* Expected Test Error = Bias^2 + Variance + Noise

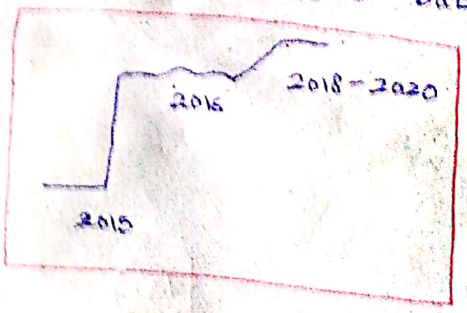
* Optimizer → Minimize the function of loss / error term

* We start with a random point on the function and move in the negative direction of the gradient of the function to reach the Local / global minima

(*) Revolutions :- 2012 → Alex Net : Image Net competition

1985 ← ANN.

2014, 2015, 2016 } CNNs / Batch Norm, OD, IS ← Object Detection, Image Segment.



self-attention : Attending to the same sequence

Masking : To prevent undesired masking

$$\text{Attention} = \text{Softmax}\left(\frac{q \cdot k^T}{\sqrt{d}}\right) \cdot v$$

Input Pipeline

1) Tokenization → Breaks a string into a sequence of tokens

2) Lookup → Assigns integer ID to each token using a vocab

3) (Input) embedding → Assigns a (learnable) vector to each integer ID in the vocab

4) Positional encoding → Encodes knowledge about the position in the sequence

Batch Gradient Descent → Fast

Stochastic Gradient Descent → slowest, Learning schedule

Mini-Batch Gradient Descent → Fastest