NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas

Department : E & ECE, IIT Kharagpur

Topic

Lecture 48: Normalization - III

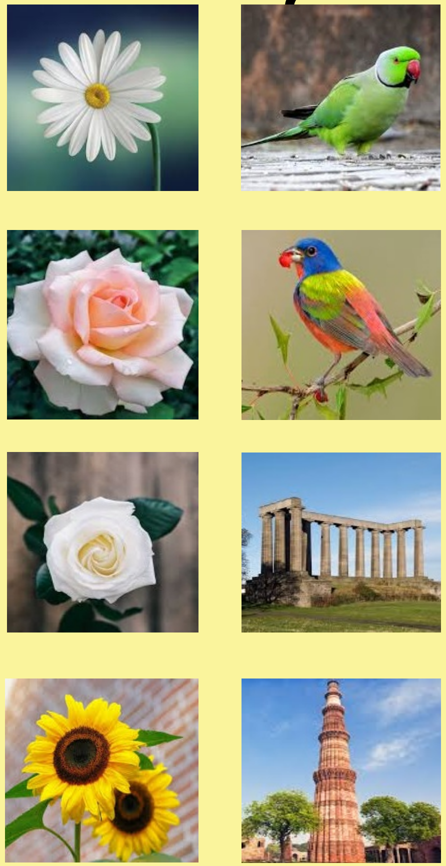
CONCEPTS COVERED

Concepts Covered:

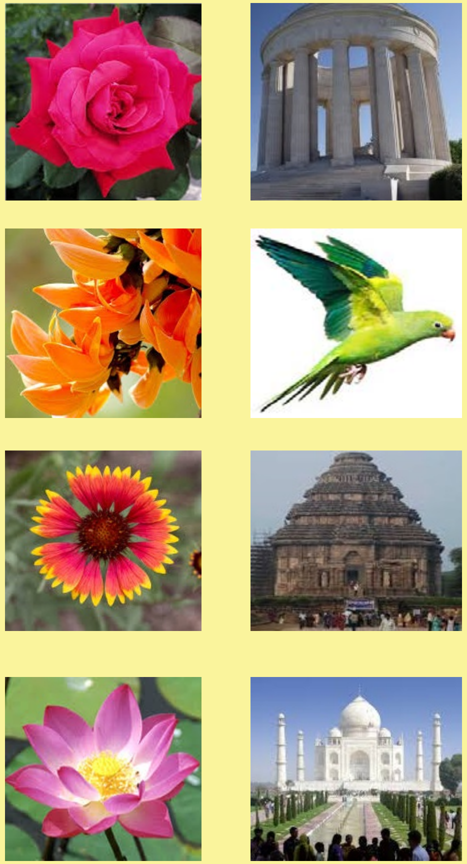
- Deep Neural Network
  - Normalization
  - Batch Normalization
  - Layer Normalization
  - Instance Normalization
  - Group Normalization

Normalization

Why normalization ?



Batch 1



Batch 2

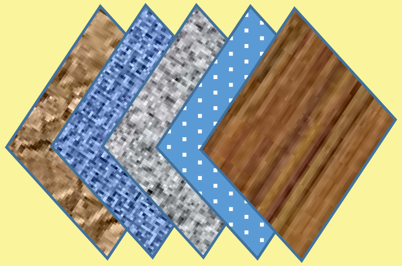
Normalization
In
Hidden Layers

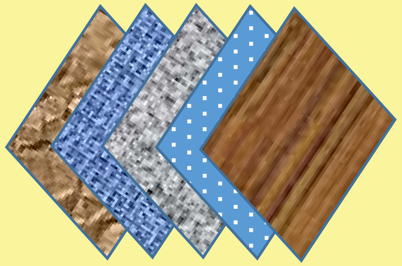
Different normalization techniques

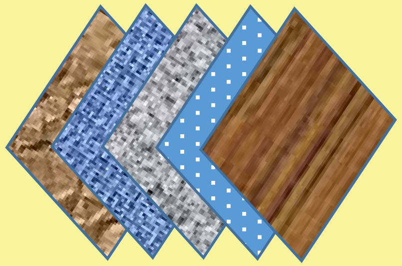
- Batch Normalization
- Layer Normalization
- Instance Normalization
- Group Normalization

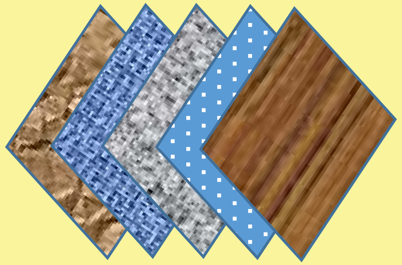
Batch Normalization

Batch Normalization

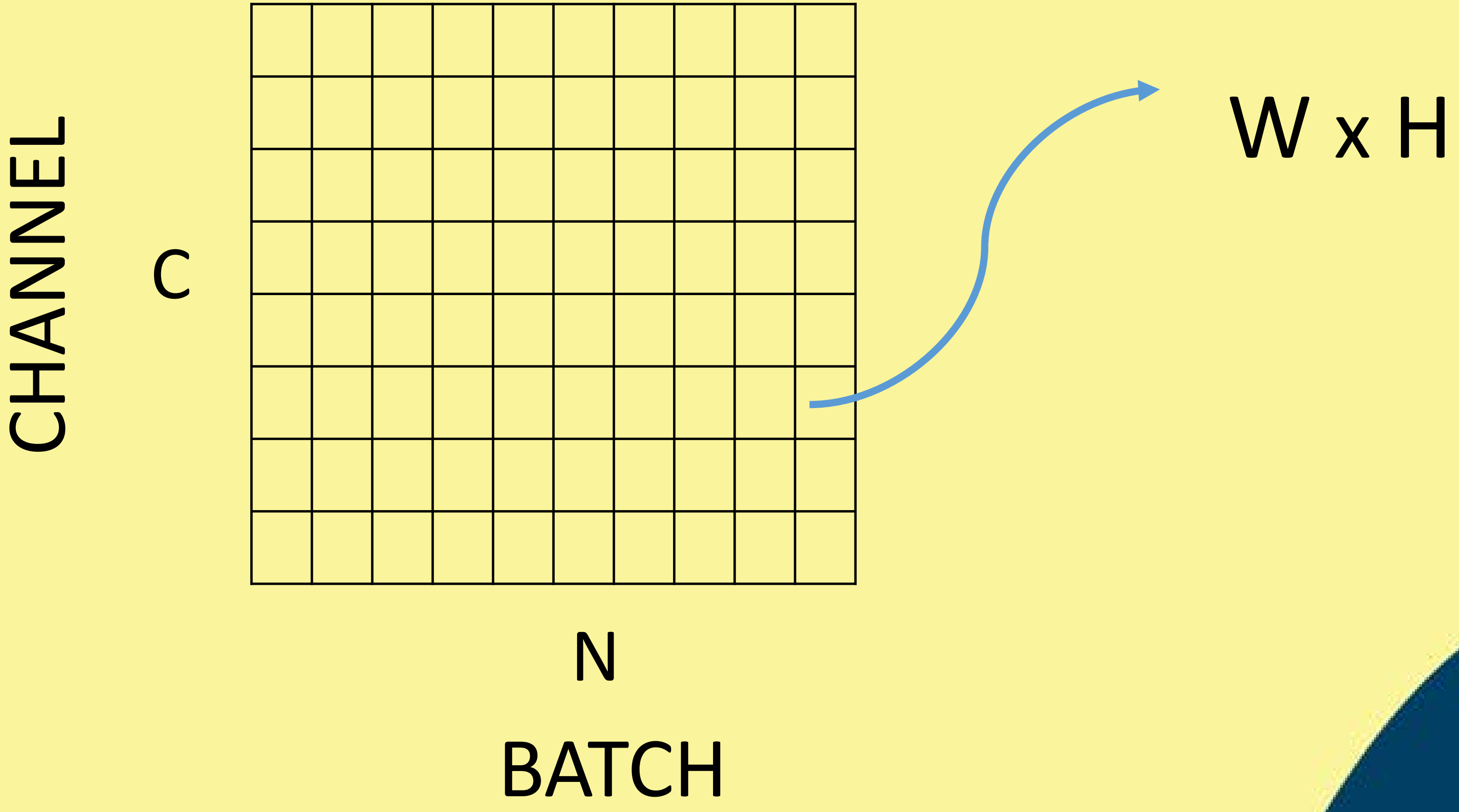




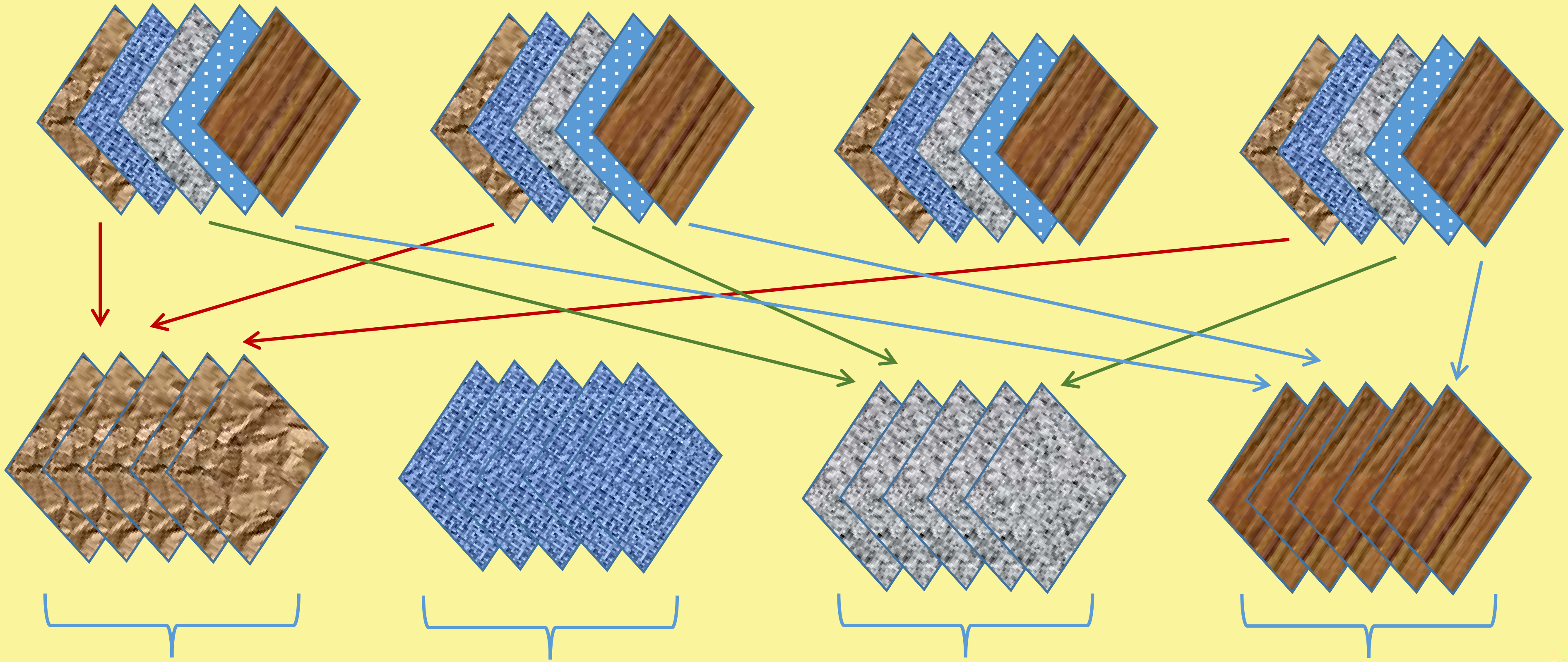




Normalization



Batch Normalization



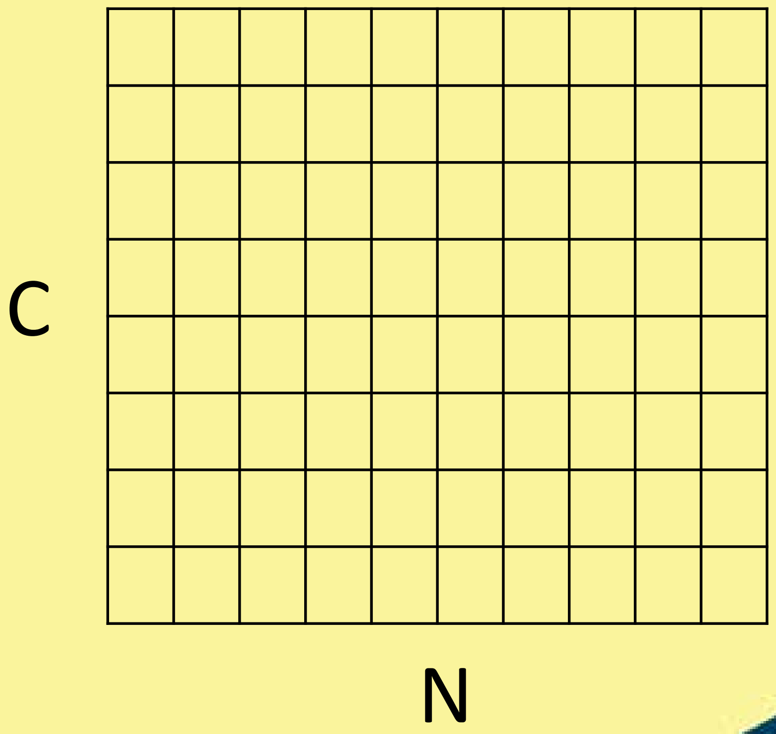
Batch Normalization

$$x \in \mathbb{R}^{N \times C \times W \times H}$$

$$\mu_C = \frac{1}{NWH} \sum_{i=1}^{N} \sum_{j=1}^{W} \sum_{k=1}^{H} x_{iCjk}$$

$$\sigma_C^2 = \frac{1}{NWH} \sum_{i=1}^{N} \sum_{j=1}^{W} \sum_{k=1}^{H} (x_{iCjk} - \mu_C)^2$$

$$\hat{x} = \frac{x - \mu_C}{\sqrt{\sigma_C^2 + \epsilon}}$$



Batch Normalization

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Batch Normalization

$$\frac{\partial \ell}{\partial \widehat{x}_i} = \frac{\partial \ell}{\partial y_i} \cdot \gamma$$

$$\frac{\partial \ell}{\partial \sigma_{\mathcal{B}}^2} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial \widehat{x}_i} \cdot (x_i - \mu_{\mathcal{B}}) \cdot \frac{-1}{2}(\sigma_{\mathcal{B}}^2 + \epsilon)^{-3/2}$$

$$\frac{\partial \ell}{\partial \mu_{\mathcal{B}}} = \left(\sum_{i=1}^{m} \frac{\partial \ell}{\partial \widehat{x}_i} \cdot \frac{-1}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \right) + \frac{\partial \ell}{\partial \sigma_{\mathcal{B}}^2} \cdot \frac{\sum_{i=1}^{m} -2(x_i - \mu_{\mathcal{B}})}{m}$$

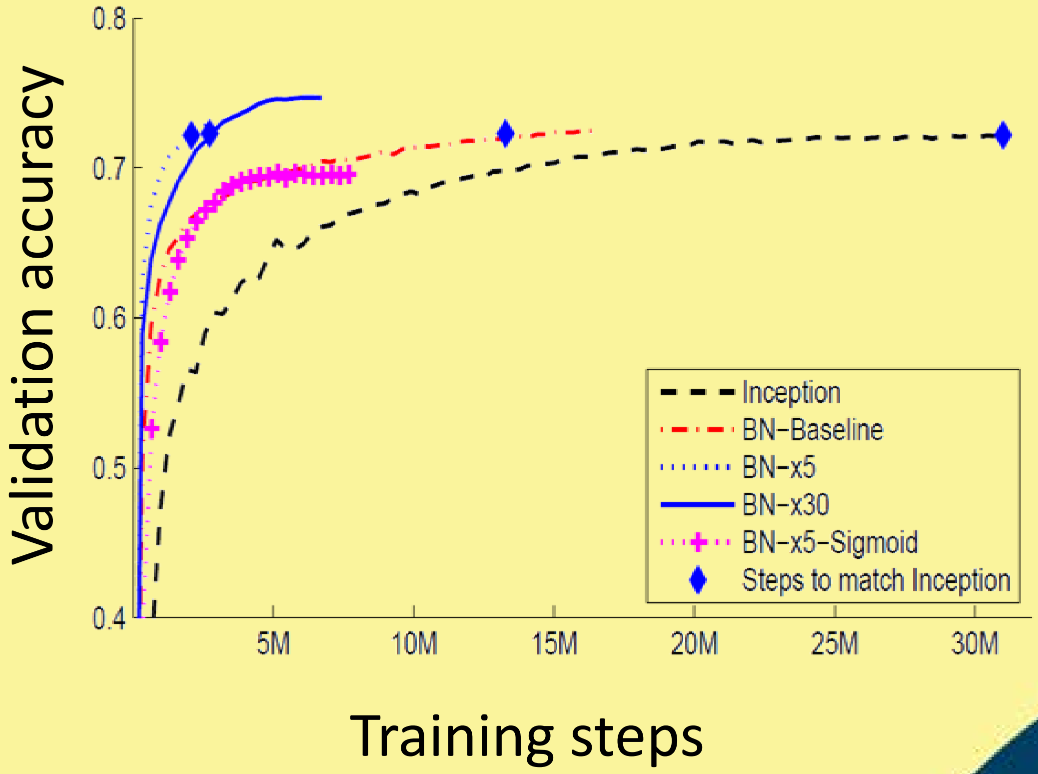
$$\frac{\partial \ell}{\partial x_i} = \frac{\partial \ell}{\partial \widehat{x}_i} \cdot \frac{1}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} + \frac{\partial \ell}{\partial \sigma_{\mathcal{B}}^2} \cdot \frac{2(x_i - \mu_{\mathcal{B}})}{m} + \frac{\partial \ell}{\partial \mu_{\mathcal{B}}} \cdot \frac{1}{m}$$

$$\frac{\partial \ell}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot \widehat{x}_i$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i}$$

Effect of Batch Normalization

- **Inception:** A network, trained with the initial learning rate of 0.0015.
- **BN-Baseline:** Same as Inception with Batch Normalization before each nonlinearity.
- **BN-x5:** The initial learning rate was
- increased by a factor of 5, to 0.0075.
- **BN-x30:** Like BN-x5, but with the initial learning rate 0.045 (30 times that of Inception).
- **BN-x5-Sigmoid:** Like BN-x5, but with sigmoid nonlinearity instead of ReLU.



Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015)

NPTEL ONLINE CERTIFICATION COURSES

Thank you