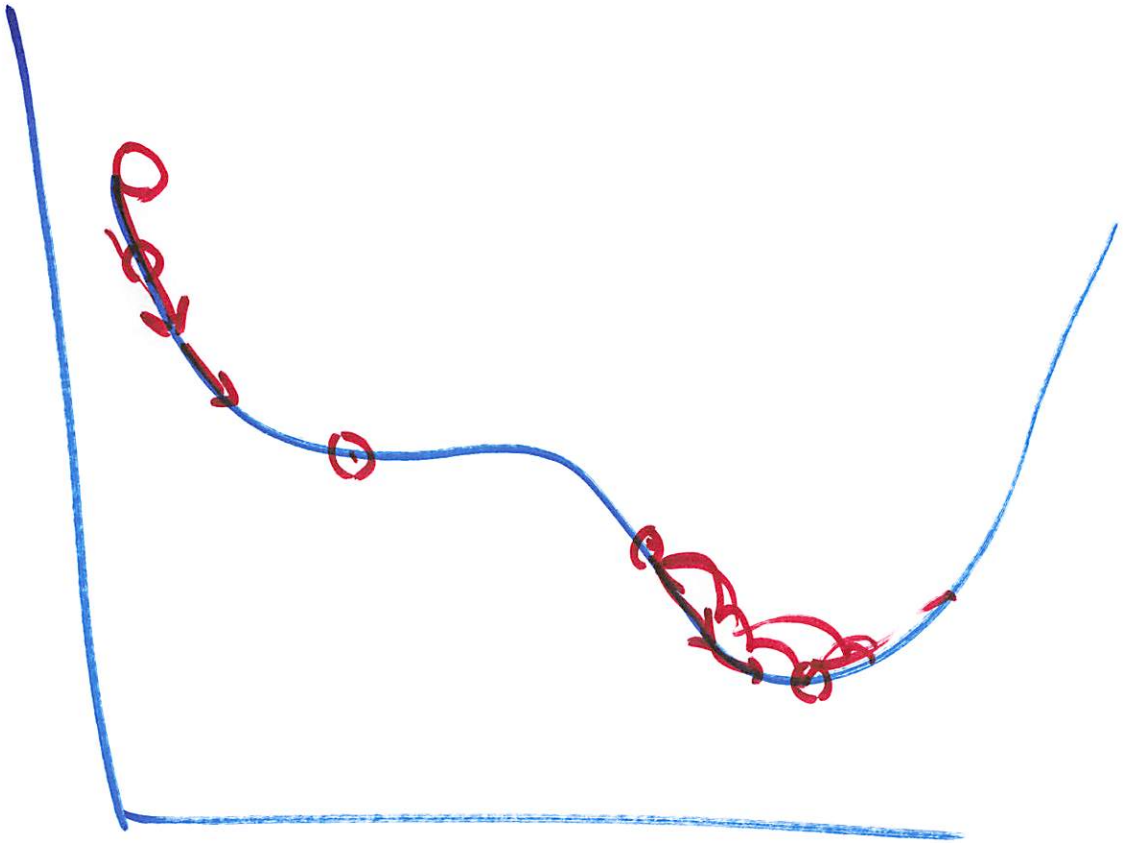
iter = 5

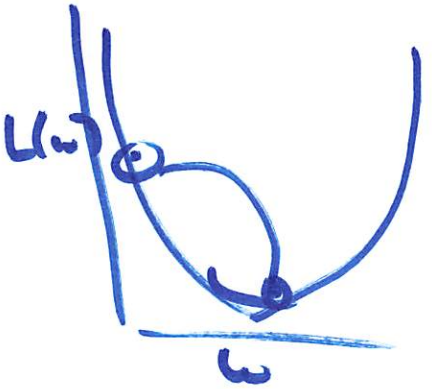
epoch

iterations = $\frac{N}{N_b}$

batch size = N_b







$$w^{n+1} = w^n - \eta \nabla L^n$$

SGD with momentum → $w^{n+1} = w^n - \eta(\alpha \nabla L^n + (1-\alpha)\nabla L^{n-1})$

Learning rate

- Choosing the Learning rate (η)
  - Too small, we will need too many iterations for convergence
  - Too large, we may skip the optimal solution
- Adaptive Learning Rate :
  - start with high learning rate and
  - gradually reduce the learning rate with each iteration.
  - Moreover, having different learning rates for different weight updates will help: Adagrad, RMS Prop

Adam

AdaDelta

Adaptive Grad

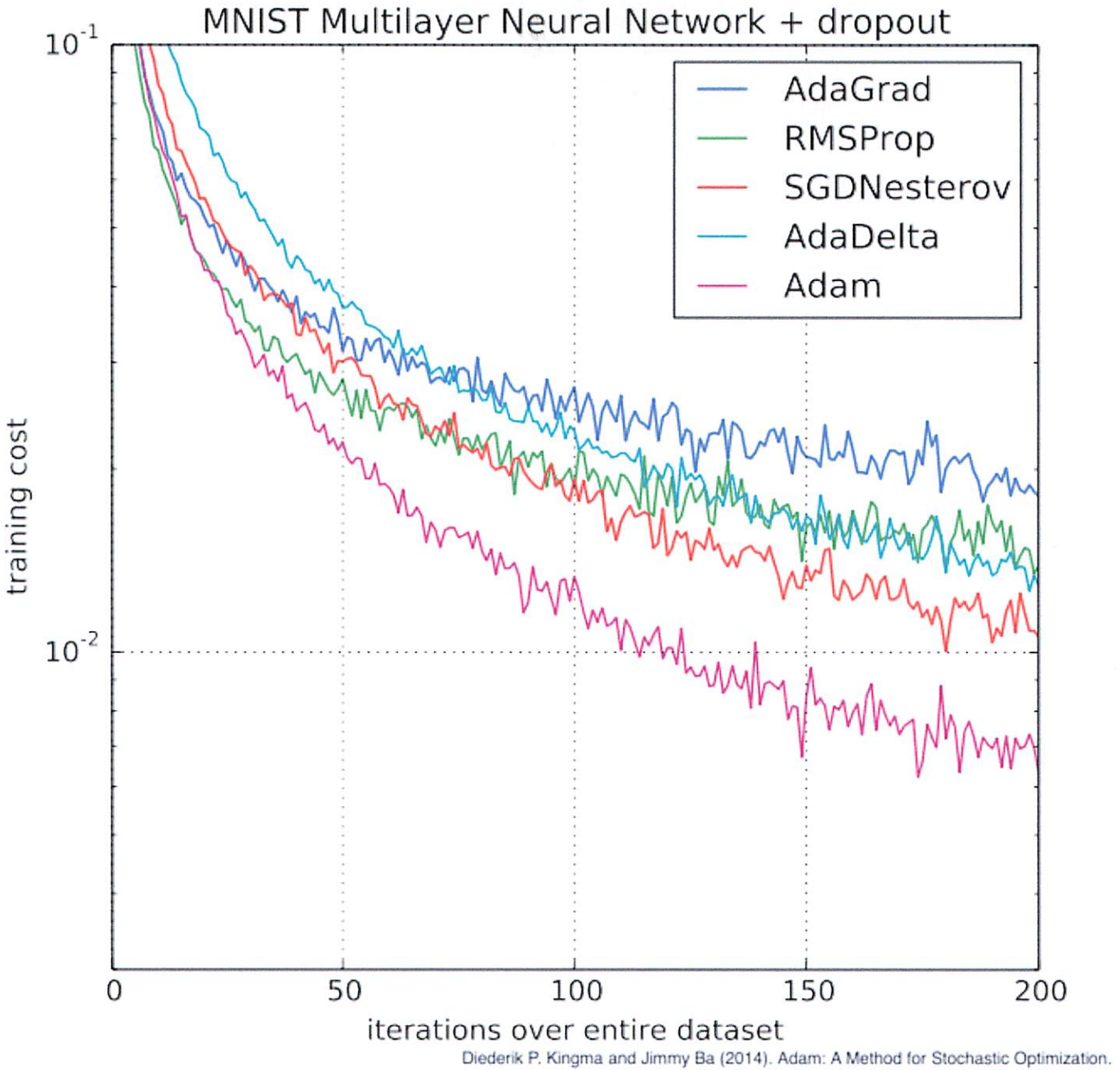
$$\omega^{n+1} = \omega^n - \frac{2}{\sqrt{s^n + \epsilon}} \nabla L^n$$

$$s^n = \sum_m (\nabla L^m)^2$$

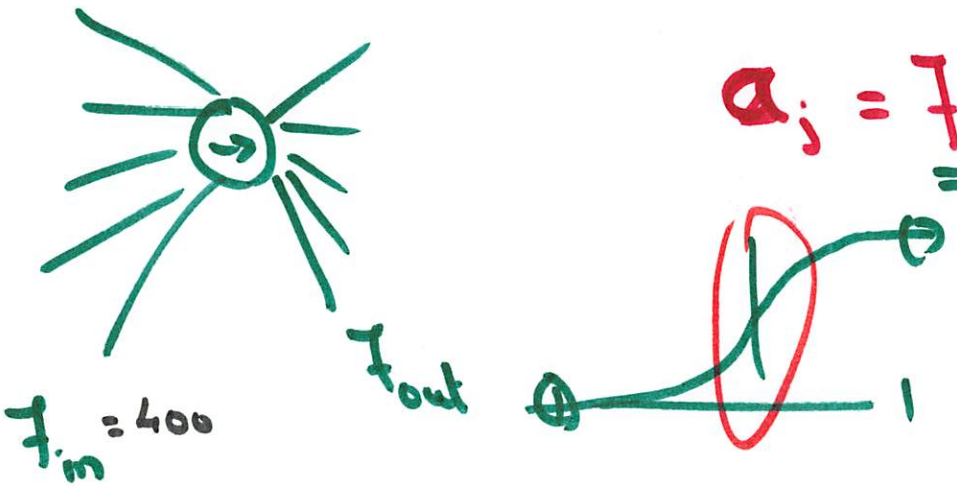
RMS Prop

$$\omega^{n+1} = \omega^n - \frac{2}{\sqrt{s^n + \epsilon}} \nabla L^n$$

$$s^n = \alpha s^{n-1} + (1-\alpha) \nabla L^n$$



Diederik P. Kingma and Jimmy Ba (2014). Adam: A Method for Stochastic Optimization.



$$a_j = \sigma(\underbrace{\Sigma w_{ij} x_i + b_j}_{z_j})$$

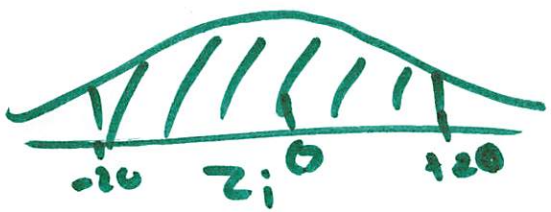
$w_{ij} = 0$

$w_{ij} = \text{Normal}(0, 1)$

Unstable gradients

∇L^n → exploding grad

∇L^n → Vanishing grad



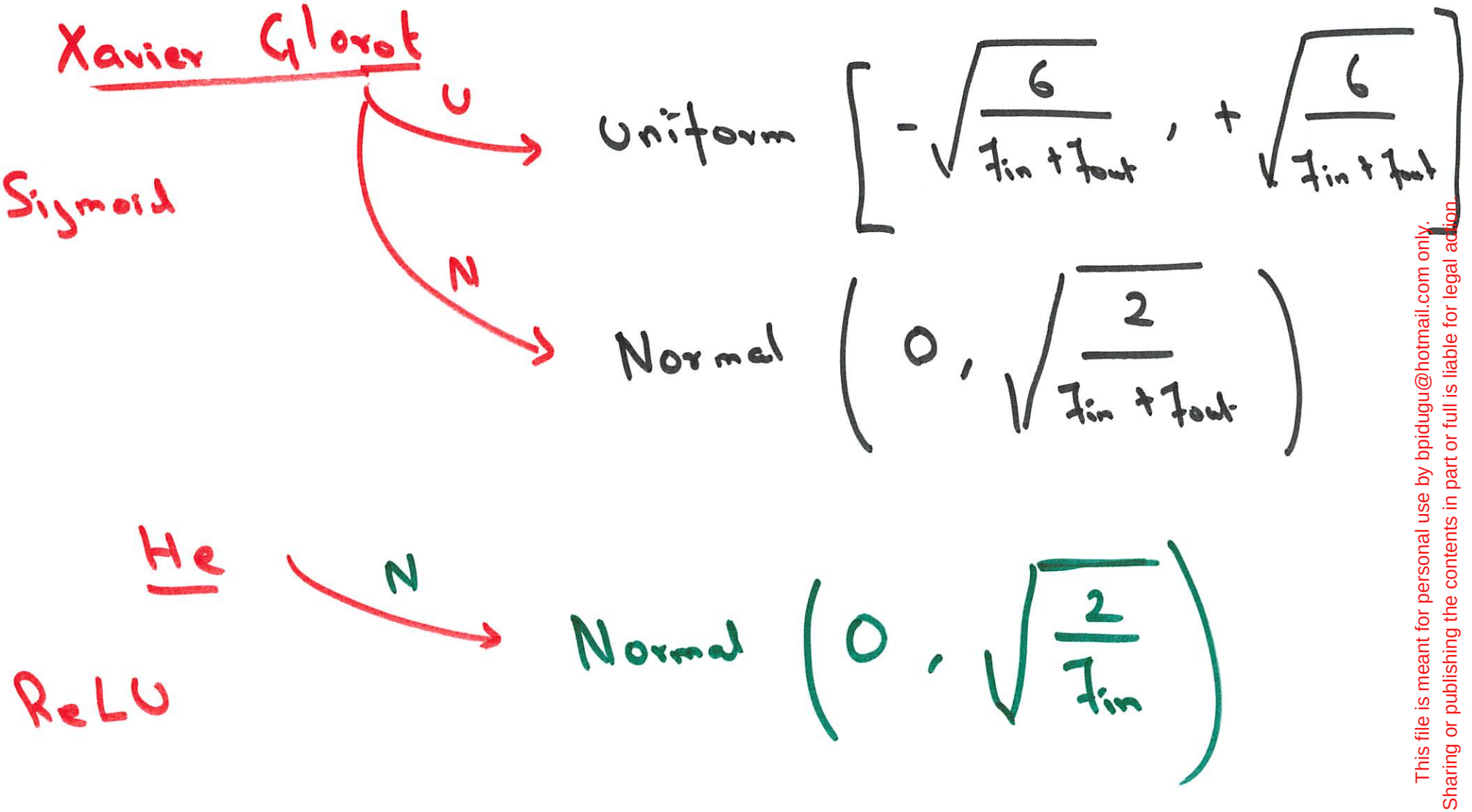
w_{ij} $N(0,1)$ ↑

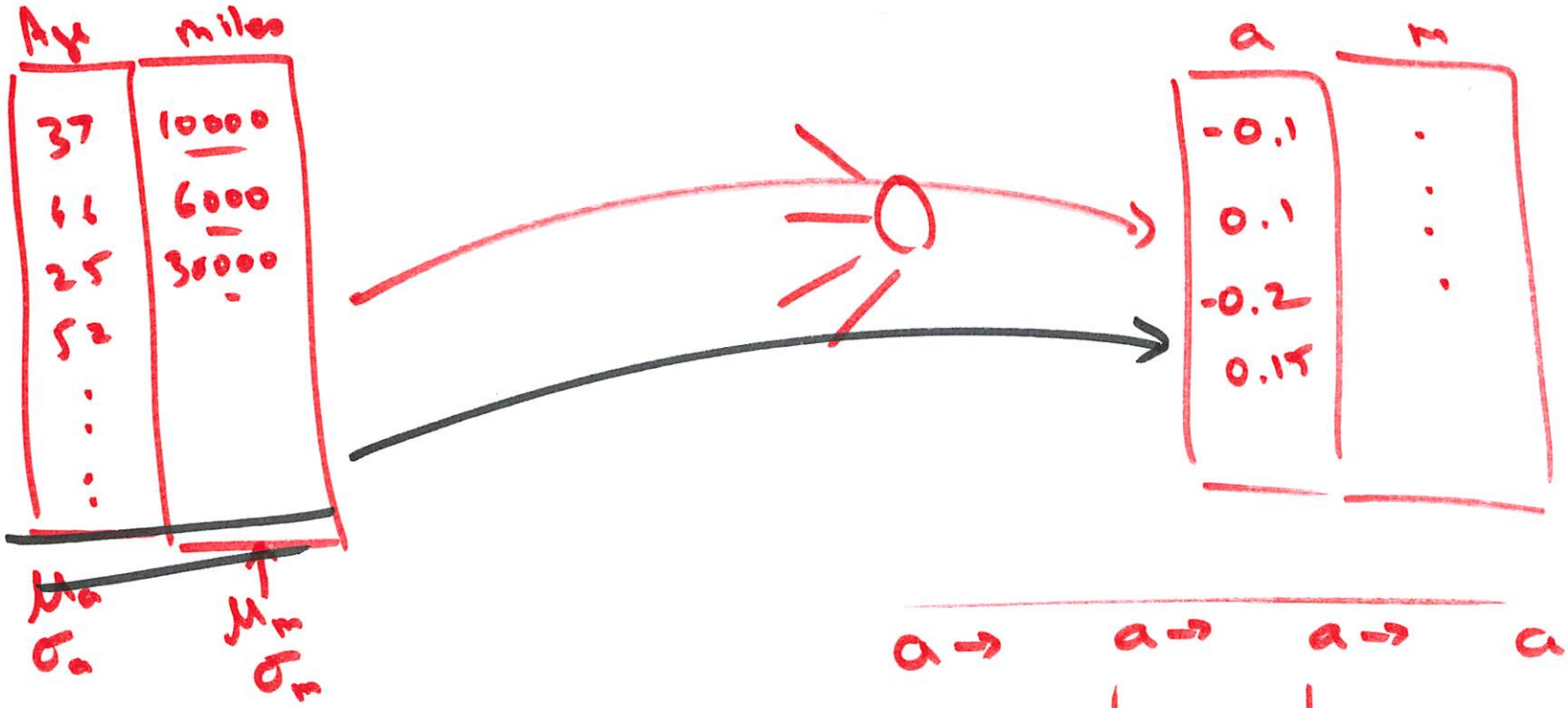
$$z_j = \sum_i w_{ij} x_i + b_j \to 0$$

x_i → 1

$\text{Var}(z_j) = 400$

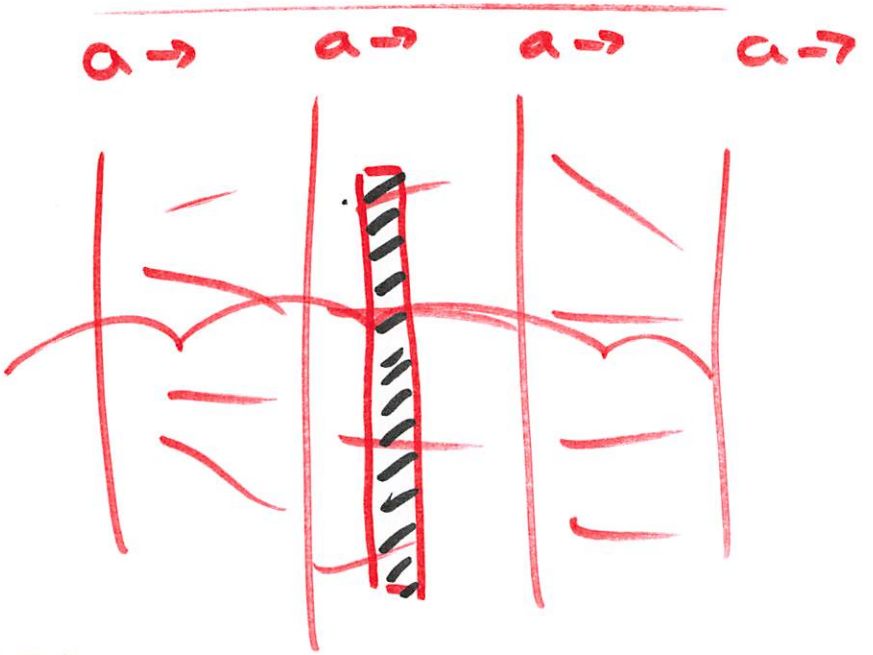
$\text{std dev}(z_j) = 20$





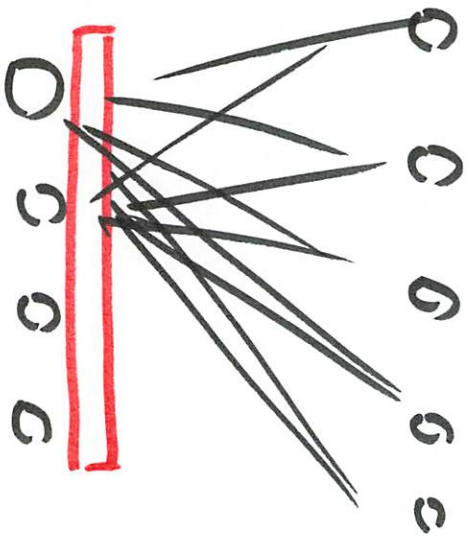
$$\hat{x}_a = \frac{x_a - \mu_a}{\sigma_a}$$

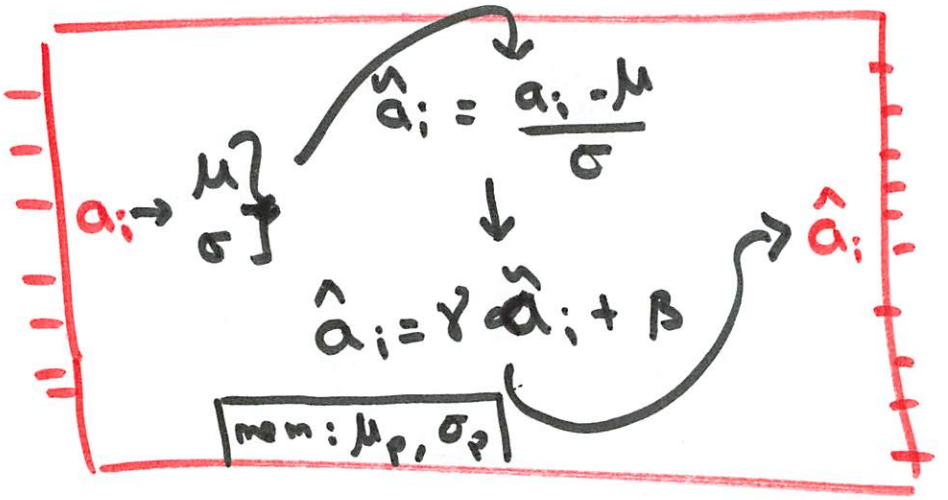
$$\hat{x}_m = \frac{x_m - \mu_m}{\sigma_m}$$

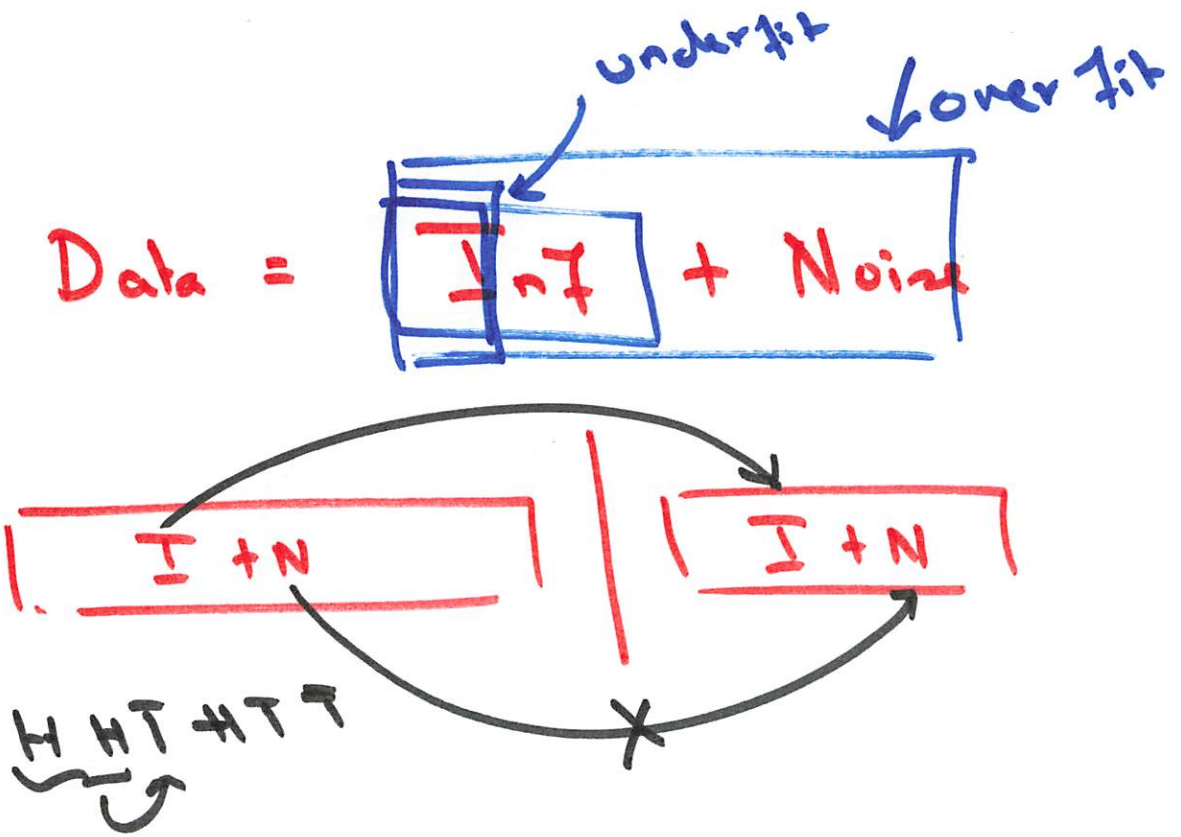


Layer → Forward → Stable

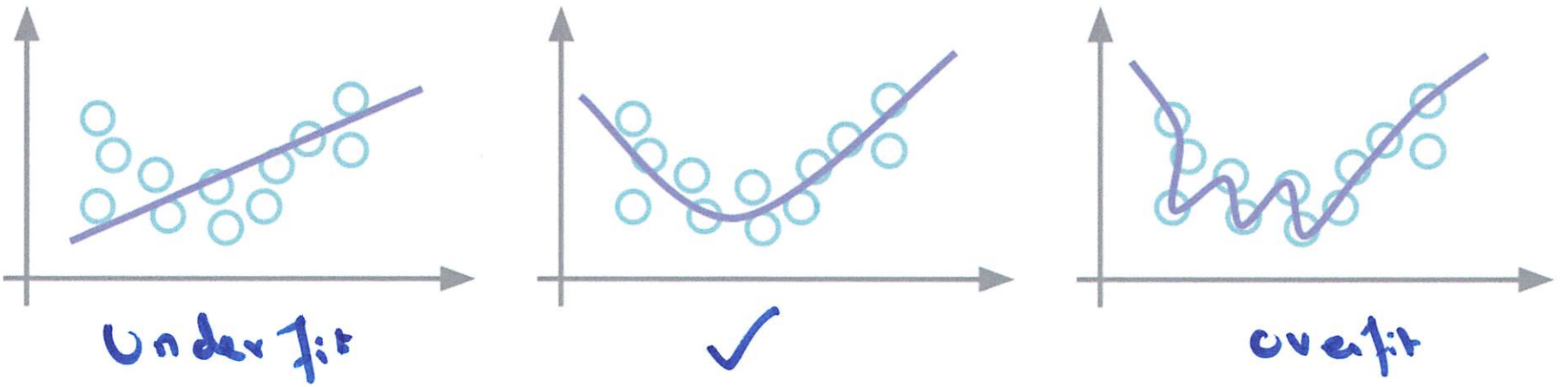
Covariate Shift

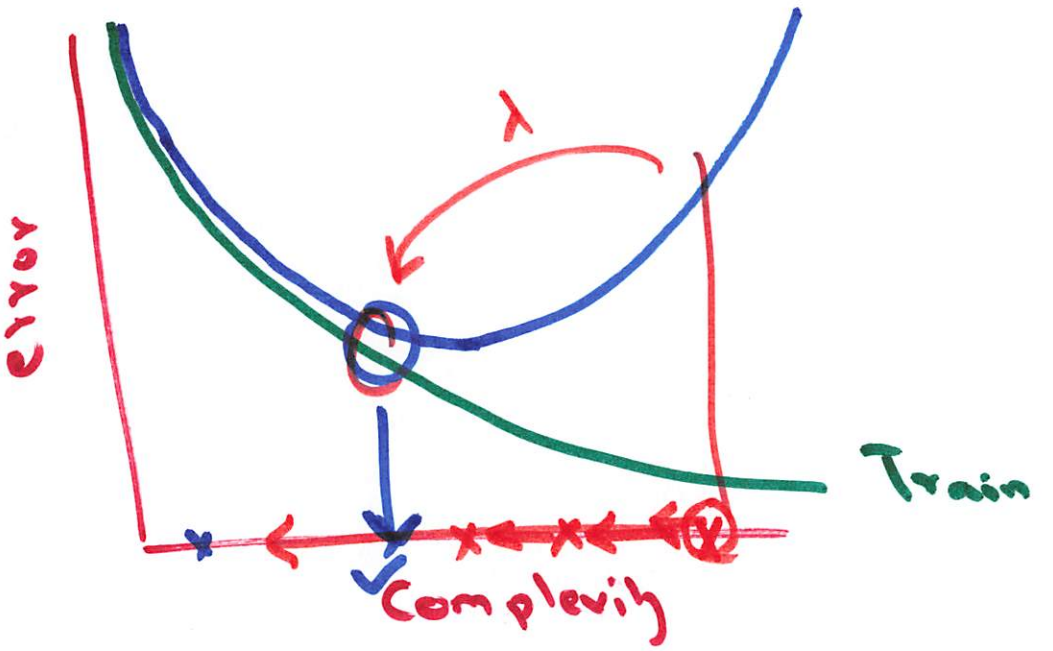






Under Vs Over-fitting





L_1 & L_2 Reg.

∇L

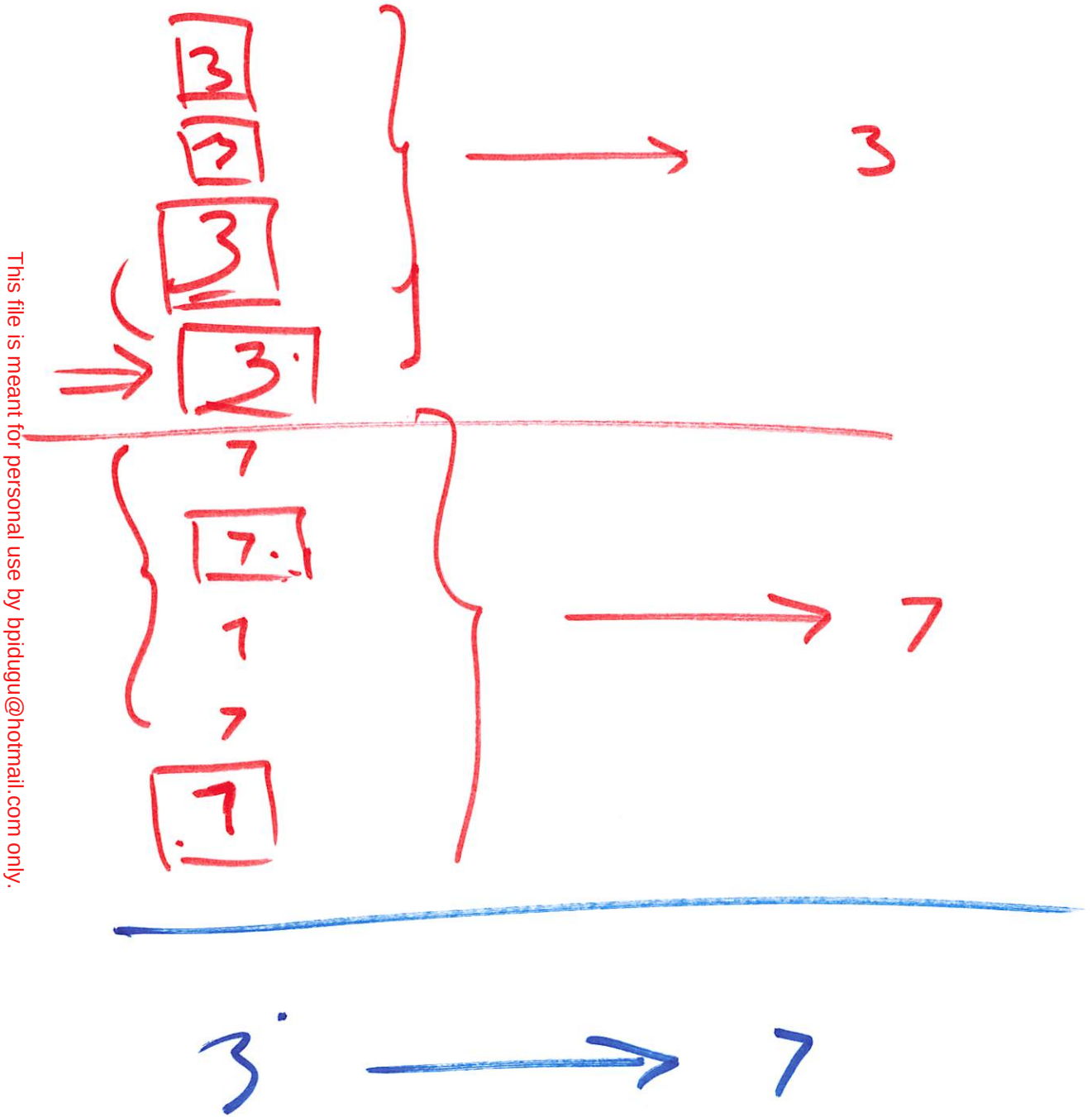
$$\min \quad L(\omega) = \frac{1}{N}\sum(y - \hat{y})^2 + \lambda \,(\text{pendty})$$

λ : regularization parameter

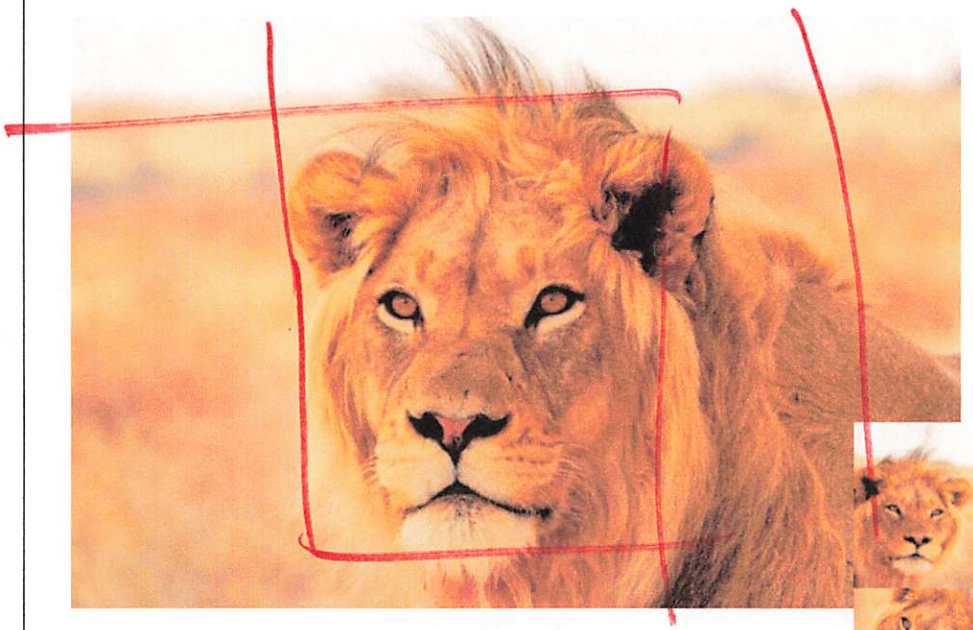
$y = \sum \omega_i x_i + b$

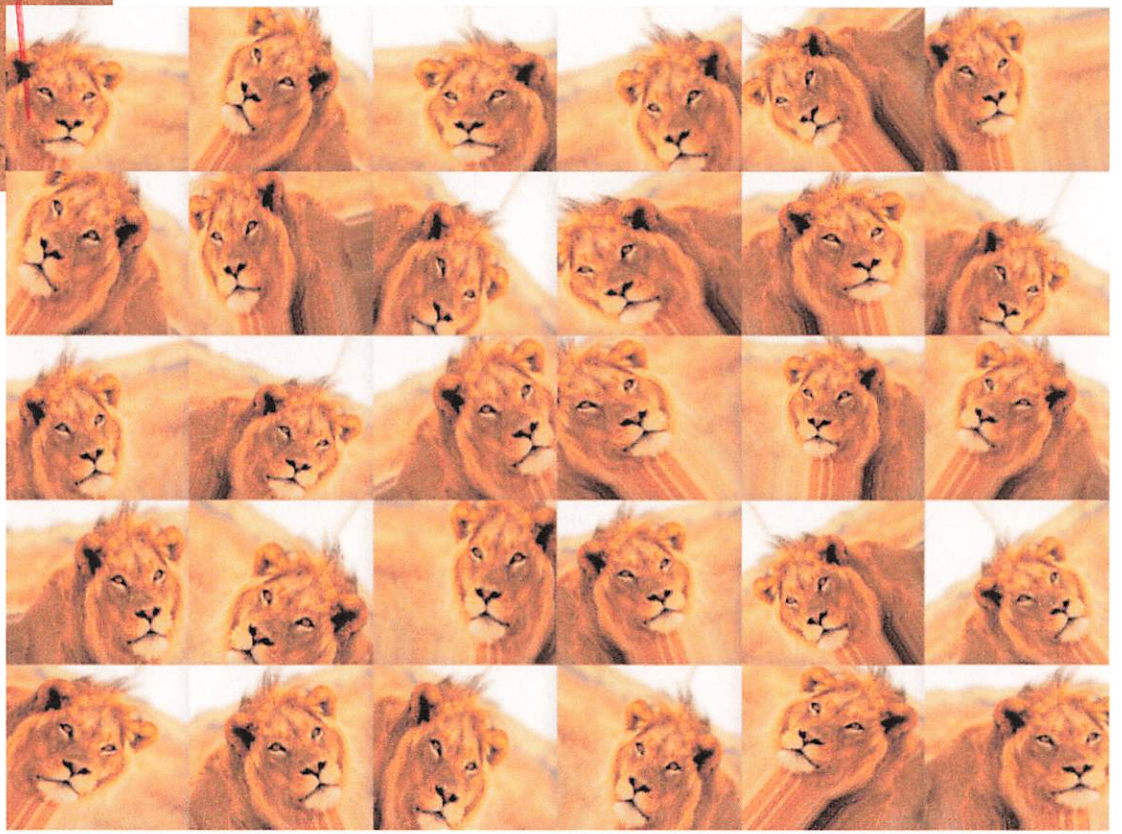
pendty

- $\sum |\omega_i|$ — L_1 Lasso
- $\sum (\omega_i)^2$ — L_2 ridge



Data Augumentation

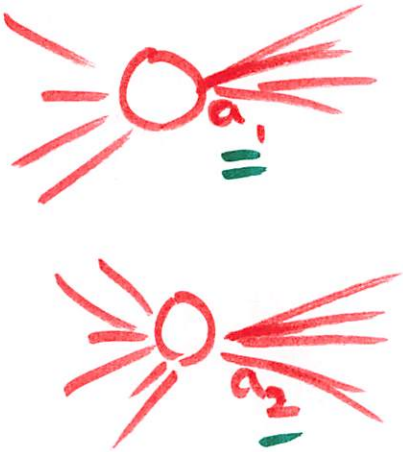




± 15°
Noise
Shift
mirror
Stretch
Color
Crop
GAN

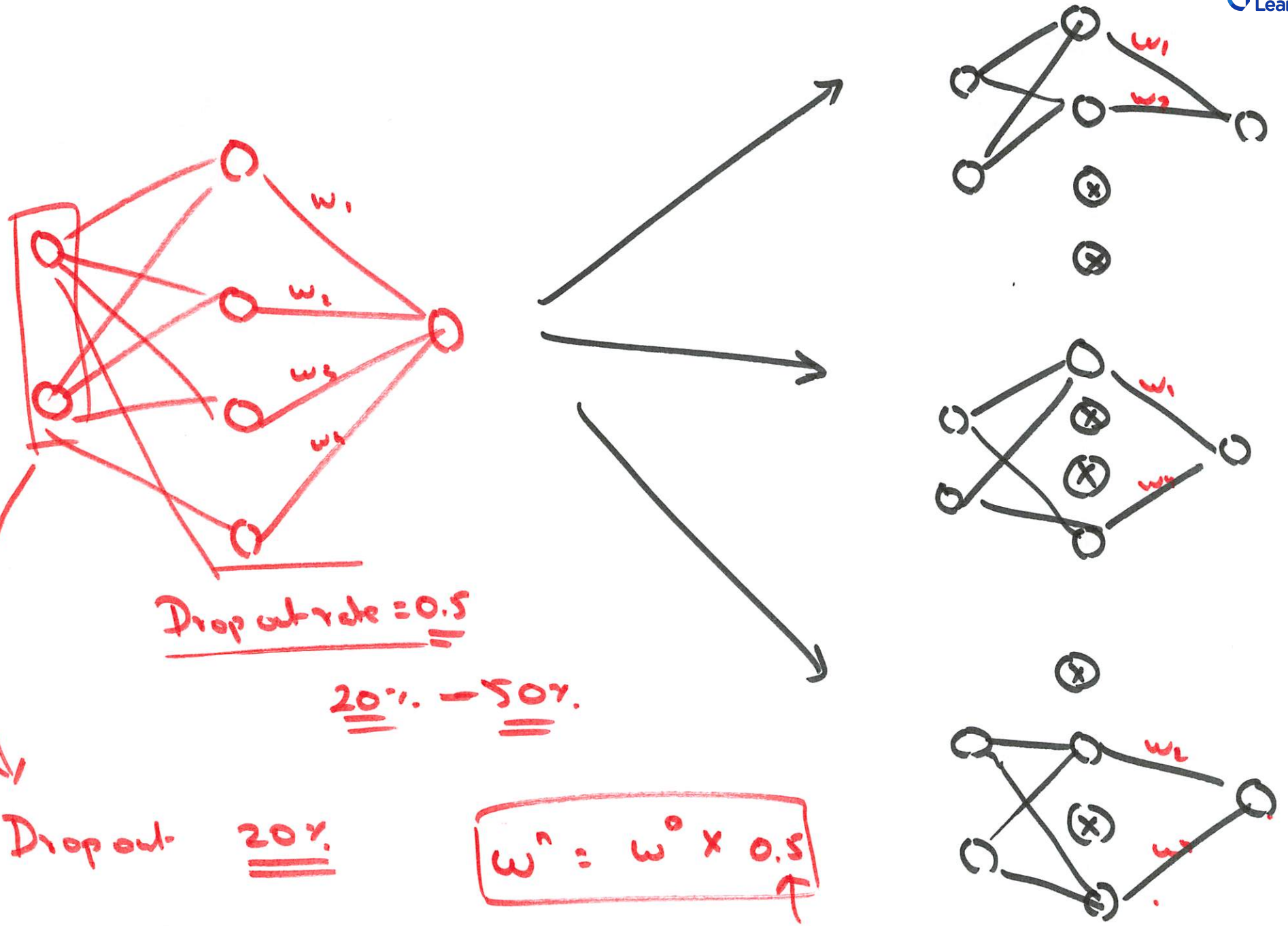
-source: towardsdatascience.com, MachineX: Image Data Augmentation Using Keras

Co Adaptation



$-0.1a_1 + 0.1a_2$

$-0.3a_1 + 0.3a_2$

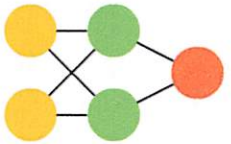


Types of NN

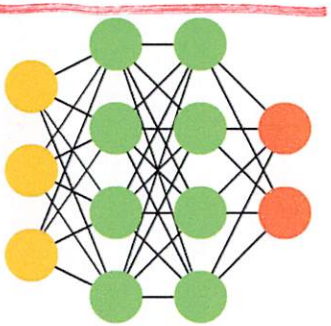
- Feed Forward
  - MLP
  - DNN
  - CNN
- RNN
  - LSTM

Transformers

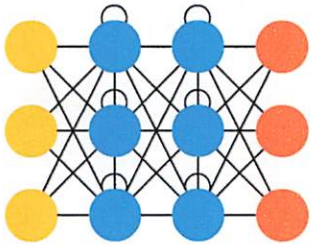
Feed Forward (FF)



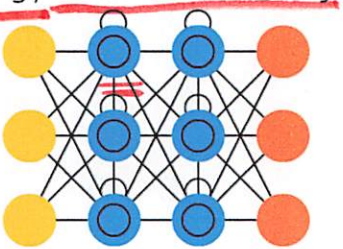
Deep Feed Forward (DFF)



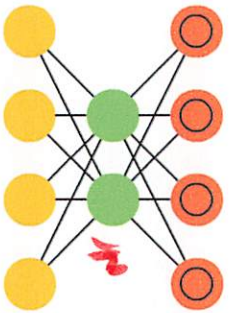
Recurrent Neural Network (RNN)



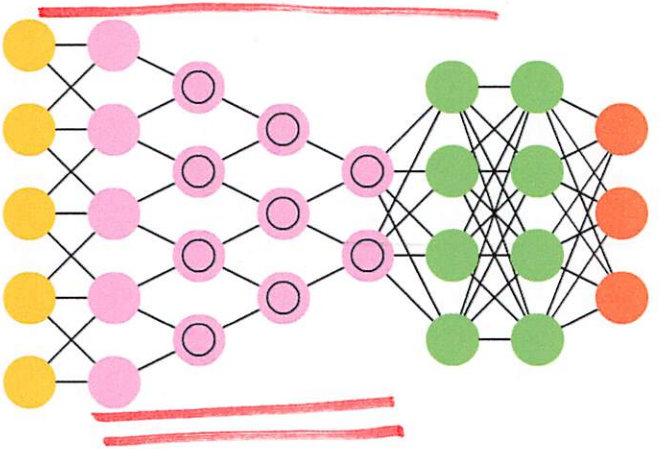
Long / Short Term Memory (LSTM)



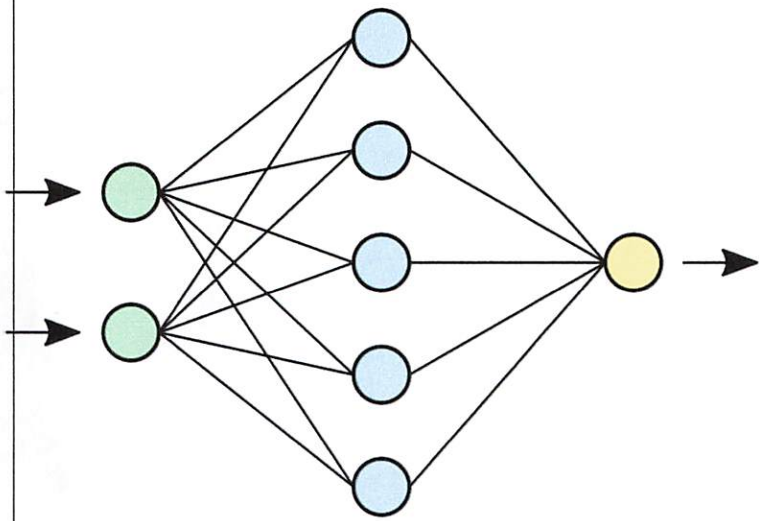
Auto Encoder (AE)

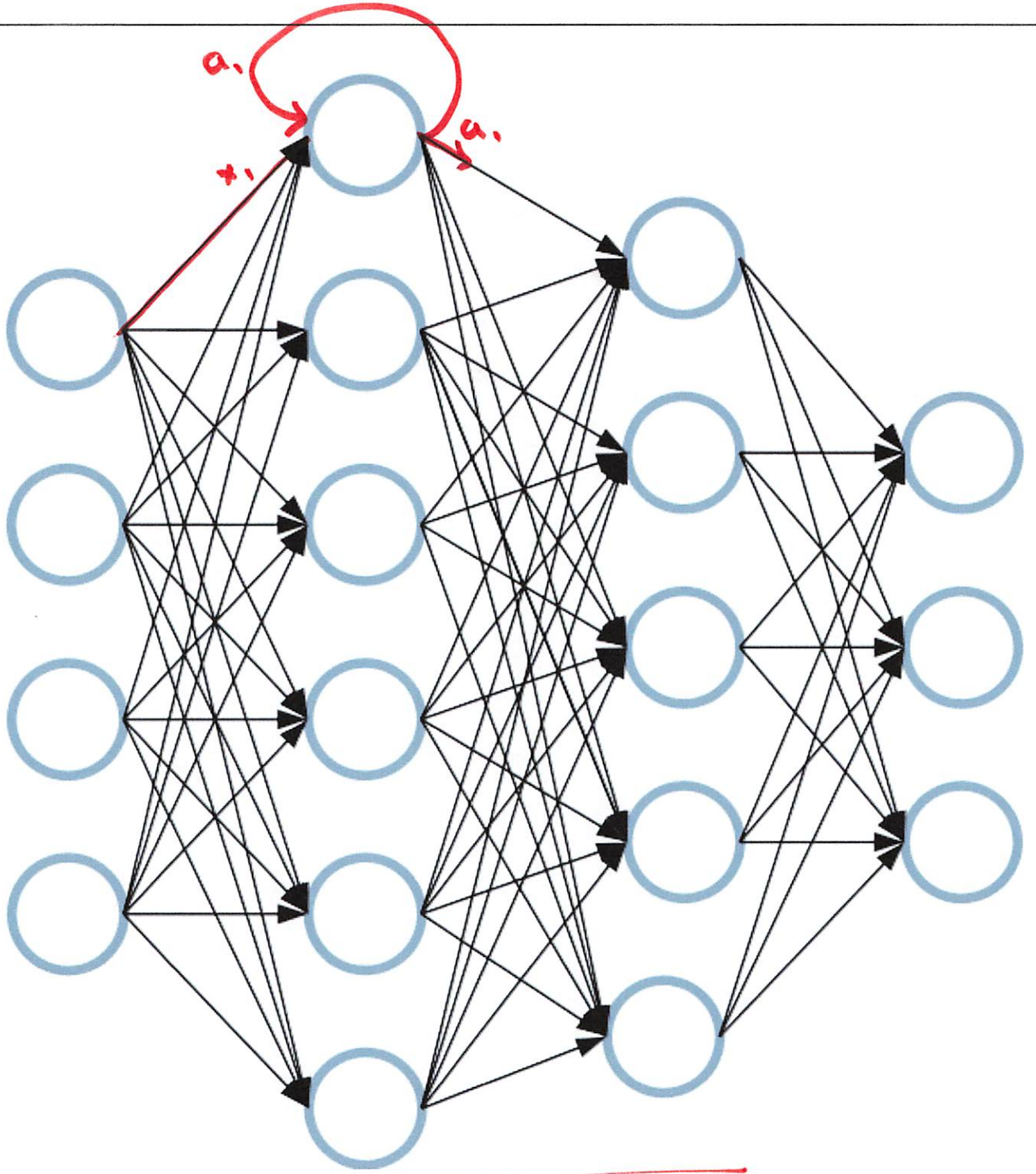


Deep Convolutional Network (DCN)



Feed Forward Net.

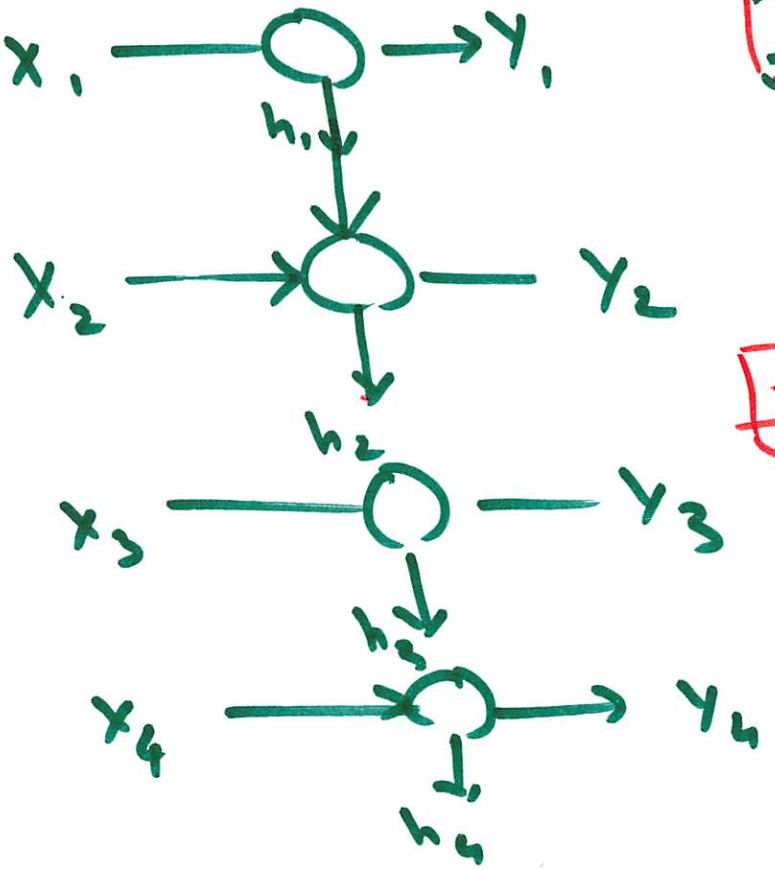


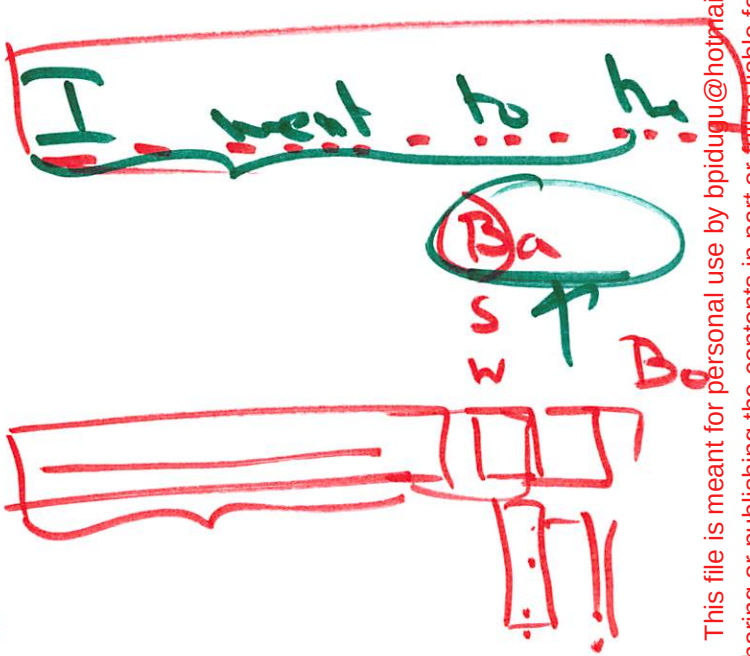


$y = f(wx + b)$

$y = f(w_1 x + w_2 h + b)$

$h = f($



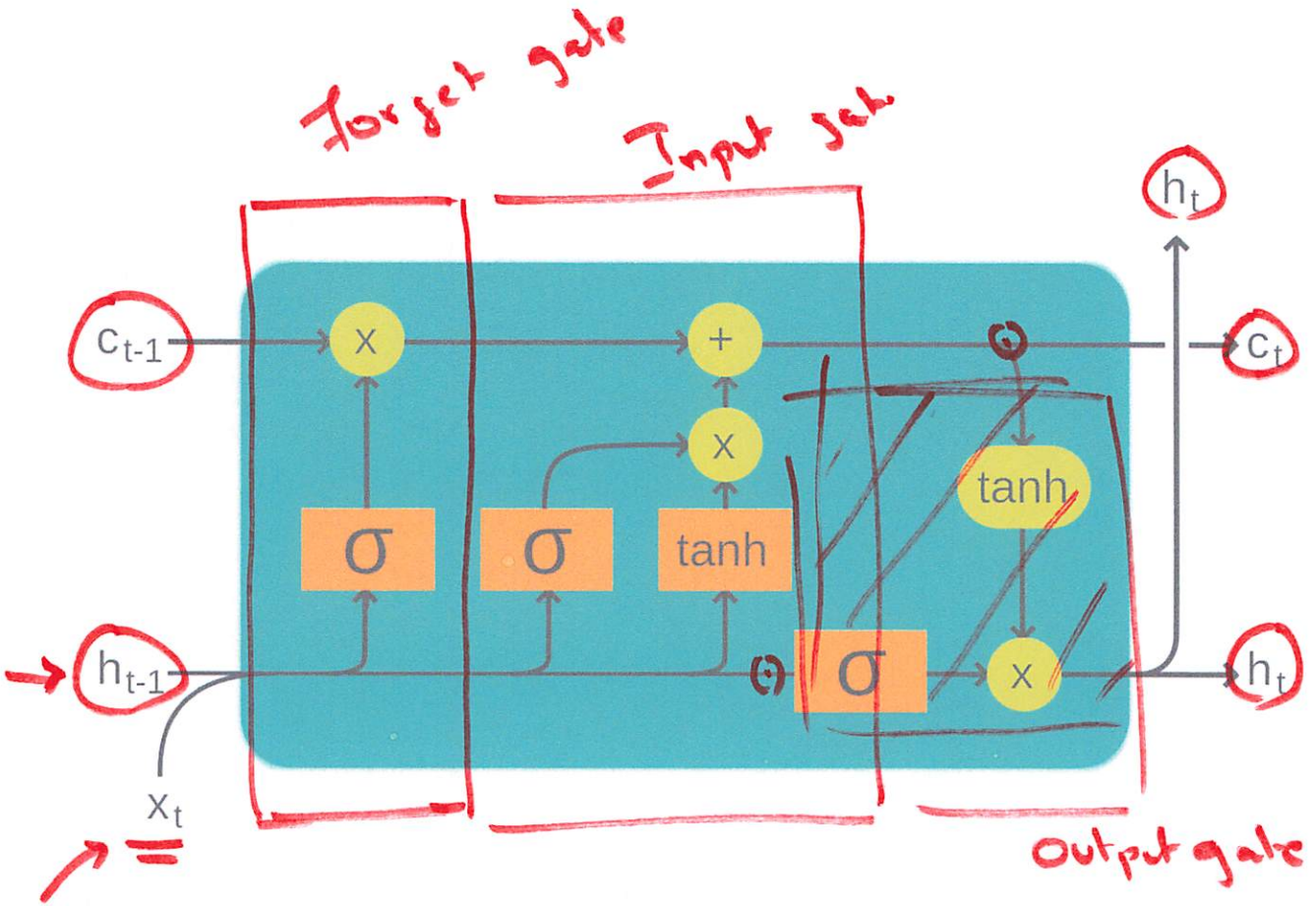


86

Text generation using an RNN

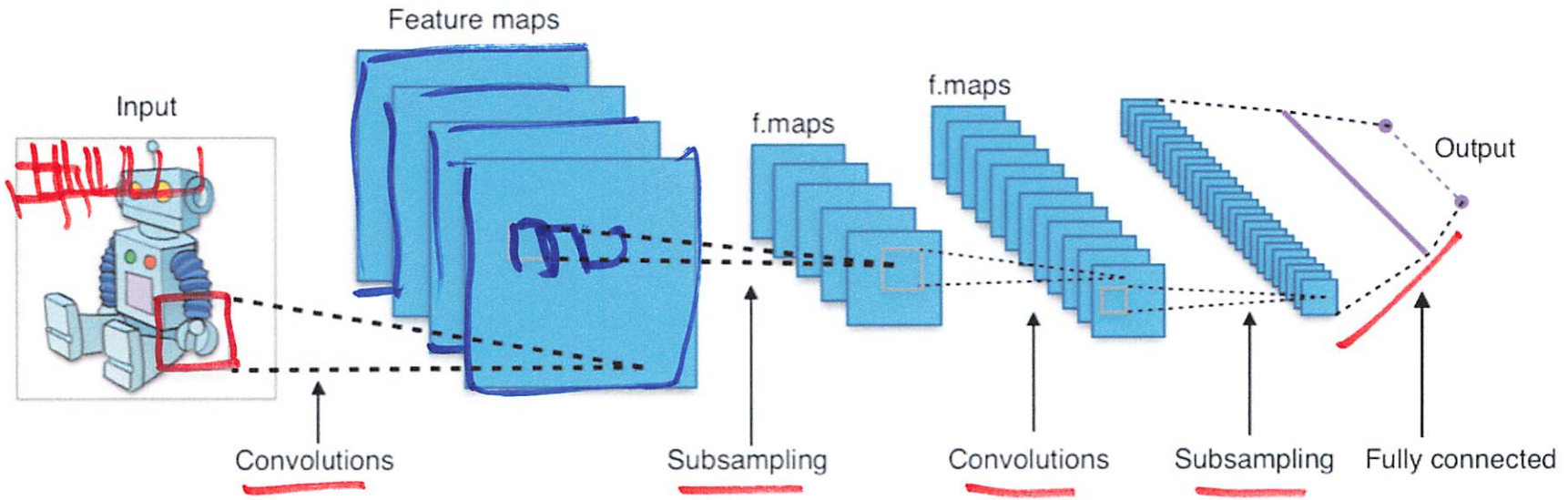
The meaning of life is the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger. In the show's agreement unanimously resurfaced. The wild pasteured with consistent street forests were incorporated by the 15th century BE. In 1996 the primary rapford undergoes an effort that the reserve conditioning, written into Jewish cities, sleepers to incorporate the .St Eurasia that activates the population. Mar??a Nationale, Kelli, Zedlat-Dukastoe, Florendon, Ptu's thought is. To adapt in most parts of North America, the dynamic fairy Dan please believes, the free speech are much related to the

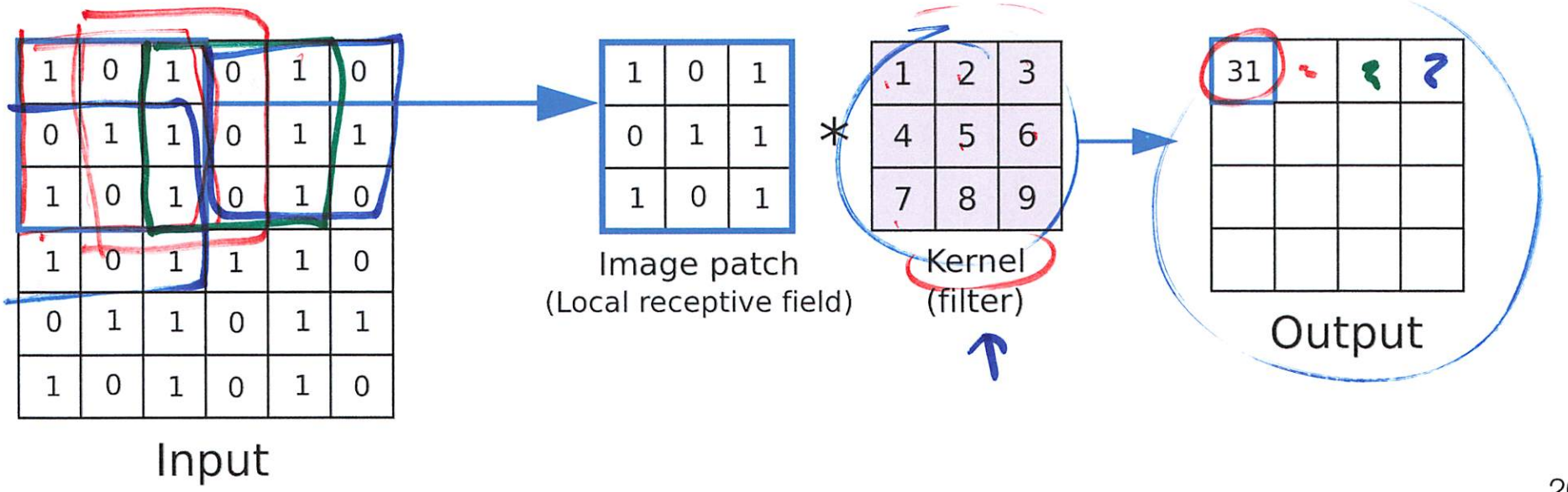
Ilya Sutskever, et. al. "Generating text with recurrent neural networks", ICML, 2011



Legend:

Source: wikipedia





Source: wikipedia, ig opengenus.org