03/10/23

Deep Learning

- Recap: CNN, VGG
- Challenges in training deep networks
- Introduction to batch normalization

- The simplest variant of the VGG model has 138M params.
- Modern DL models have in excess of 500M params !
- This leads to challenges in training such models.
- We will discuss popular methods to deal with such challenges.
- Batch - Normalization
  - Recall the SGD method
  - It helped when in the # training samples is large
  - However, a drawback is noisy estimates of gradient
  - We also have to deal with internal covariate shifts due to the mini batches
- Batch-norm addresses these problems by "whitening" the inputs to a layer in the n/w.
- Let $\mathcal{B} = \{ \underline{x}^{(1)}, \ldots, \underline{x}^{(m)} \}$ be the mini batch of samples in SGD : with $\underline{x} \in \mathbb{R}^d$

  For each dim $k = 1 \ldots d$ do the following

$$\mu_k = \frac{1}{m} \sum_{i=1}^{m} x_k^{(i)}$$

$$\sigma_k^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_k^{(i)} - \mu_k)^2$$

  For each sample $i = 1 \ldots m$ do the following

$$\hat{x}_k^{(i)} = \frac{x_k^{(i)} - \mu_k}{\sqrt{\sigma_k^2 + \epsilon}} \quad ; \ \epsilon \text{ is a small constant}$$

$$y_k^{(i)} = \gamma_k \cdot \hat{x}_k^{(i)} + \alpha_k$$

γ_k, α_k are learnable params

α_k → bias

Scaling & used to ensure the expressivity of the model

- Note that the effective input of minibatch samples to the layer is now

$$\mathcal{B}_{BN} = \{ y^{(1)}, \ldots y^{(m)} \}$$

Also note that the dim of y is the same as x

- During training, the sample mean and variance of the entire dataset is computed and saved.
- During inference, an input sample is effectively whitened using the "global" mean and variance.