7/3/19

EE6337: Deep Learning

Today: • Review

- Adam
- Batch Normalization
- Regularization: norm penalty, early stopping, dropout

• Adam: Adaptive Moments - a combination of RMS prop & momentum

- $\underline{s} \leftarrow \rho_1 \underline{s} + (1-\rho_1) g$ $\quad \rho_1, \rho_2$: hyper parameters
- $\underline{r} \leftarrow \rho_2 \cdot \underline{r} + (1-\rho_2) . g \odot g$
- $\hat{\underline{s}} = \dfrac{s}{1-\rho_1^t}$ (bias correction) $\quad t$: iteration count
- $\hat{\underline{r}} = \dfrac{r}{1-\rho_2^t}$ (bias correction).
- $\Delta\theta = \dfrac{-\varepsilon}{\sqrt{\delta + \hat{\underline{r}}}} \odot \hat{\underline{s}}$.
- $\theta \leftarrow \theta + \Delta\theta$

Batch normalization: $\hat{y}_i = w_1 \cdots w_l x$

$\underline{w} = [w_1 \cdots w_l]^T$

$\underline{w} \leftarrow \underline{w} - \varepsilon g \qquad g = [g_1 \cdots g_l]^T$

$\hat{y}_{i+1} = (w_1 - \varepsilon g_1) \cdots (w_l - \varepsilon g_l) \cdot x$

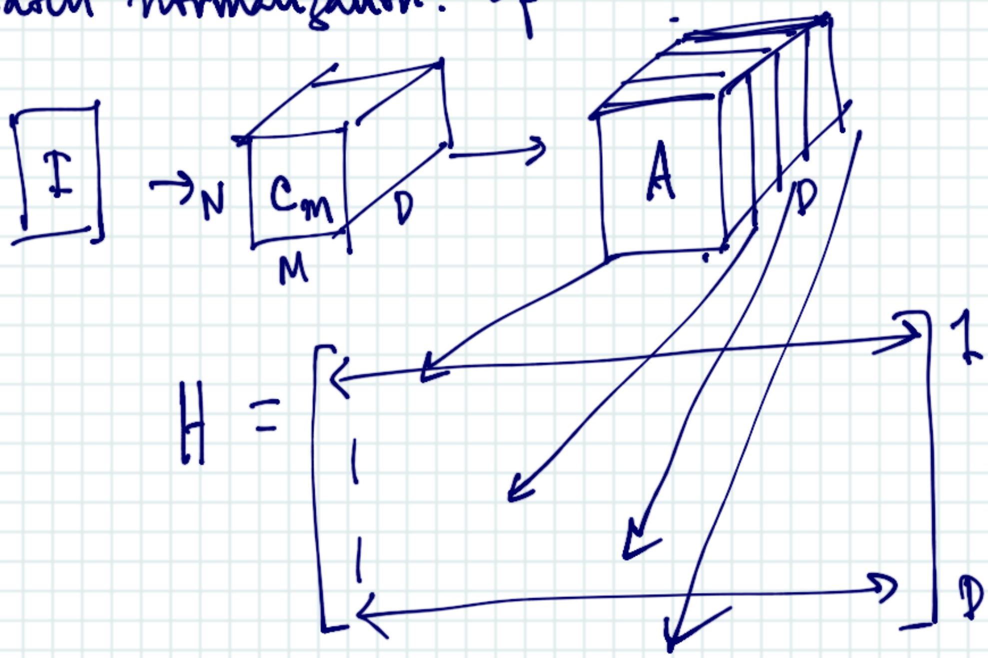
$$= \cdots + \varepsilon^2 \underbrace{g_1 g_2} \prod_{i=3}^{l} w_i + \varepsilon^l g_1 \cdots g_l$$

observation: The choice of ε should be such that the contribution of the second order gradient relation is reduced, while also reducing the contribution of the other powers of ε.

This is a hard problem. ($h_i = \underline{h_{i-1}} \cdot w_i$, h_i: output at ith layer, w_i: weight of ith layer)

How can batch normalization help?

Batch normalization: If



$$H' = \frac{H - \mu}{\sigma}$$