⑤

Batch Normalization

Batch normalization is an algorithmic method which makes the training of deep neural network faster and more stable.

It consists of normalization activation vectors from hidden layers using the mean and variance of the current batch. This normalization step is applied right before (or right after) the non-linear function.

$$\text{normalize} = \text{mean} = \mu = 0$$
$$\text{sd} = \sigma = 1$$

Covariate Shift – It refers to the situation where the distribution of input features changes b/w the training and testing phase. The relationship b/w the input features and the target variable remains same but the distribution of the input features itself changes.

Internal Co-variate Shift – Change in the distribution of network activation due to the change in network parameters during training. So the distribution changes when it is passed thro an entire network, at the higher level, the distribution will be different. normalization ensures that the distribution will be normal at each hidden layer.

How does batch normalization works ?

Step 1 :- During training, bn. operates on mini batches of input data for each layer in the neural network, it computes the mean and variance of the activations along each feature dimensions across the mini batch.

Step 2 :- normalizes the action function by subtracting the mean and dividing by the standard deviation.

It ensures that the mean is zero and the standard deviation is one.

Applying normalization before the activation function is much more popular than the other way.

$$Z_{11} = w_1 x_1 + w_2 x_2 + b.$$

$$g(Z_{11}) = \boxed{a_{11}}$$

$$Z_{11} = Z_{11}^{N} \longrightarrow g(Z_{11}^{N}) = a_{11}$$

Norm $= \boxed{\dfrac{Z_{11} - \mu}{\sigma}}$

$$Z_{11}^{i} = \frac{Z_{11}^{i} - \mu_B}{\sigma_B + \epsilon}$$

→ Error term is added to avoid being zero.

Step 3 :-

After normalization, batch normalization introduces learnable parameters, scale (γ) and shift (β) for each feature dimension

$$\boxed{Z_{11}^{BN} = \gamma Z_{11}^{N} + \beta}$$

initial value $\gamma = 1$ $\beta = 0$

$$g(Z_{11}^{BN}) = a_{11}$$

These params allow the model to learn optimal scaling and shifting of the normalized activation, enabling it to adapt to the data distribution.

→ gradients of the loss w.r.t scale & shift params are computed and used to update these param via optimization algorithm like stochastic gradient descent.

→ when making prediction on the new data, batch norm calculates the mean & variance using statistics computed over the entire training dataset.

The scale & shift params learned during training are then used to normalize the activations of the inference data.

Advantages :- Stabilizes training.
→ Enables higher learning rates.
→ Acts as regularization.
→ Improves generalization.