L09: 16.05.2022

5.4.2 Batch normalization

Input normalization : done once for the dataset

Hidden layers : data distribution changes due to the update of $\underline{\theta}$

- over layers
- over time during training

$\Big\}$ internal covariate shift

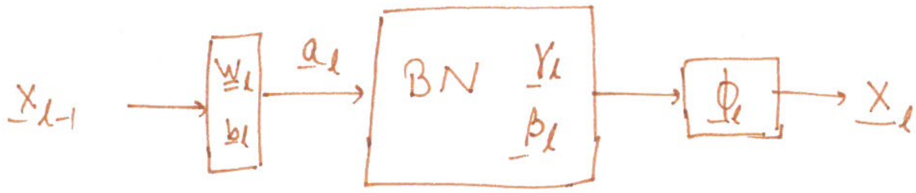
→ slow down training if nothing is done

Soln:

Batch normalization (BN):

like Input normalization, but

- for hidden layers (we have a choice on which layer to apply)
- for each mini batch → stabilize training



BN layer

Slide 520
521

Q (Ilias) If we first do a zero mean unit variance scaling and then again boost the features then we still obtain the dynamic range with a different μ and σ. So what is the use of using an affine transform of the activation o/p?

Q Mathematically, we can write this whole term as one term in place of $\underline{\underline{W}}$ and $\underline{b}$, so how does it make a difference?

5.5 Parameter initialization

SGD does a local search, so we need a good starting point. The sol$\underline{n}$ depends on the initial value $\underline{\theta}^0 = \ell_0$

Different parameter ~~ite~~ initializations :

Slide 5-24

2. Random initialization:

Only done for $\underline{\underline{W}}_l$. Bias are set to 0. eg

- normal distribution $N(\,)$

$$[\underline{\underline{W}}_l]_{ij} \sim \text{iid } \sigma . N(0,1) = N(0, \sigma^2)$$

to control the data dynamic range.

This leads to assymetry.

- uniform distribution $U(\,)$

$$[\underline{\underline{W}}_l]_{ij} \sim \text{iid } \sigma\, U(-1,1) = U(-\sigma, \sigma)$$

$\sigma =$ const $\forall i, j, l \longrightarrow$ not optimum

3. He initialization (3)

as 2), but $\sigma_l \sim \frac{1}{\sqrt{M_{l-1}}} \longrightarrow$ constant activation flow after initialization

layer l after initialization:

$$\underline{a}_l = \underline{\underline{W}}_l \cdot \underline{x}_{l-1} + \underline{b}_l \xrightarrow{b_l = 0} \underline{\underline{W}}_l \cdot \underline{x}_{l-1}$$

or

$$\underset{M_l \times 1}{\underline{a}} = \underset{M_l \times M_{l-1}}{\underline{\underline{W}}} \cdot \underset{M_{l-1} \times 1}{\underline{X}}$$

$$a_i = \sum_{j=1}^{M_{l-1}} W_{ij} x_j$$

Assumption:

- X_j iid – zero mean, variance σ_x^2
- W_{ij} iid " " " σ_w^2
- X_j and W_{ij} independent.

$\Rightarrow$ • $E(a_i) = \sum_j \underbrace{E(W_{ij})}_{0} \underbrace{E(x_j)}_{0} = 0$

• $Var(a_i) = E(a_i^2) - \underbrace{(E(a_i))^2}_{0}$

$$= E\left[\left(\sum W_{ij} X_j \right)^2 \right]$$

$$= E\left[\sum_j \sum_k W_{ij} W_{ik} X_j X_k \right]$$

$$= \sum_j \sum_k E\left[W_{ij} W_{ik} \right] E\left[X_j X_k \right] = \sum_{j=1}^{M_{l-1}} \sigma_w^2 \sigma_x^2 = M_{l-1} \sigma_w^2 \sigma_x^2$$

Therefore constant activation flow in fwd pass:

$$\mathrm{Var}(a_i) = \text{const} \quad \forall\, i, l$$

$$\Rightarrow \sigma_{w,l} \approx \frac{1}{\sqrt{M_{l-1}}} \qquad M_{l-1} : \text{fan in}$$

4. Glorot initialization

Constant gradient flow in backward pass

$$\left\| \frac{\partial L}{\partial \underline{a}_l} \right\| = \text{const} \quad \forall\, l$$

$$\Rightarrow \sigma_{w,l} \sim \frac{1}{\sqrt{M_l}} \qquad M_l : \text{fan-out}$$

Compromise : $\sigma_{w,l} \sim \dfrac{1}{\sqrt{M_{l-1} + M_l}}$

5. 6 Improved model

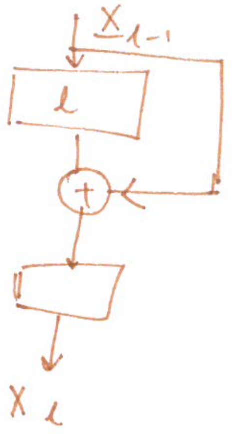
Model - architecture of NN

A. A better activation function $\phi_l(x)$:

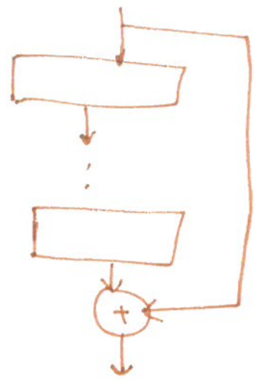
- ReLU instead of sigmoid.
- leaky ReLU instead of ReLU as we have small gradient for -ve values.

B. Skip-connections, shortcuts, residual n/w

over one layer



over multiple layer



- forward pass:

  Combine low level features of shallow layer with high level features of deep layer

  see ch 7, ch 10

- backward pass

$$\underline{x}_l = \phi_l\left(\underline{\underline{W}}_l \cdot \underline{x}_{l-1} + \underline{b}_l\right) + \underline{x}_{l-1}$$

$$\frac{\partial x_l}{\partial \underline{x}_{l-1}} = \underbrace{\frac{\partial \phi_l}{\partial x_{l-1}}} + \underline{\underline{I}}$$

This can have a vanishing gradient term, but $\underline{\underline{I}}$ guarentees that propagation will always happen. Hence we need shortcuts.

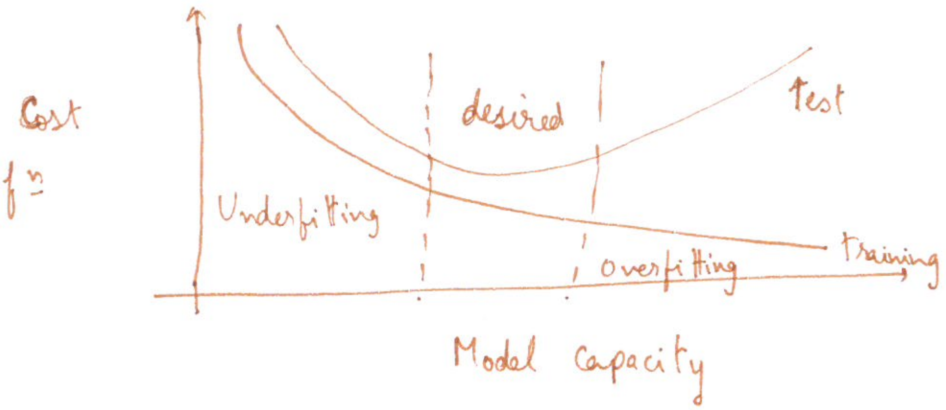
(c) Many other architecture improvements see ch 10.

6 Overfitting and regularization

6.1 Model capacity, underfitting and overfitting

Slide 6-1

Slide 6-2



Slide 6-3