2025-07-29 - Compare and Contrast among activation functions, loss functions, optimizers and regularization for choosing the appropriate method for the given application.

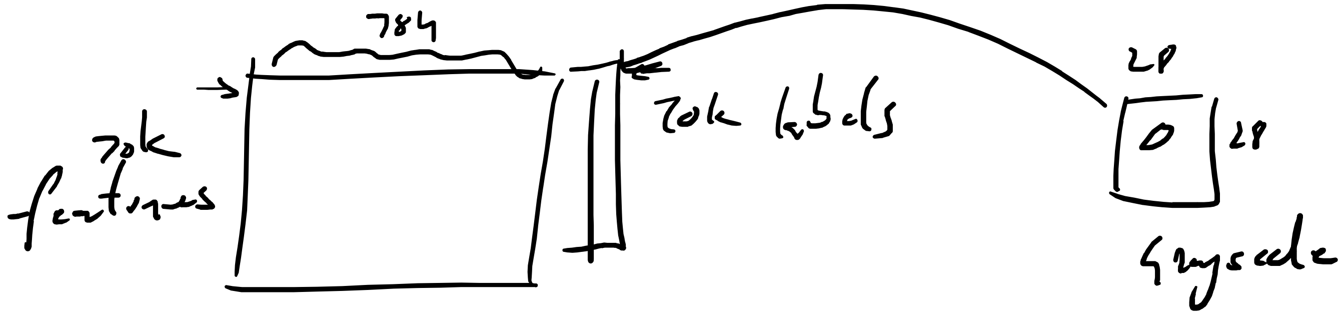
29 July 2025 11:34

=> Digit Classification (MNIST) :

0 - 9 Digits (Handwritten)

70000 × 28×28

=> Grayscale : [0 - 255)



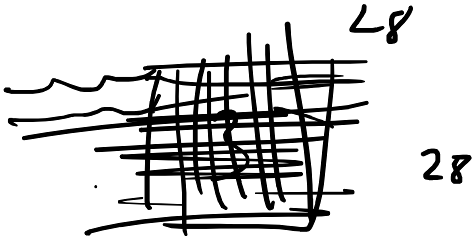
28×28 = 784

(70000 × 784)

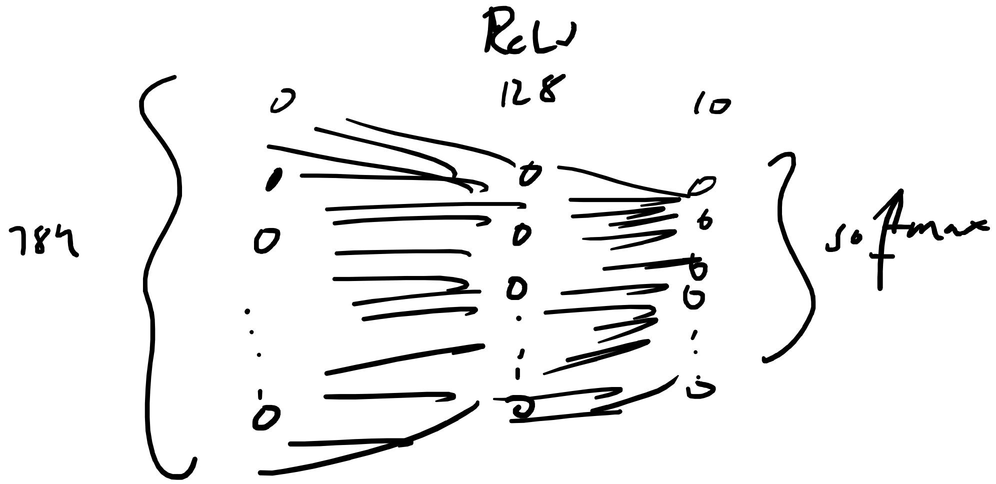
Train & Test

60k

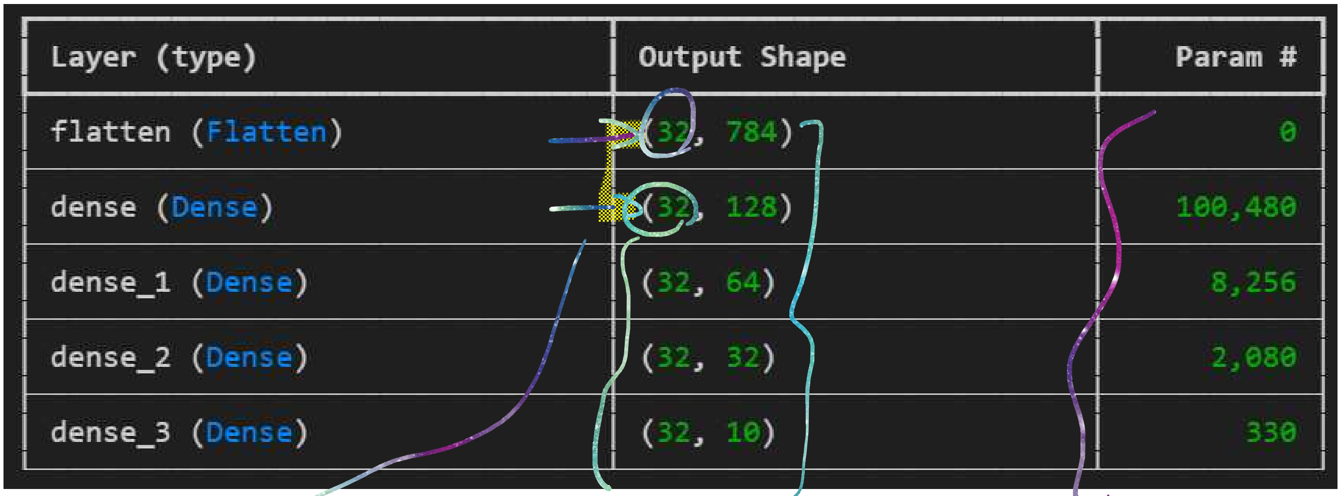
10k



28×28 ≈ 784



Parametric Calculation:



Layer (type)	Output Shape	Param #
flatten (Flatten)	(32, 784)	0
dense (Dense)	(32, 128)	100,480
dense_1 (Dense)	(32, 64)	8,256
dense_2 (Dense)	(32, 32)	2,080
dense_3 (Dense)	(32, 10)	330

=> Batch Size Features Parameters

$$\text{Parameter} = (\text{Previous Features} \times \text{Current Features}) + \text{Curr Features}$$

$$= (784 \times 128) + 128 = 100{,}480$$

=> Overfitting: Network trains very closely to the training data.

Overfitting

Add more training data

Regularization
Early Stopping
Batch Normalization

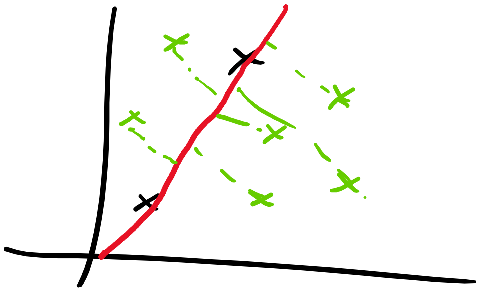
=> Regularization : Penalty

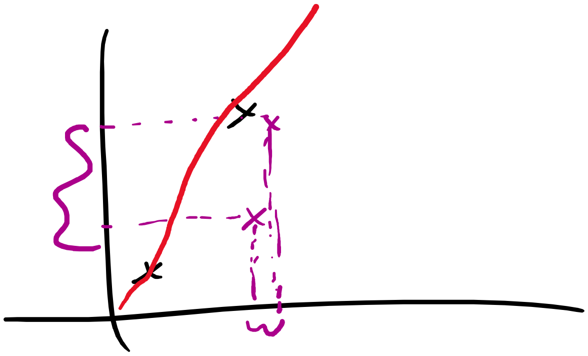
L1 - Lasso

L2 - Ridge ✓

L1+L2 - Elastic

=> Ridge





$$\Rightarrow \quad \hat{y} = w \cdot x + b$$

$$\rightarrow L = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \sim \text{Error}$$

Cost Fn

$$C = \frac{1}{N} \sum_{i=1}^{N} L(y_i - \hat{y}_i) + \frac{\lambda}{2N} \sum_{i=1}^{N} \|w_i\|^2$$

Hyperparameter Penalty

$$\frac{\lambda}{2N}\left[w_1^2 + w_2^2 + \cdots\cdots w_N^2\right]$$