Improving Deep Neural Networks

Video 1: Problems in Deep Neural Networks

Analytics
Vidhya

In Air

Challenges in Neural Networks

Problems with Large Neural Networks

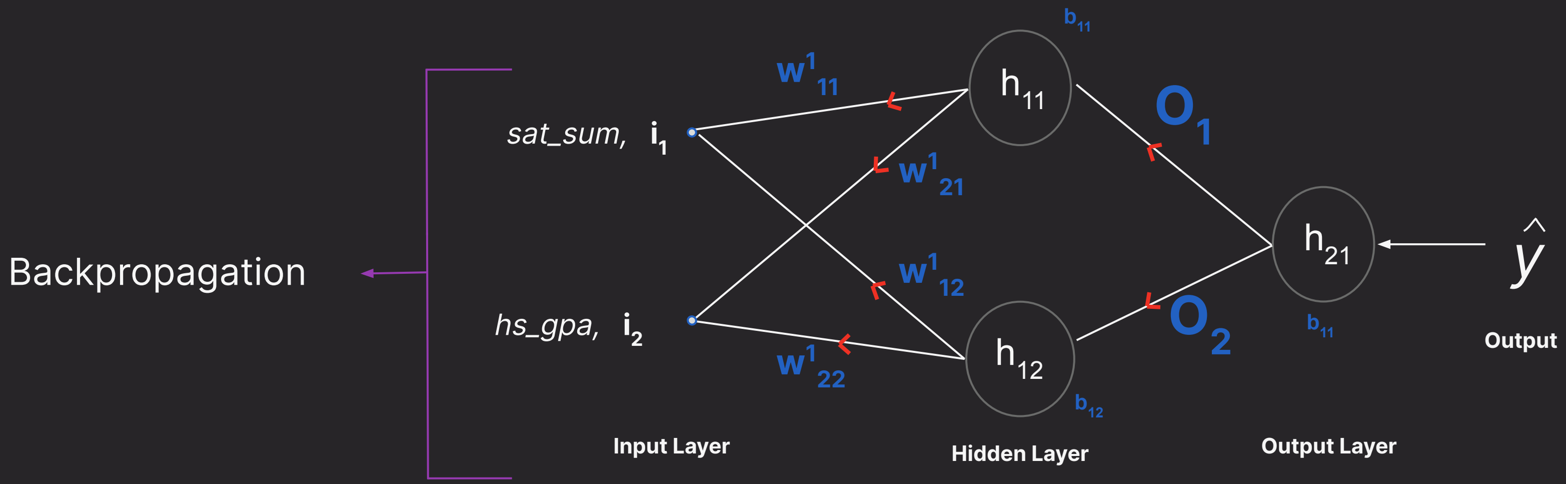
- Vanishing and Exploding Gradients
- Overfitting

Challenges in Neural Networks

Problems with Large Neural Networks

- **Vanishing and Exploding Gradients**
- Overfitting

Challenges in Neural Networks



How does too small or too large weight affect neural network learning ?

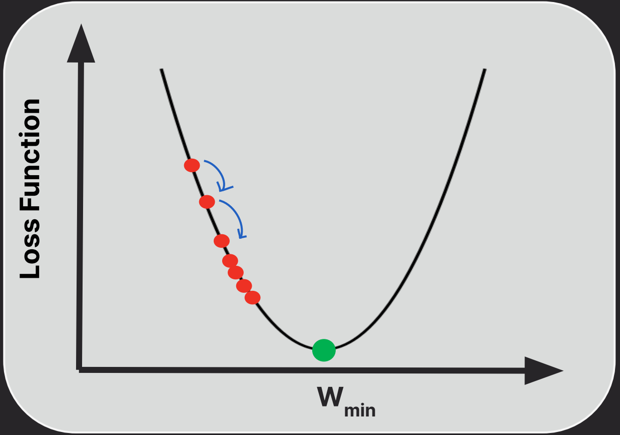
Vanishing Gradients

- Update to the weights is small

$$w_{new} = w_{old} - eeta * dL/dw$$

Becomes closer to 0

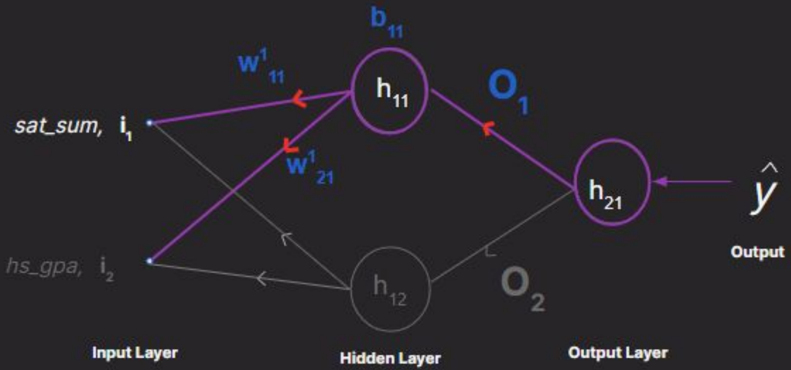
$$w_{new} = w_{old}$$



Small Learning Rate

Why do deep neural networks face vanishing gradient problem ?

Chain Rule of Differentiation

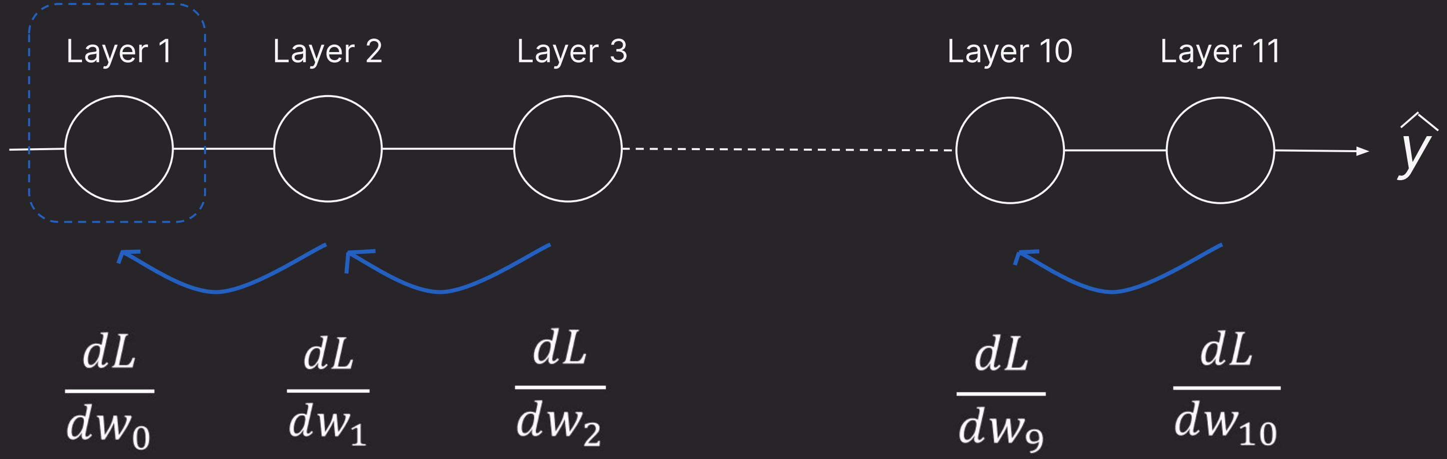


$$\frac{dL}{dw_{11}^{1}} = \frac{dL}{dO_1} * \frac{dO_1}{dw_{11}^{1}}$$

$$\frac{dL}{dw_{21}^{1}} = \frac{dL}{dO_1} * \frac{dO_1}{dw_{21}^{1}}$$

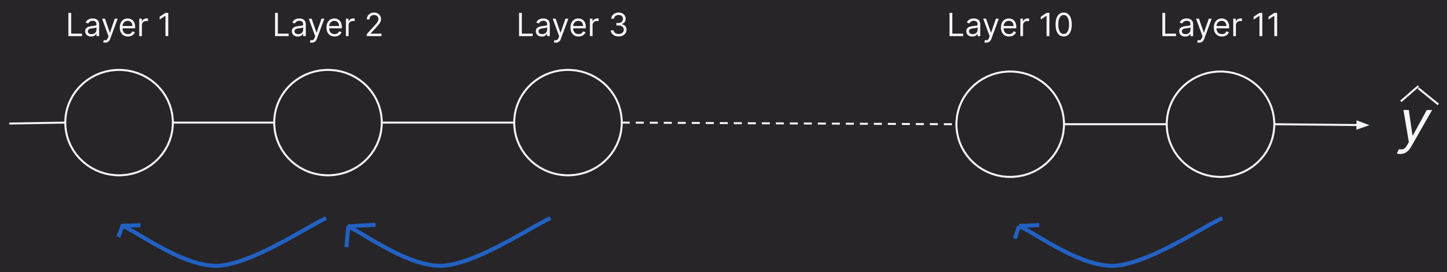
$$\frac{dL}{db_{11}} = \frac{dL}{dO_1} * \frac{dO_1}{db_{11}}$$

Chain Rule of Differentiation



- Could be represented as multiplication of all previous Gradients $\left(\frac{dL}{dw_n}\right)^{10}$

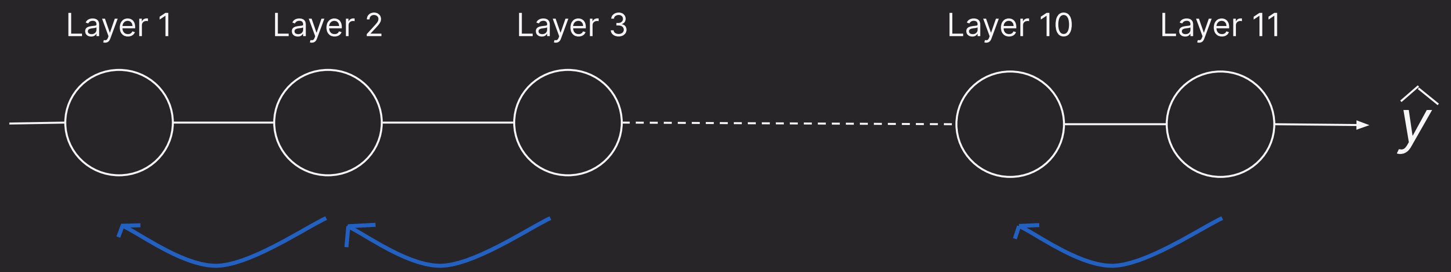
Vanishing Gradient



- **Assumption:** $dw_n = 0.1$ → $dw_0 * 0.1^{10}$ → A very small value ; close to 0

- Could be represented as multiplication of all previous Gradients $\left(\frac{dL}{dw_n}\right)^{10}$

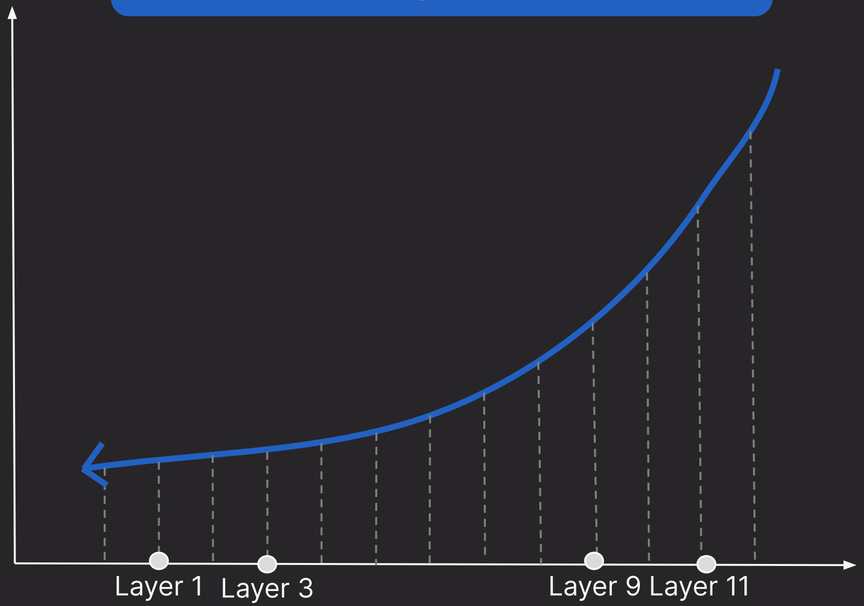
Exploding Gradient



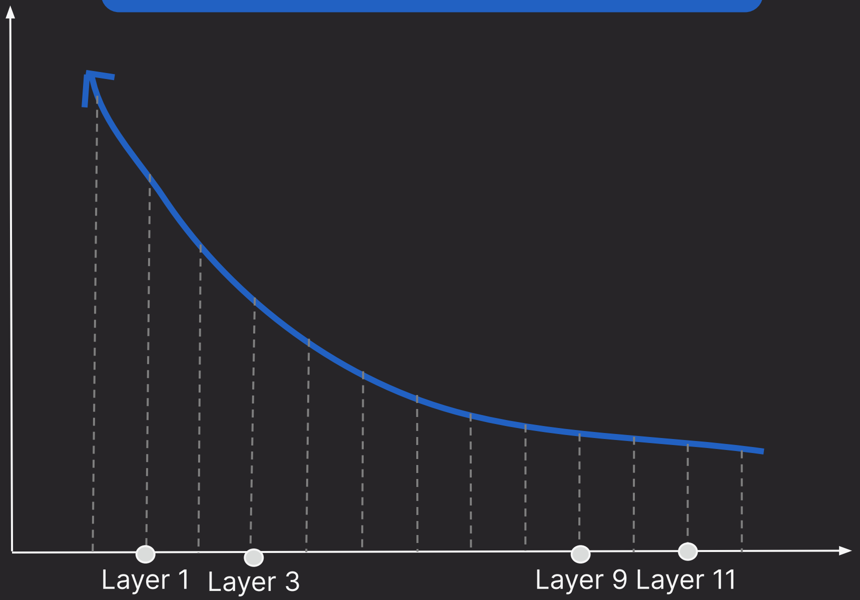
Assumption: $dw_n = 2$ → $dw_0 * 2^{10}$ → Extremely large value for weight update → Parameters can overflow, giving NaN values

Could be represented as multiplication of all previous Gradients $\left(\frac{dL}{dw_n}\right)^{10}$

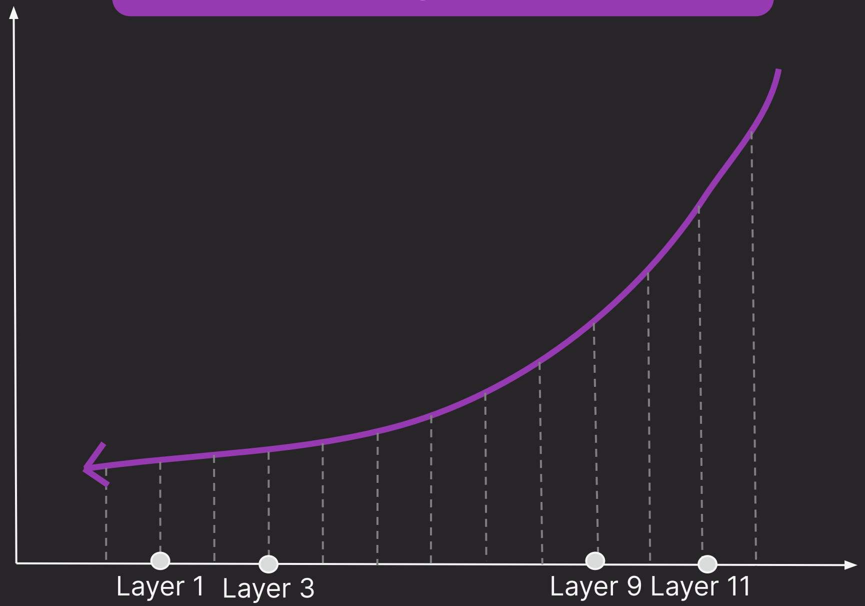
Vanishing Gradient



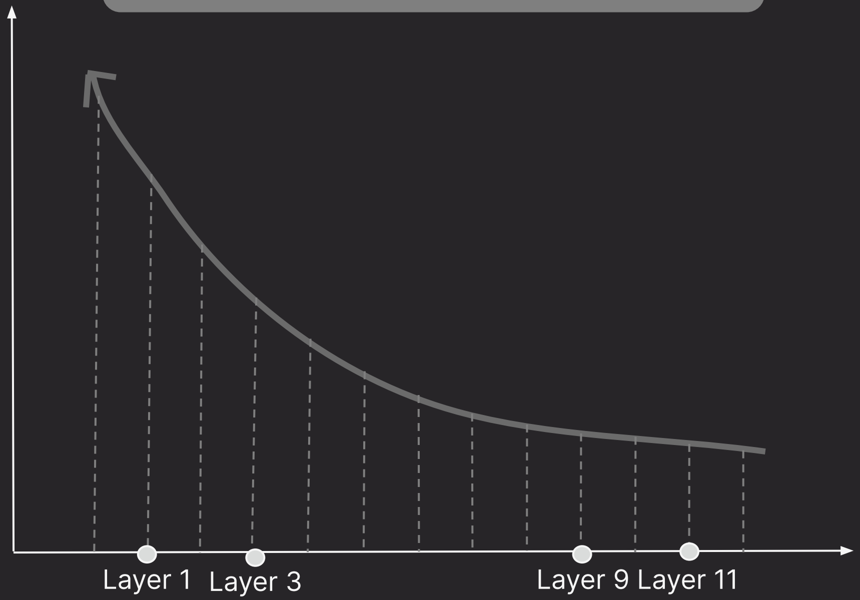
Exploding Gradient



Vanishing Gradient



Exploding Gradient



How to identify which gradient the model is suffering from ?

Exploding Gradient	Vanishing Gradient
There is an exponential growth in the model parameters of the lower or initial layers.	The parameters of the higher layers change significantly whereas the parameters of lower or initial layers would not change much or not at all.
The model weights may become NaN during training	The model weights may become close to 0 during training.
During training the model may abruptly increase loss values by a large amount.	During training the model learns very slowly and the training stagnates at a very early stage.