Hyperparameters tuning 3 week 01.01.2019

Them first

α

$\beta \sim 0.9$

#hidden units

mini-batch size

#layers

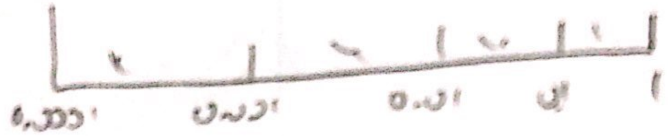
learning rate decay

$\beta_1, \beta_2, \varepsilon$

$\hookrightarrow 0.9$ $\hookrightarrow 0.999$ $\hookrightarrow 10^{-8}$

Appropriate scale for hyperparameters

$\alpha = 0.0001 \ldots 1$



$r = -4 * \text{np.random.rand()} \leftarrow r \in [-4, 0]$

$\alpha = 10^r \leftarrow 10^{-4} \ldots 10^0$

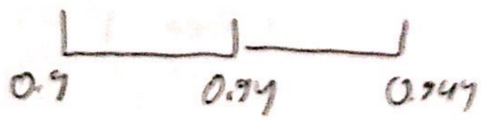
10^a 10^b $a = \log_{10} 0.0001$

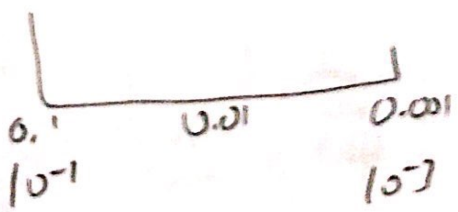
$b = \log_{10} 1$

Hyperparameters for exponentially weighted averages

$\beta = 0.9 \ldots 0.999$

$1 - \beta = 0.1 \ldots 0.001$





$r \in [-3, -1]$

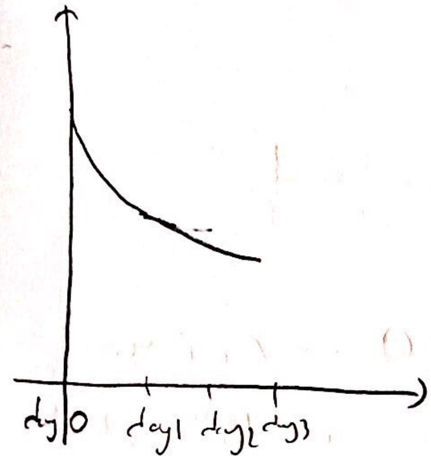
$1 - \beta = 10^r$

$\beta = 1 - 10^r$

Hyperparameters tuning in practice: Pandas vs. Caviar

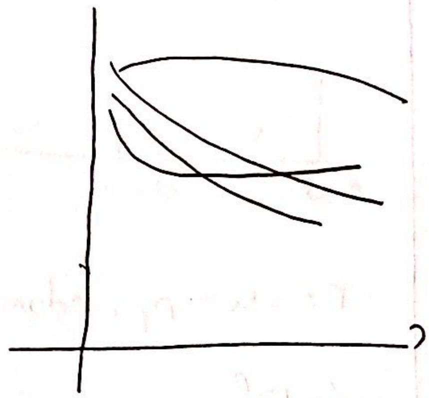
Babysitting one model

- Dek code
- Kaynak c7



Training many models in parallel

• try different model parallel



Normalizing activations in a network (Sergey Ioffe and Christian Szegedy)

Batch Normalization

$$\mu = \frac{1}{m} \sum_i z^{(i)}$$

$$\sigma^2 = \frac{1}{m} \sum_i (z_i - \mu)^2$$

$$z^{(i)}_{norm} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \varepsilon}}$$

$$\tilde{z}^{(i)} = \overbrace{\gamma z^{(i)}_{norm} + \beta}^{\text{learnable parameters of model}}$$

if

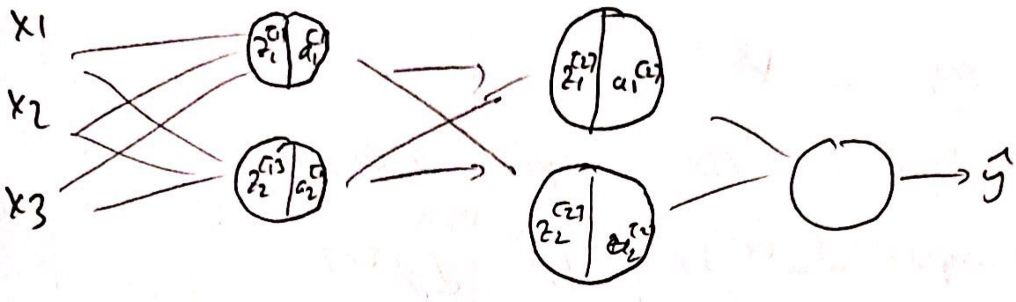
$\gamma = \sqrt{\sigma^2 + \varepsilon}$

$\beta = \mu$

then $\tilde{z}^{(i)} = z^{(i)}$

01.0

Adding Batch Norm to a network



$$X \xrightarrow{W^{[1]}, b^{[1]}} Z^{[1]} \xrightarrow[\text{Batch Norm (BN)}]{\beta^{[1]}, \gamma^{[1]}} \tilde{Z}^{[1]} \to a^{[1]} = g^{[1]}(\tilde{Z}^{[1]}) \xrightarrow{W^{[2]}, b^{[2]}} Z^{[2]} \xrightarrow[BN]{\beta^{[2]}, \gamma^{[2]}} \tilde{Z}^{[2]} \to a^{[2]}$$

Parameters: $W^{[1]}, b^{[1]} \dots W^{[l]}, b^{[l]}$ → Bu betanın weightleri optimize etmek için kullanılan momentum ve adam ile daha yok!!

$\beta^{[1]}, \gamma^{[1]} \dots \beta^{[l]}, \gamma^{[l]}$

$d\beta^{[l]}$ $\quad \beta^{[l]} = \beta^{[l]} - \alpha \, d\beta^{[l]}$

Working with mini-batches

if you using batch norm b parameters eliminate!!

$W^{[l]}, \cancel{b^{[l]}}, \beta^{[l]}, \gamma^{[l]}$

$(n^{[l]},1)$ $(n^{[l]},1)$ $(n^{[l]},1)$

number of hidden units

$z^{[l]} = W^{[l]} a^{[l-1]} + \cancel{b^{[l]}}$

$z^{[l]} = W^{[l]} a^{[l-1]}$

$z^{[l]}_{norm}$

$\tilde{z}^{[l]} = \gamma^{[l]} z^{[l]}_{norm} + \beta^{[l]}$

implementing gradient descent

for $t=1$ ... num Mini batches

Compute forward prop on $X^{\{t\}}$

each hidden layer use BN to replace $z^{[l]}$ with $\tilde{z}^{[l]}$

Use back prop to compute $dw^{[l]}, \cancel{db^{[l]}}, d\beta^{[l]}, d\gamma^{[l]}$

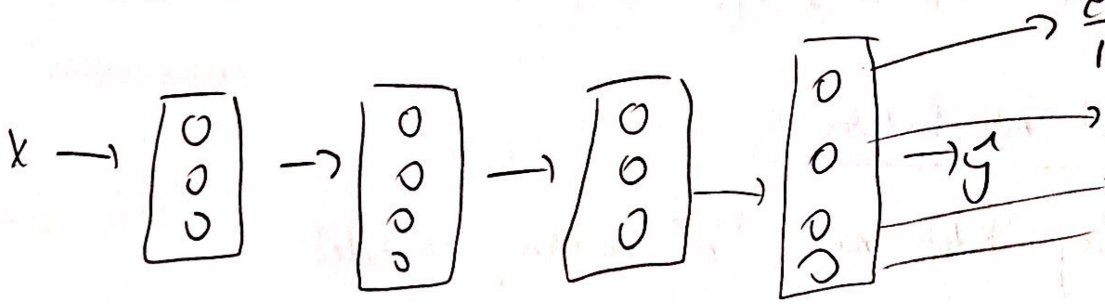
Update parameters $w^{[l]} := w^{[l]} - \alpha\, dw^{[l]}$

$$\beta^{[l]} := \beta^{[l]} - \alpha\, d\beta^{[l]}$$

$$\gamma^{[l]} := \gamma^{[l]} - \alpha\, d\gamma^{[l]}$$

works momentum, RMSprop/Adam

Softmax regression Classification fazla olduğunda kullanılır



$$\frac{e^5}{176.3} = \begin{bmatrix} 0.842 \\ 0.042 \\ 0.002 \\ 0.114 \end{bmatrix}_{4\times1}$$

Sonuç bu

$z^{[l]} = w^{[l]} a^{[l-1]} + b^{[l]}$

Activation function

$t = e^{(z^{[l]})}$

$$a^{[l]} = \frac{e^{z^{[l]}}}{\sum_{j=1}^{4} t_i}, \quad a_i^{[l]} = \frac{t_i}{\sum_{j=1}^{4} t_i}$$

$$\underset{(4,1)}{a^{[l]}} = g^{[l]}(\underset{(4,1)}{z^{[l]}})$$

$$z^{[l]} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \end{bmatrix}$$

$$t = \begin{bmatrix} e^5 \\ e^2 \\ e^{-1} \\ e^3 \end{bmatrix} = \begin{bmatrix} 148.4 \\ 7.4 \\ 0.4 \\ 20.1 \end{bmatrix}, \quad \sum_{j=1}^{4} t_j = 176.3$$

$$a^{[l]} = \frac{t}{176.3}$$

Training a softmax classifier 01.03

— hard max

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Softmax regression generalizes logistic regression to C classes

$$y^{(1)} = \underset{(4,1)}{\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}} \qquad a^{[L](1)} = \underset{(4,1)}{\hat{y}^{(1)}} = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.1 \\ 0.4 \end{bmatrix} \qquad C=4$$

$$\mathcal{L}(\hat{y}, y) = -\sum_{j=1}^{4} y_j \log \hat{y}_j$$

$$-\underbrace{y_2}_{1} \log \hat{y}_2 = -\log \hat{y}_2$$ → diğerleri 0 olur.

$$J(w^{[1]}, b^{[1]} \ldots) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$

$$Y = [y^{(1)}, y^{(2)} \ldots, y^{(m)}] \qquad \hat{Y} = [\hat{y}^{(1)} \ldots \hat{y}^{(m)}]$$

$$= \underset{(4,m)}{\begin{bmatrix} 0 & 0 & \\ 1 & 0 & \ldots \\ 0 & 1 & \\ 0 & 0 & \end{bmatrix}} \qquad = \underset{(4,m)}{\begin{bmatrix} 0.3 & \\ 0.2 & \\ 0.1 & \ldots \\ 0.4 & \end{bmatrix}}$$