NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas

Department : E & ECE, IIT Kharagpur

Topic

Lecture 46: Normalization

CONCEPTS COVERED

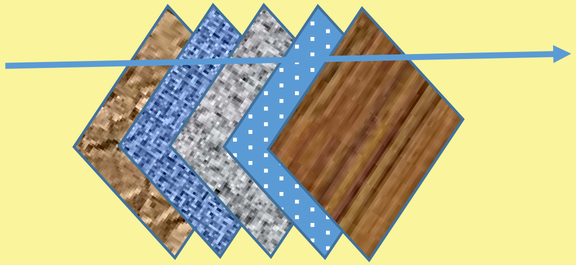
Concepts Covered:

- Deep Neural Network
  - Gradient Descent Challenges
  - Normalization
  - Batch Normalization
  - Layer Normalization
  - Instance Normalization
  - Group Normalization

Normalization

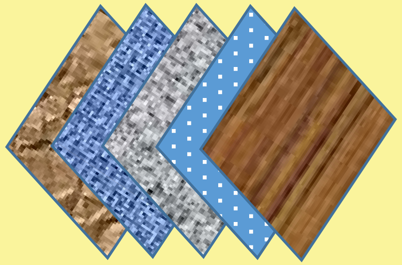
Local Response Normalization (Inter-Channel)

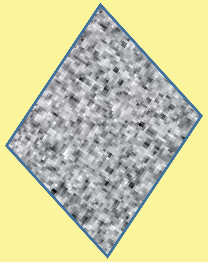
$$b_{x,y}^{i} = \frac{a_{x,y}^{i}}{\left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} \left(a_{x,y}^{j}\right)^2\right)^{\beta}}$$



Local Response Normalization (Intra-Channel)

$$b_{x,y}^{i} = \frac{a_{x,y}^{i}}{\left(k + \alpha \sum_{p=\max(0,x-n/2)}^{\max(W,x+n/2)} \sum_{q=\max(0,y-n/2)}^{\min(H,y+n/2)} \left(a_{p,q}^{i}\right)^{2}\right)^{\beta}}$$



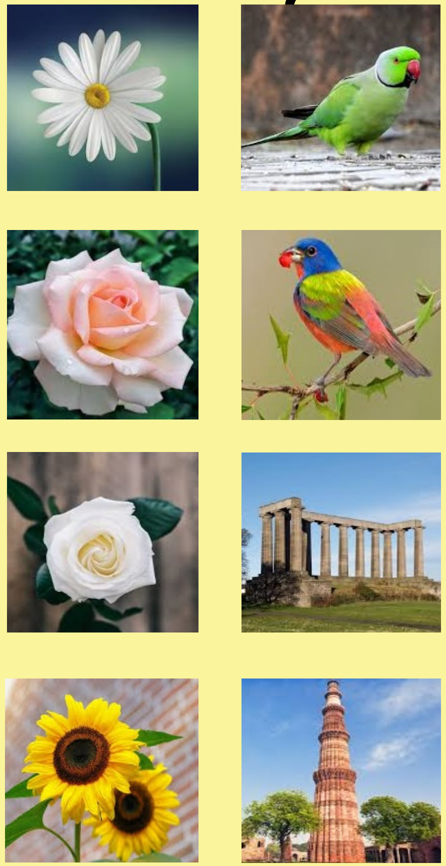


Normalization

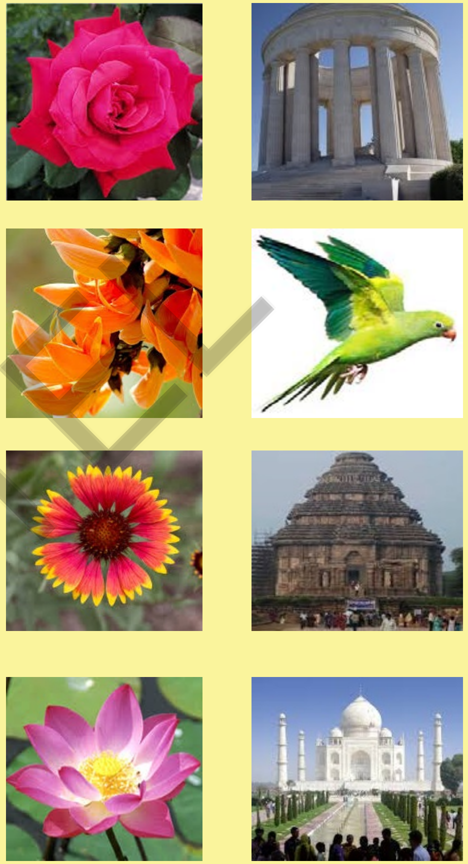
- Normalization that address the problem of covariate shift.
- Makes learning process faster.
- Different layers learn independently of others.

What does a classifier learn?

Why normalization ?



Batch 1



Batch 2

NPTEL ONLINE CERTIFICATION COURSES

Thank you

NPTEL ONLINE CERTIFICATION COURSES

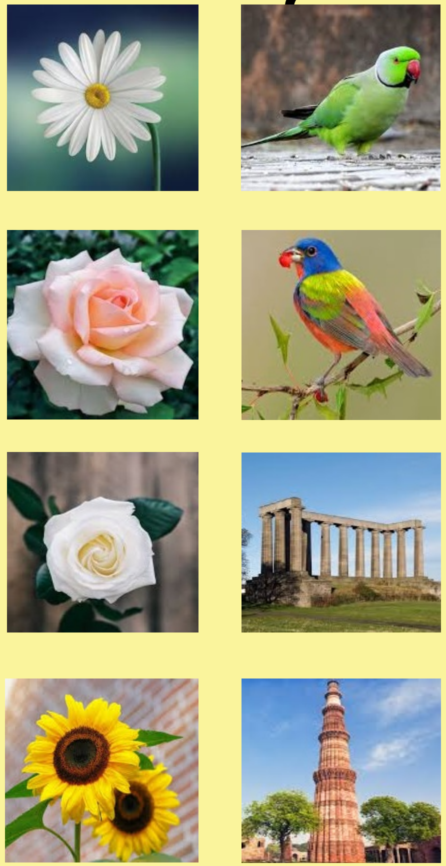
Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas
Department : E & ECE, IIT Kharagpur

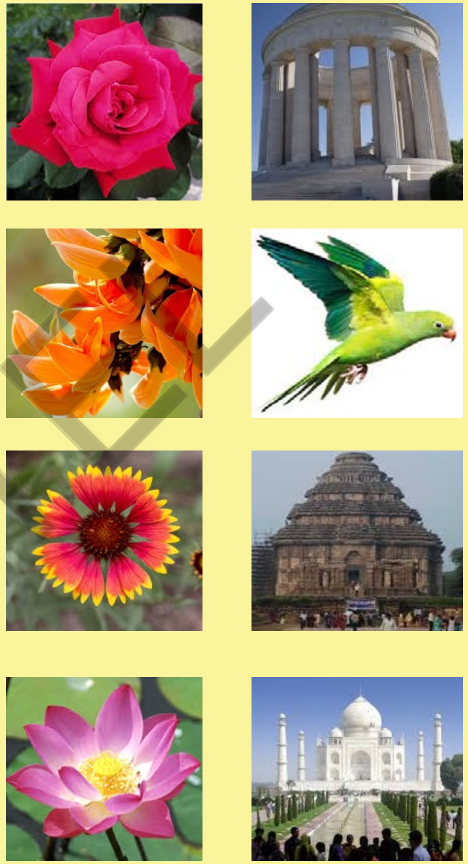
Topic

Lecture 47

Why normalization ?



Batch 1



Batch 2

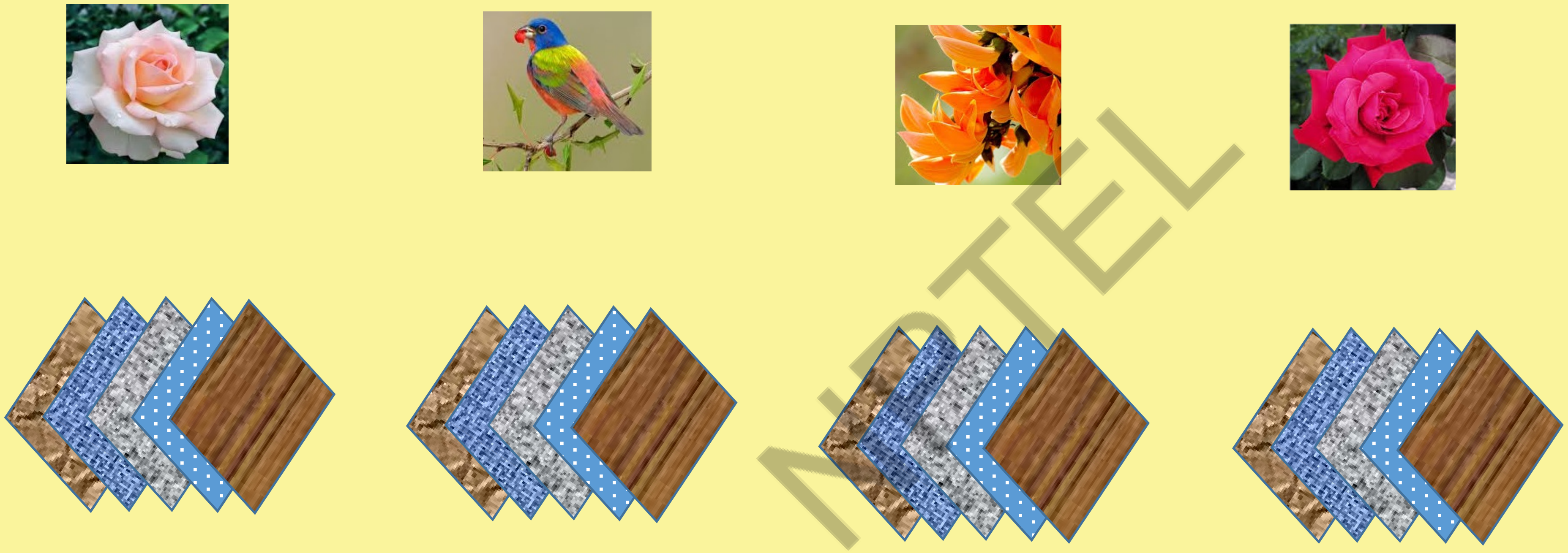
Normalization In Hidden Layers

Different normalization techniques

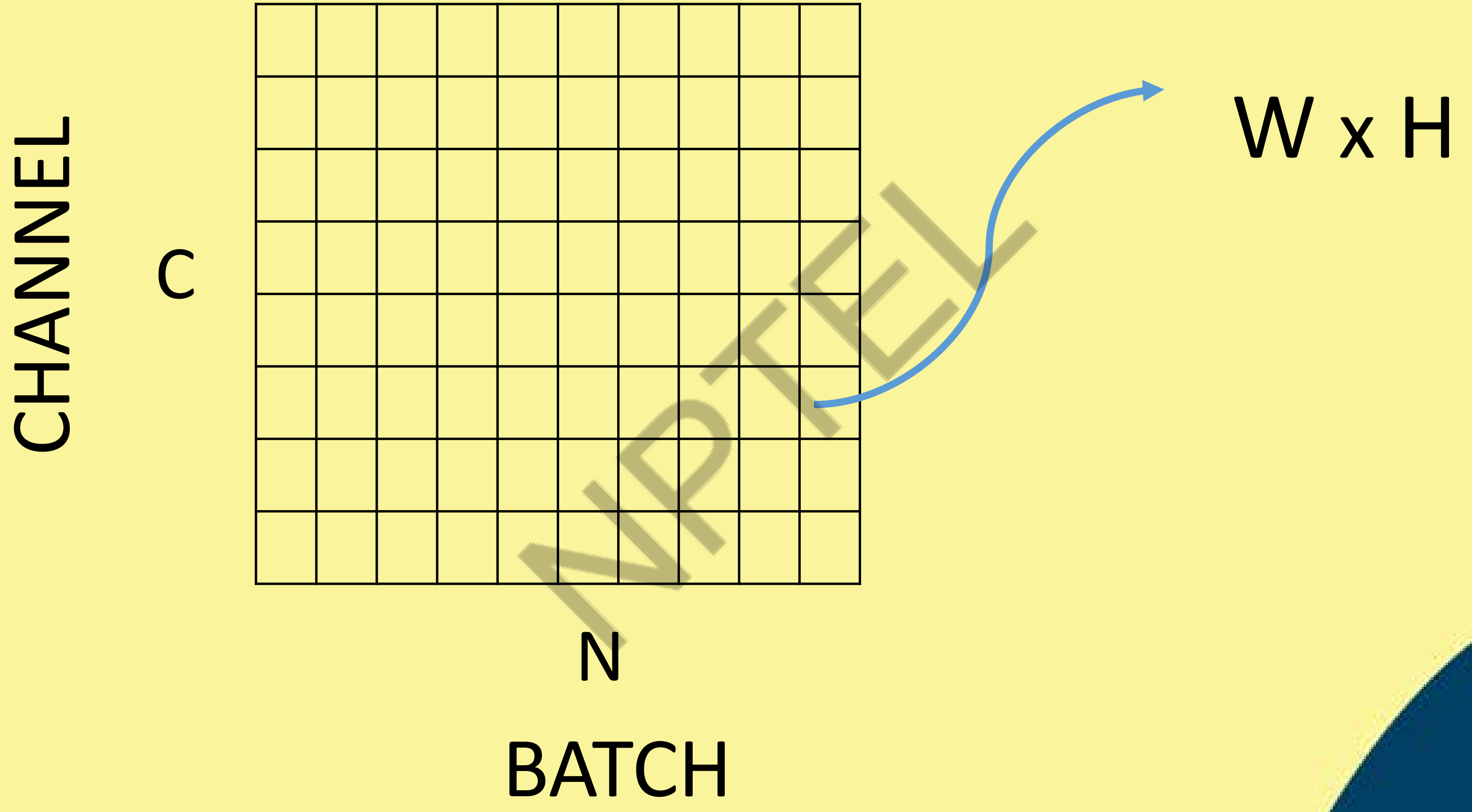
- Batch Normalization
- Layer Normalization
- Instance Normalization
- Group Normalization

Batch Normalization

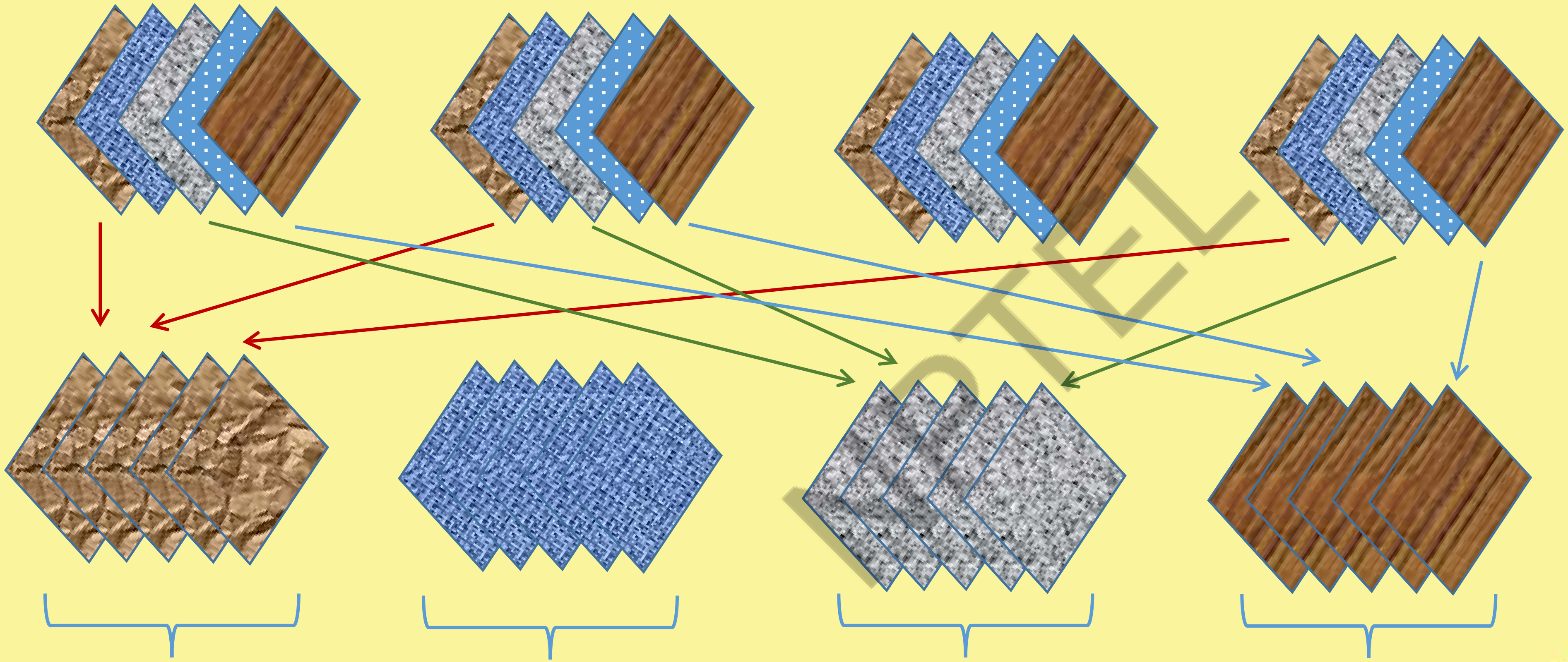
Batch Normalization



Normalization



Batch Normalization



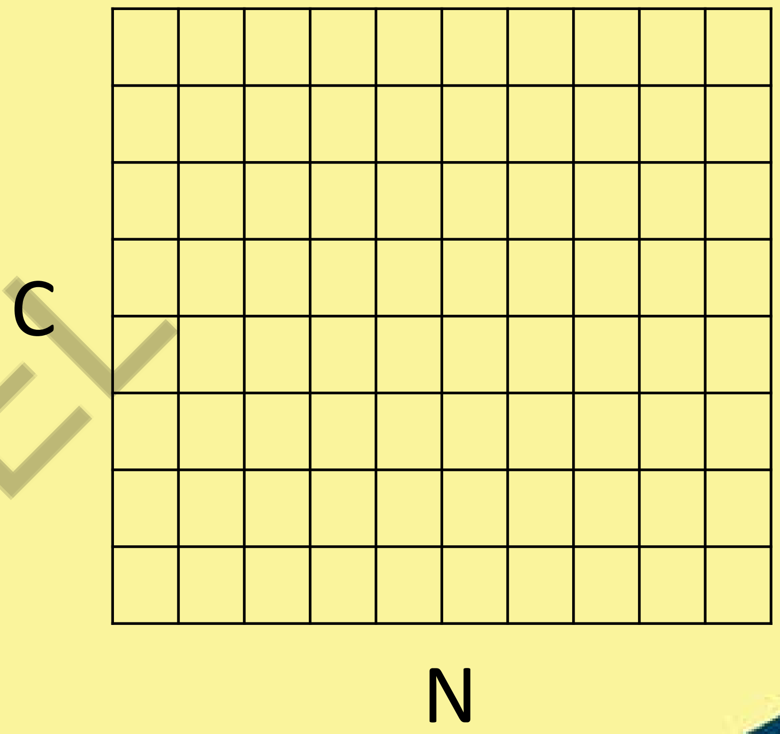
Batch Normalization

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

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

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

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



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

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

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

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

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

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

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

NPTEL ONLINE CERTIFICATION COURSES

Thank you

NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas

Department : E & ECE, IIT Kharagpur

Topic

Lecture 48: Normalization - III

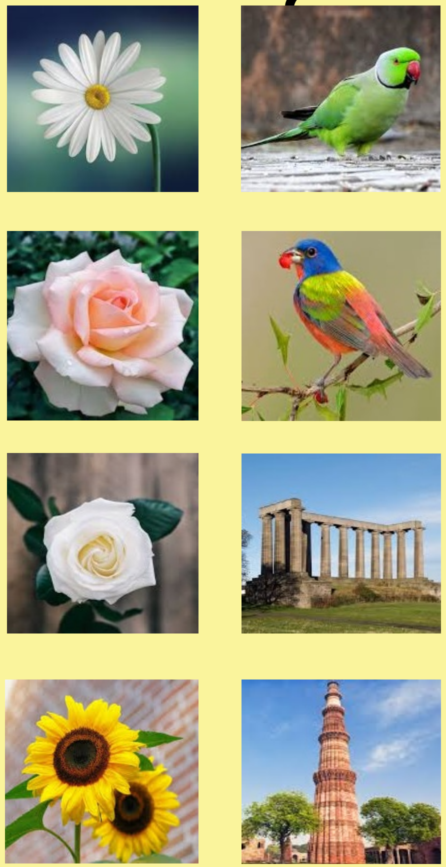
CONCEPTS COVERED

Concepts Covered:

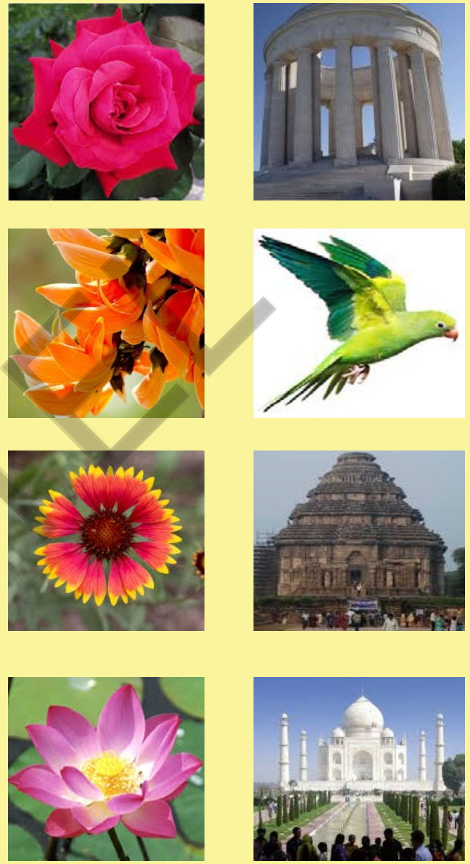
- Deep Neural Network
  - Normalization
  - Batch Normalization
  - Layer Normalization
  - Instance Normalization
  - Group Normalization

Normalization

Why normalization ?



Batch 1



Batch 2

Normalization In Hidden Layers

Different normalization techniques

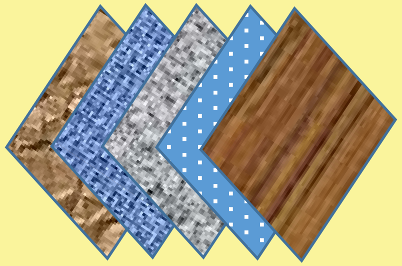
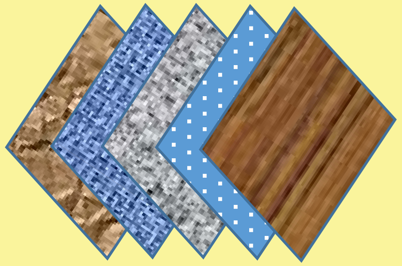
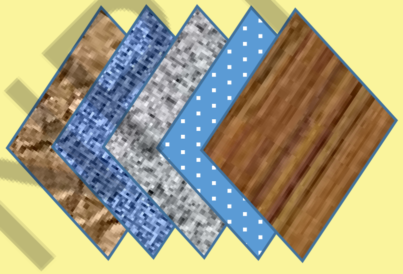
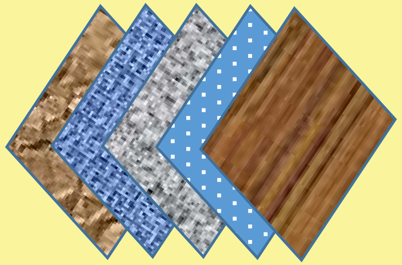
- Batch Normalization
- Layer Normalization
- Instance Normalization
- Group Normalization

Batch Normalization

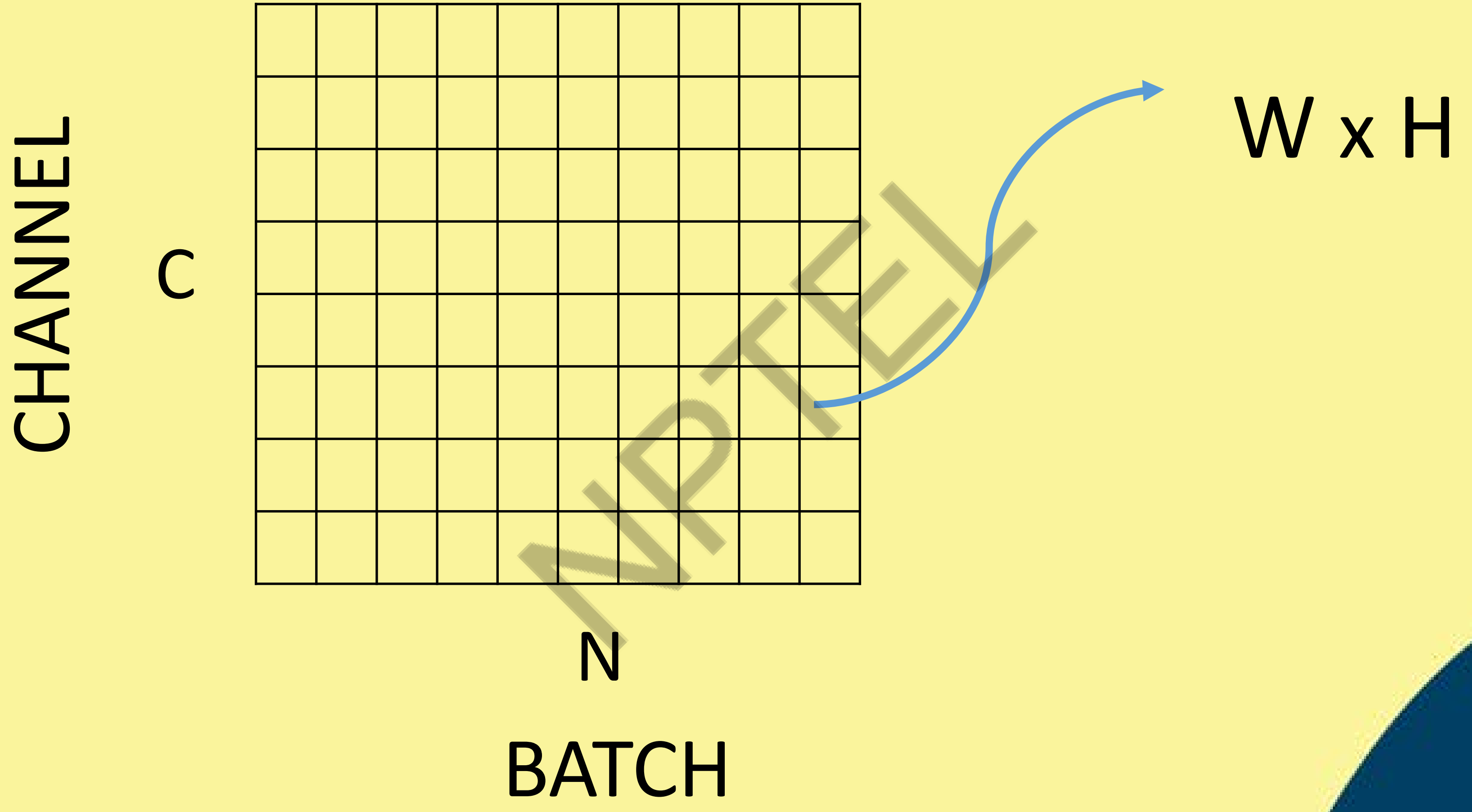
Batch Normalization



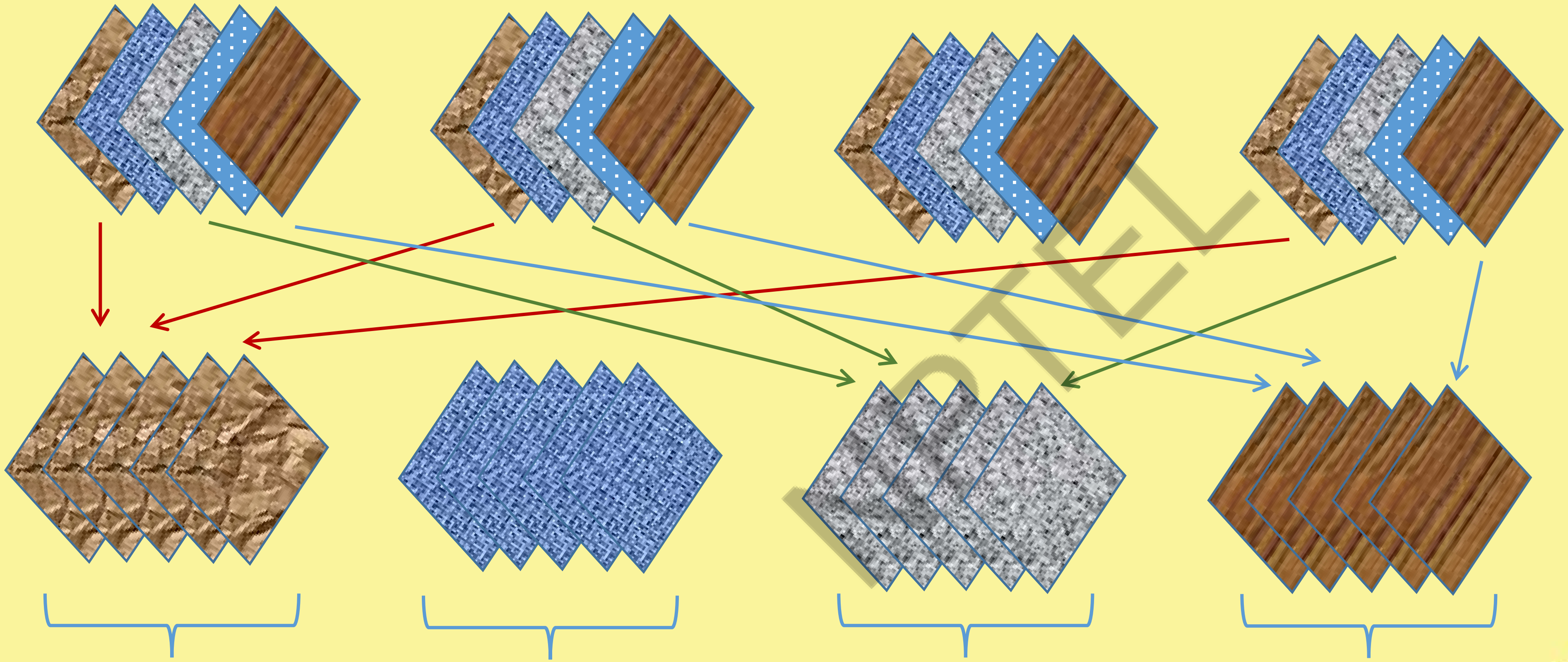


Normalization



Batch Normalization



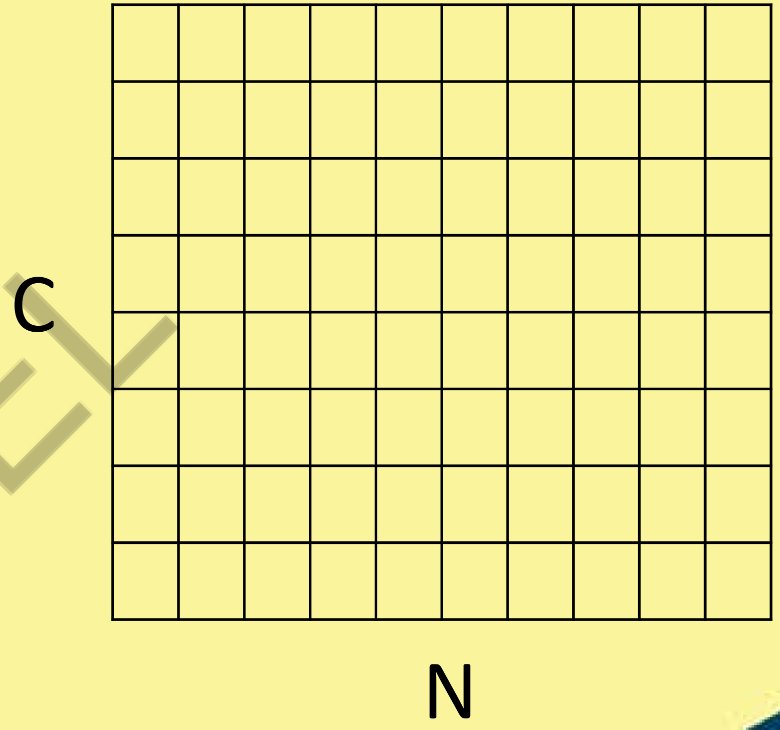
Batch Normalization

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

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

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

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



Batch Normalization

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1\ldots m}\}$;
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

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

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

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

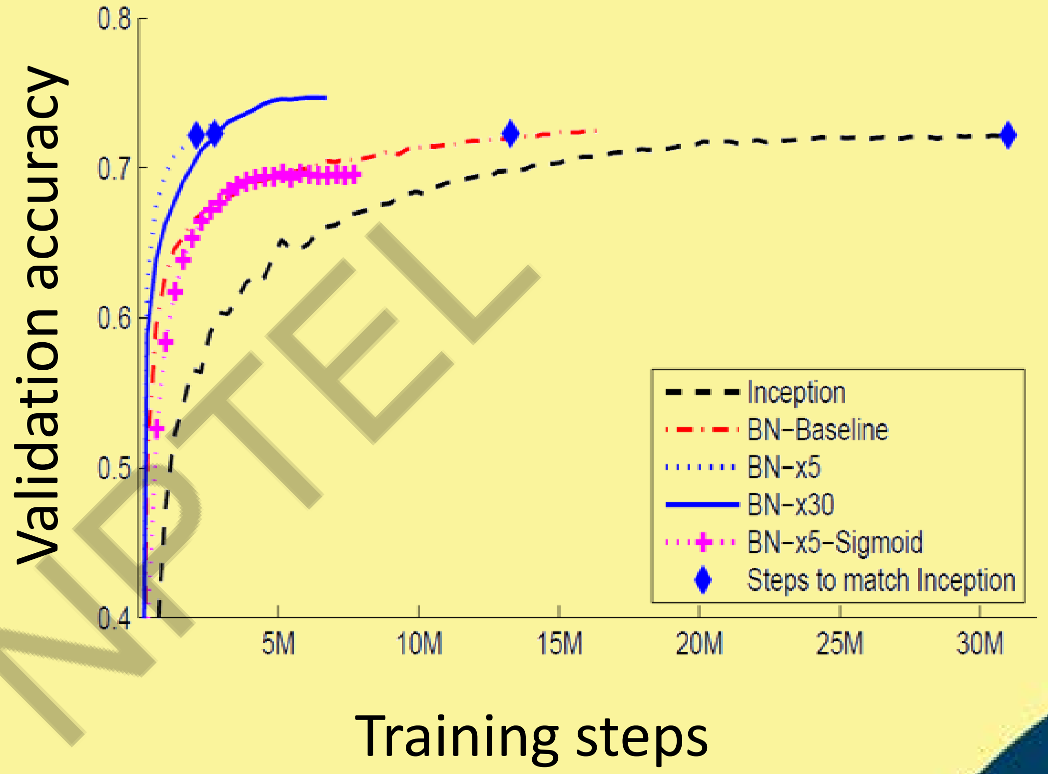
$$\frac{\partial \ell}{\partial x_i} = \frac{\partial \ell}{\partial \widehat{x}_i} \cdot \frac{1}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} + \frac{\partial \ell}{\partial \sigma_{\mathcal{B}}^2} \cdot \frac{2(x_i - \mu_{\mathcal{B}})}{m} + \frac{\partial \ell}{\partial \mu_{\mathcal{B}}} \cdot \frac{1}{m}$$

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

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

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

NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas

Department : E & ECE, IIT Kharagpur

Topic

Lecture 49: Normalization - IV

CONCEPTS COVERED

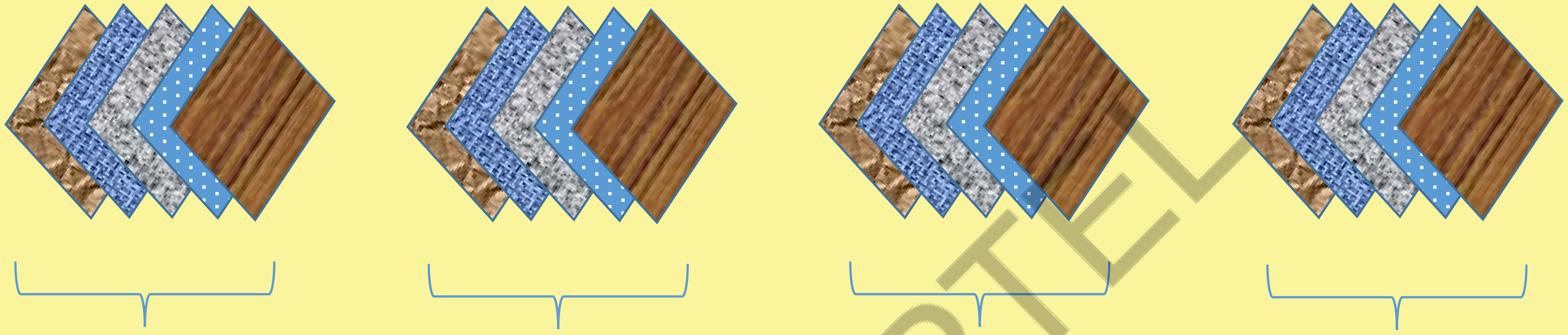
Concepts Covered:

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

Normalization

Layer Normalization

Layer Normalization



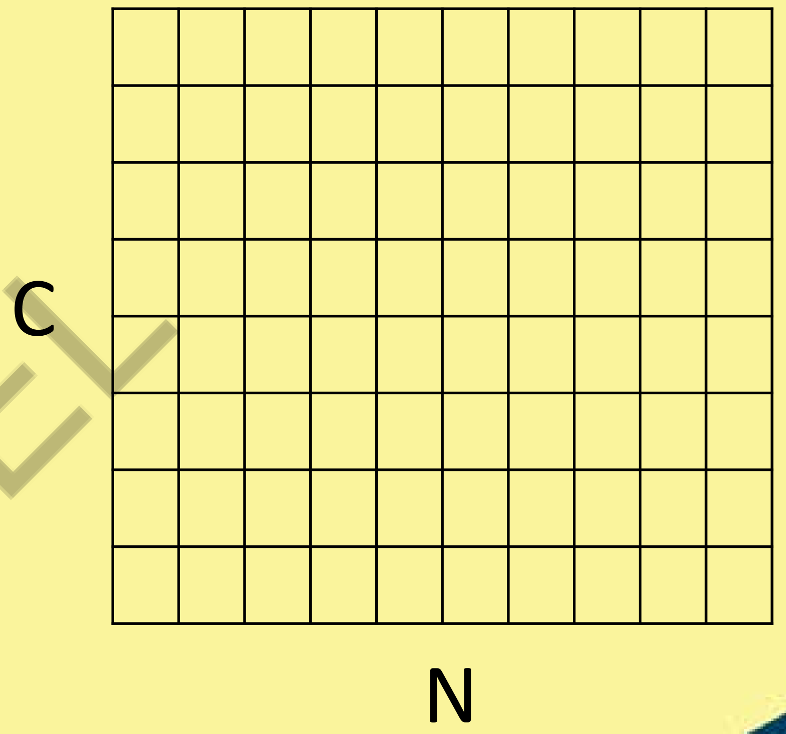
Layer Normalization

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

$$\mu_N = \frac{1}{CWH} \sum_{i=1}^{C} \sum_{j=1}^{W} \sum_{k=1}^{H} x_{Nijk}$$

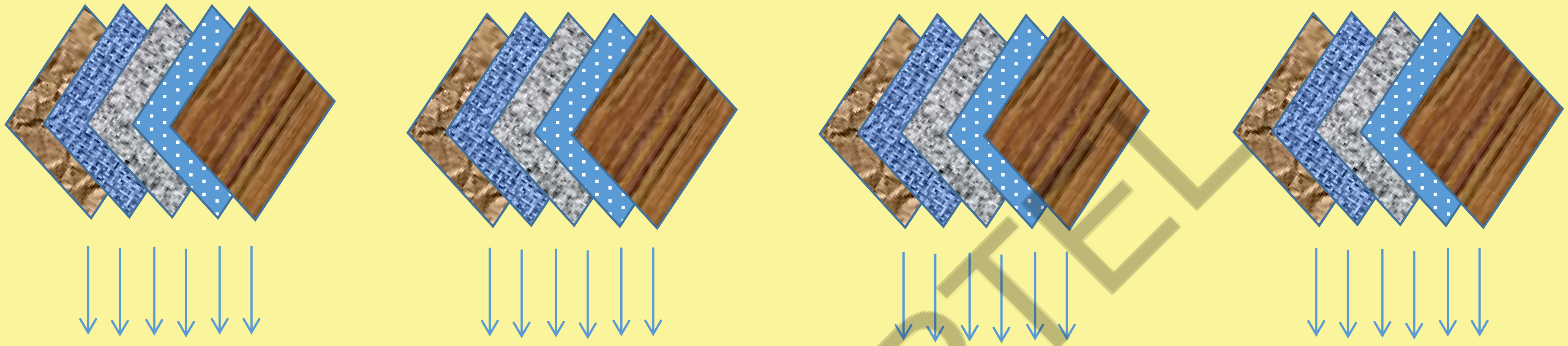
$$\sigma_N^2 = \frac{1}{CWH} \sum_{i=1}^{C} \sum_{j=1}^{W} \sum_{k=1}^{H} (x_{Nijk} - \mu_N)^2$$

$$\hat{x} = \frac{x - \mu_N}{\sqrt{\sigma_N^2 + \epsilon}}$$



Instance Normalization

Instance Normalization



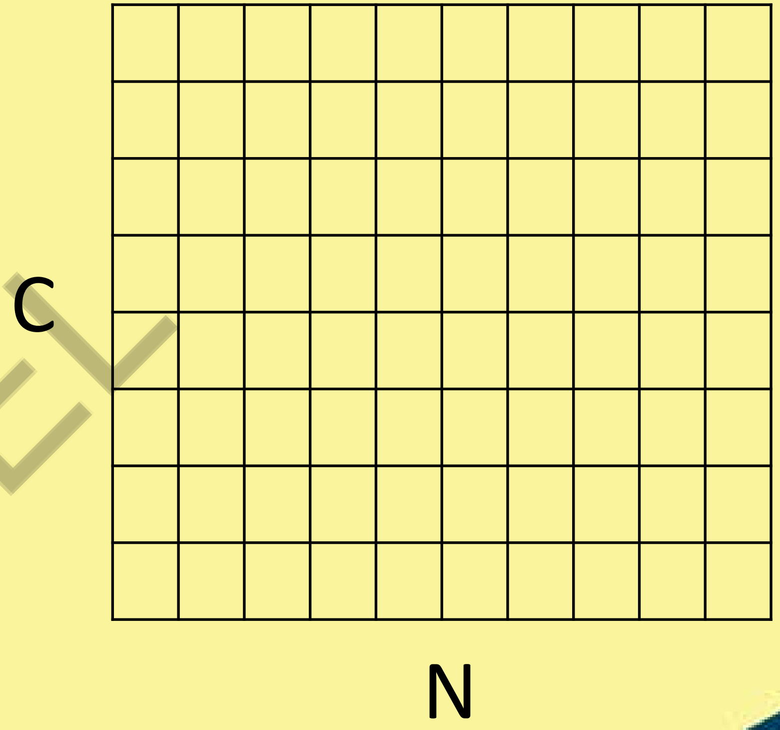
Instance Normalization

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

$$\mu_{NC} = \frac{1}{WH}\sum_{j=1}^{W}\sum_{k=1}^{H} x_{Nijk}$$

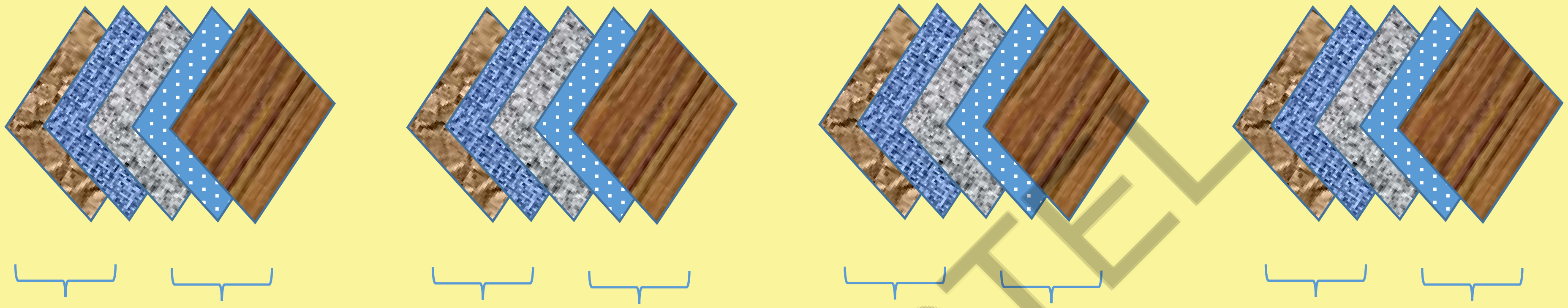
$$\sigma_{NC}^2 = \frac{1}{WH}\sum_{j=1}^{W}\sum_{k=1}^{H} (x_{Nijk} - \mu_N)^2$$

$$\hat{x} = \frac{x - \mu_{NC}}{\sqrt{\sigma_{NC}^2 + \epsilon}}$$



Group Normalization

Group Normalization



Group Normalization

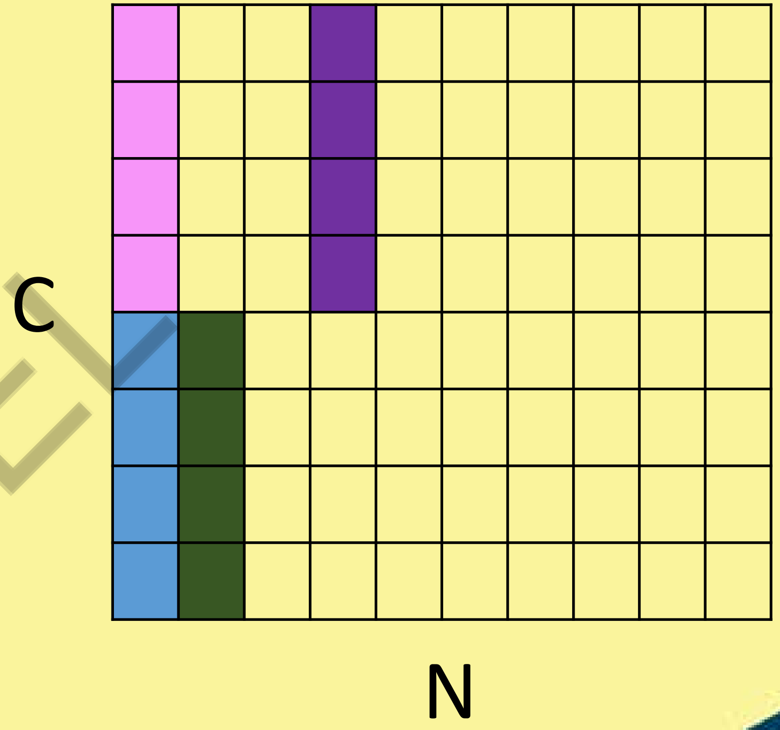
$x \in \mathbb{R}^{N\times C\times W\times H} \rightarrow \mathbb{R}^{N\times G\times C'\times W\times H} \quad C = G.C'$

G=number of groups

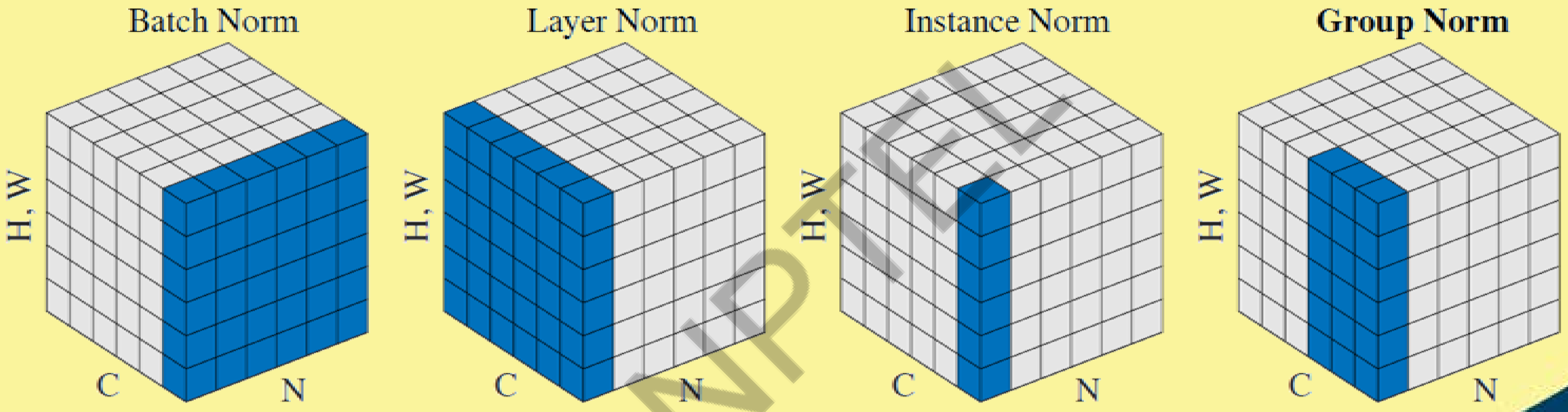
C'=number of channel per group

$$\mu_{NG} = \frac{1}{C'WH}\sum_{i=1}^{C'}\sum_{j=1}^{W}\sum_{k=1}^{H} x_{NGijk}$$

$$\sigma_{NG}^2 = \frac{1}{C'WH}\sum_{i=1}^{C'}\sum_{j=1}^{W}\sum_{k=1}^{H} (x_{NGijk} - \mu_{NG})^2$$

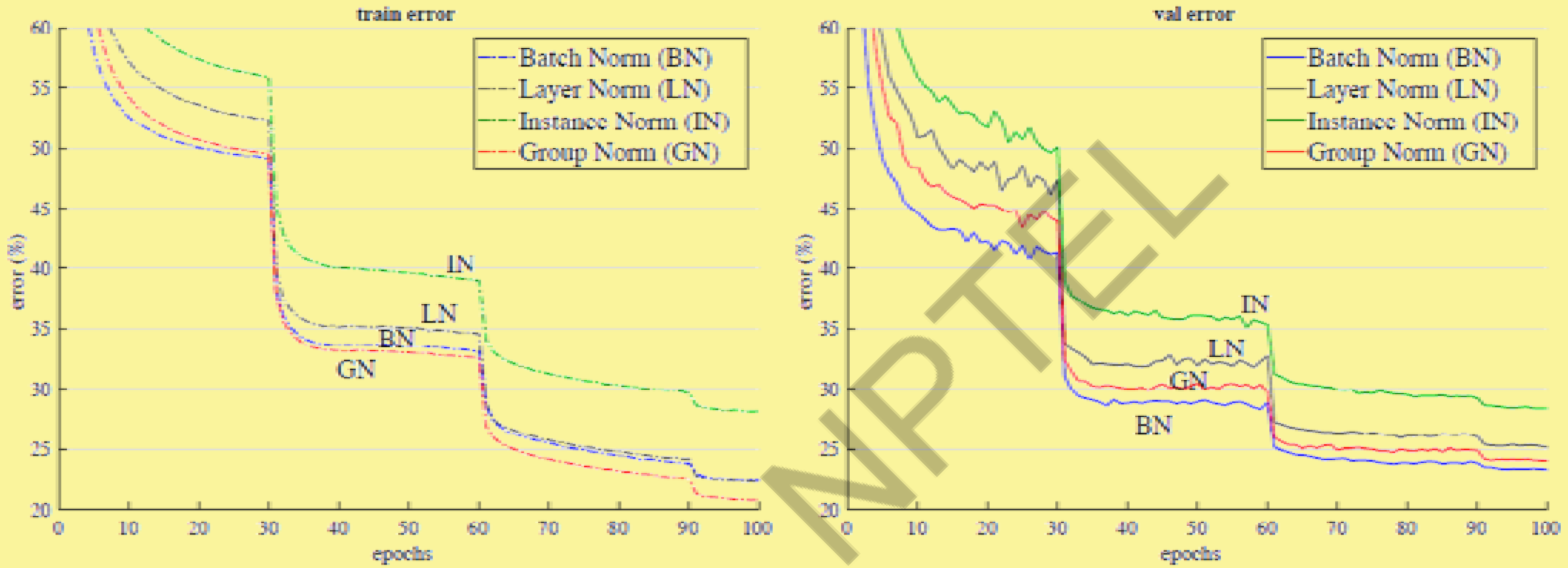


$$\hat{x} = \frac{x - \mu_{NG}}{\sqrt{\sigma_{NG}^2 + \epsilon}}$$



https://mlexplained.com/2018/11/30/an-overview-of-normalization-methods-in-deep-learning/

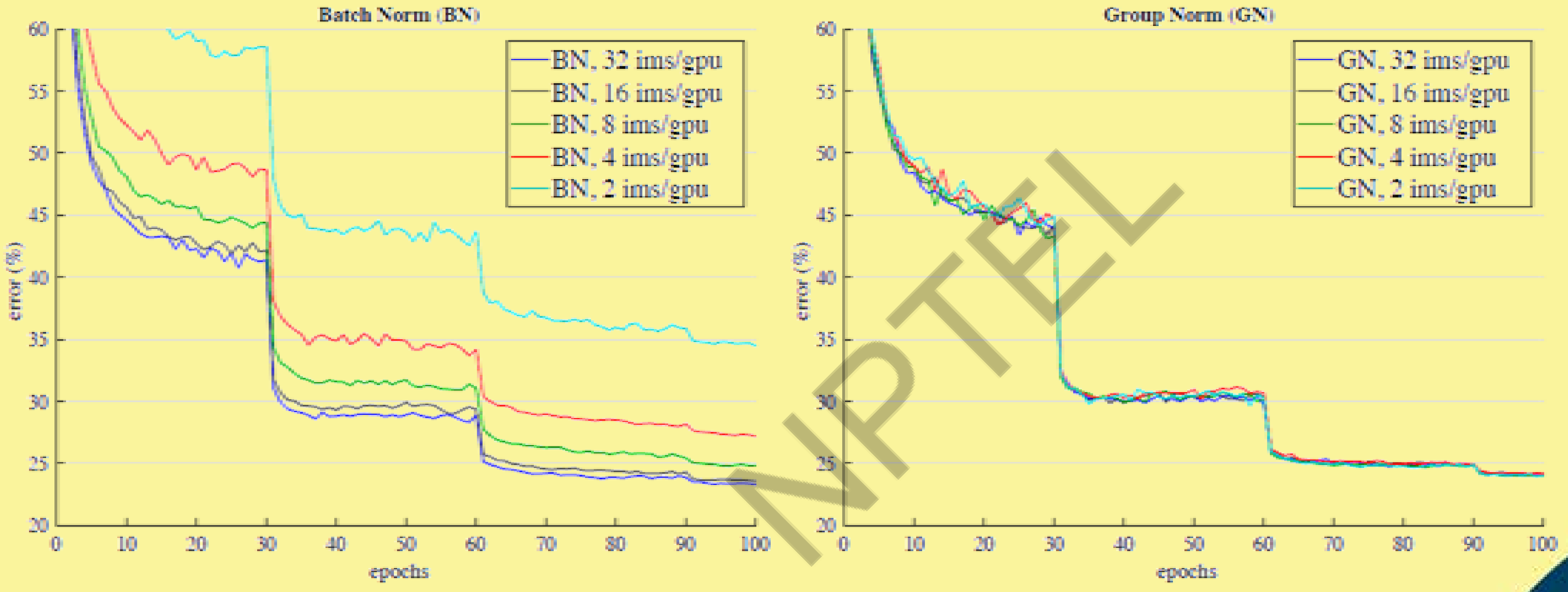
BN/LN/IN/GN Normalization



Model Name: Resnet-50, Dataset: Imagenet, Batch size: 32

Wu, Yuxin, and Kaiming He. "Group normalization." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

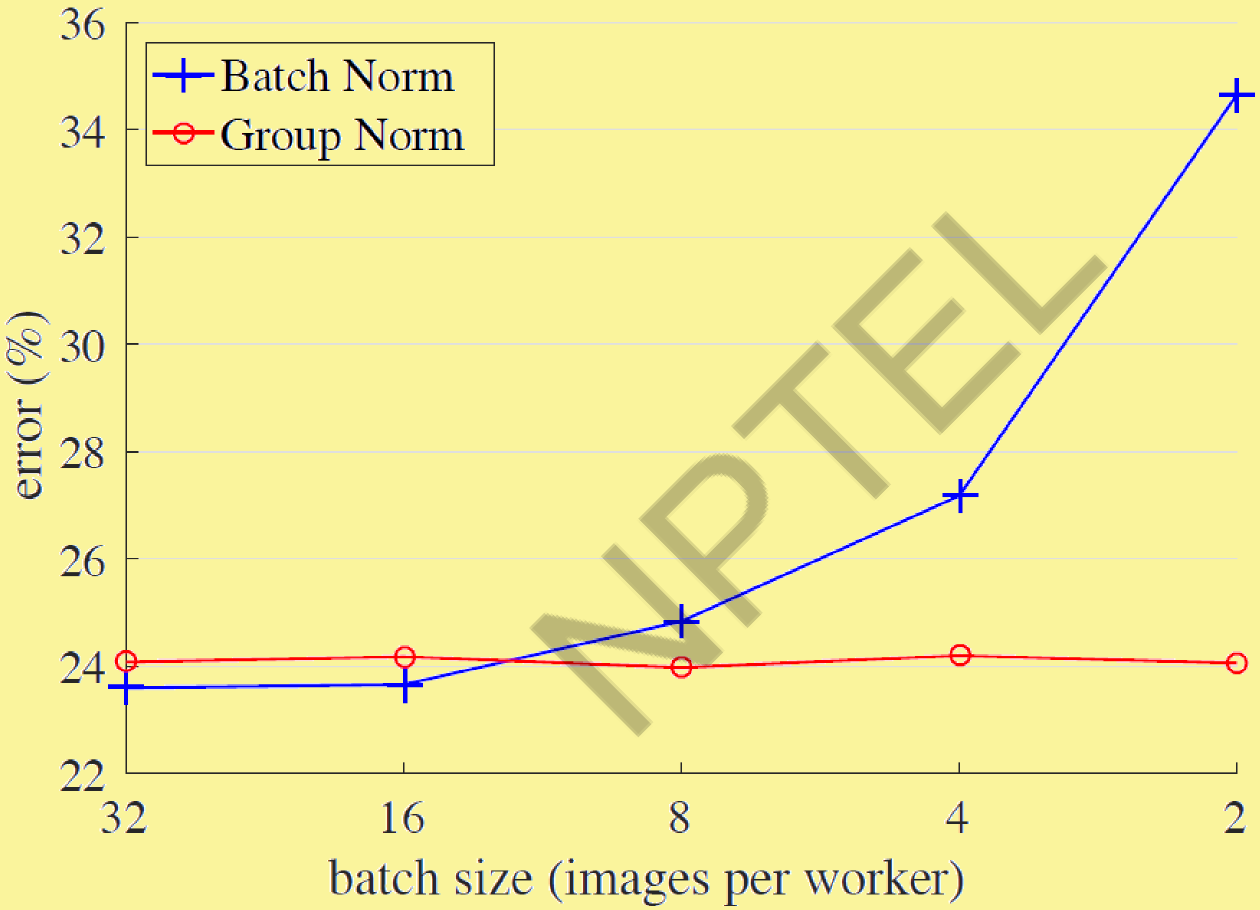
Batch/Group Normalization



Model Name: Resnet-50, Dataset: Imagenet

Wu, Yuxin, and Kaiming He. "Group normalization." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

Batch/Group Normalization



Wu, Yuxin, and Kaiming He. "Group normalization." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

NPTEL ONLINE CERTIFICATION COURSES

Thank you

NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas
Department : E & ECE, IIT Kharagpur

Topic

Lecture 50: Training Tricks

CONCEPTS COVERED

Concepts Covered:

- Deep Neural Network
  - Normalization
  - Underfitting/Ovefitting
  - Regularization
  - Dropout
  - Early Stopping

Regularization
Early stopping

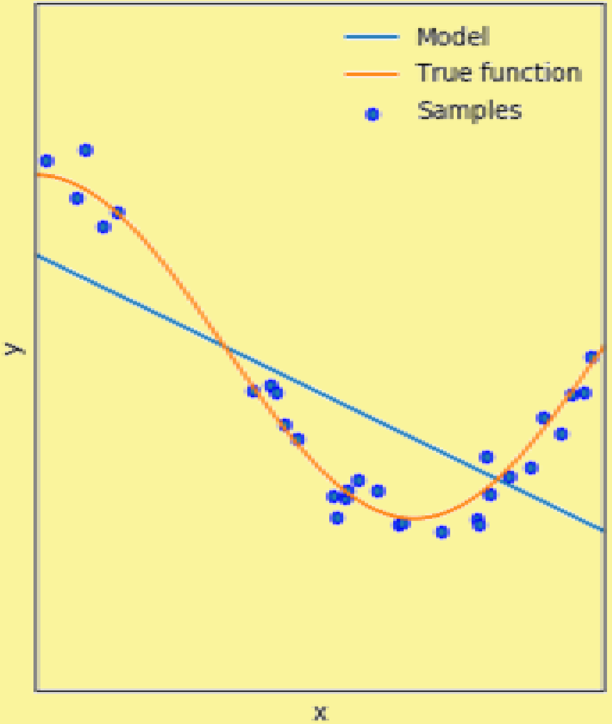
Overfitting/Underfitting

- Overfitting occurs when a statistical model or machine learning algorithm captures the noise of the data.
- Intuitively, overfitting occurs when the model or the algorithm fits the data too well.

- A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data.

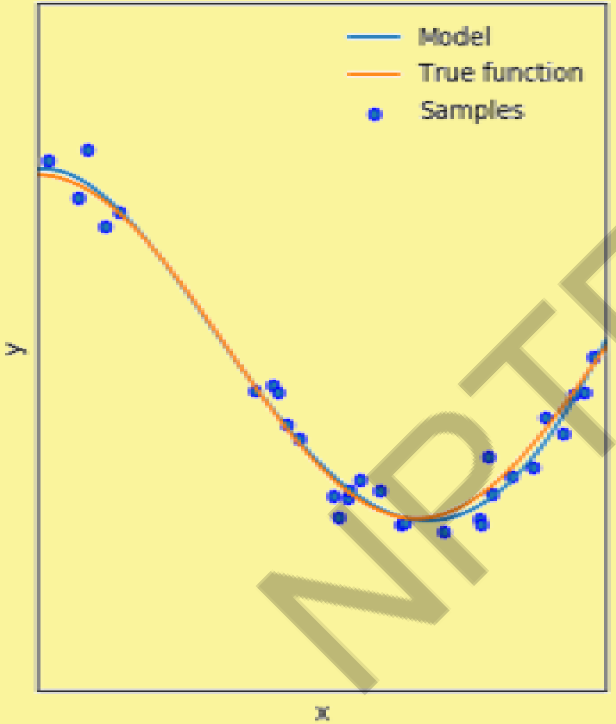
Overfitting/Underfitting: Regression

Degree: 1



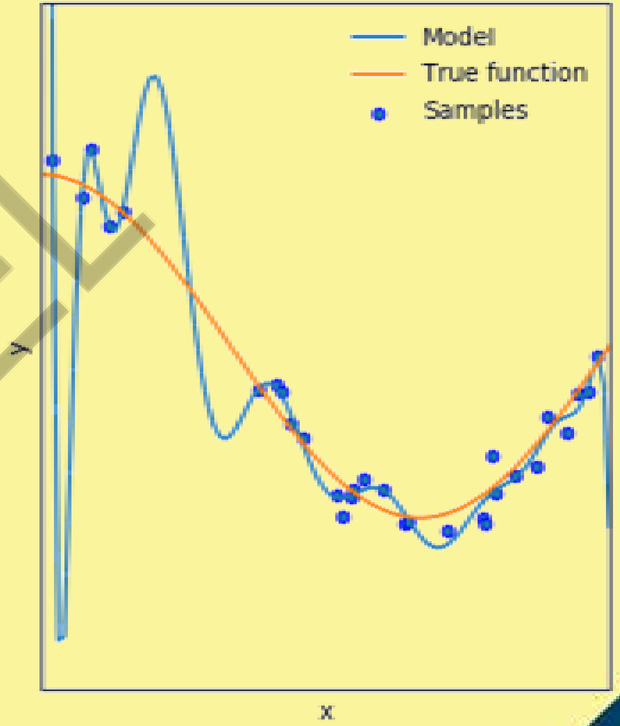
Underfit

Degree: 4



Perfectly fit

Degree: 15



Overfit

Image Source: Internet

Overfitting/Underfitting: Classification

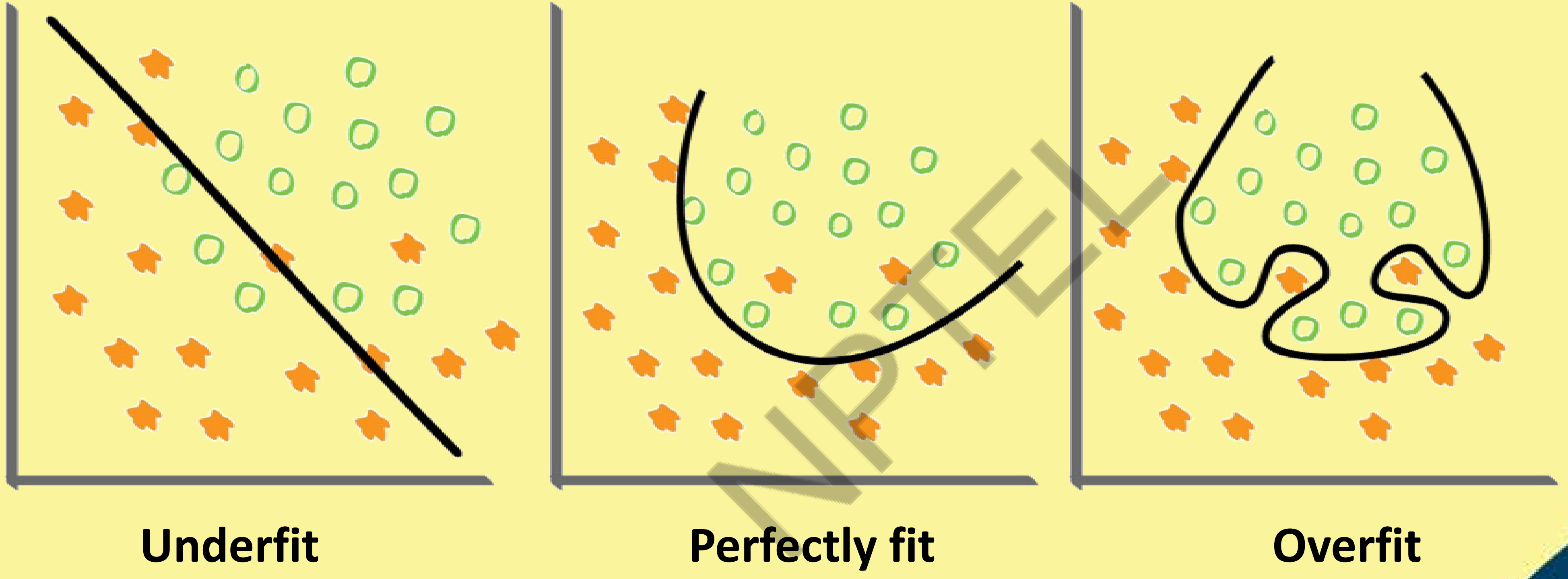


Image Source: Internet

Regularization

- Regularization is a way to prevent overfitting.
- L1 and L2 are the most common types of regularization used in training deep models.
- General cost function with regularization for training is defined as: Cost function = Loss + Regularization term
- Due to this regularization term, the numerical values of weights decrease because it assumes that a neural network with smaller weights leads to simpler models.
- So this helps to reduce overfitting.

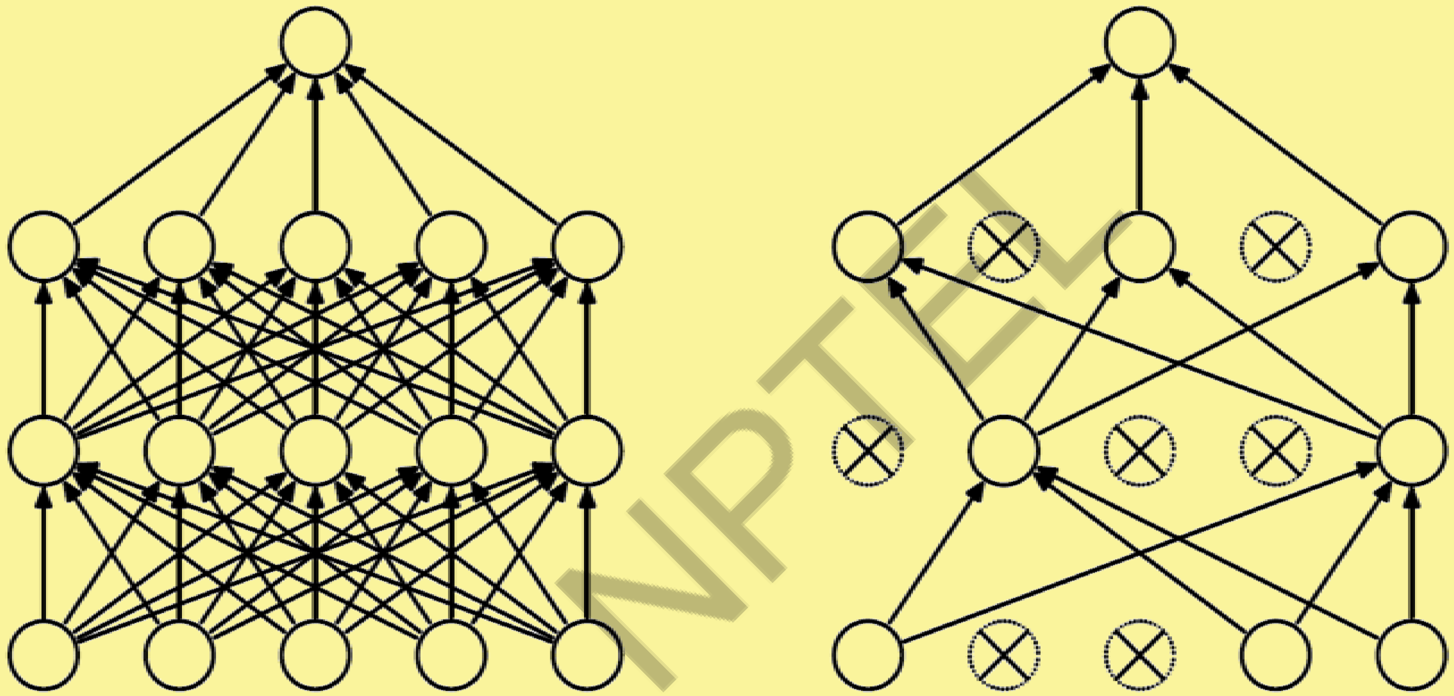
Regularization: L1 & L2

- L1 regularizer: Cost function = Loss + $\lambda \sum |w|$
  - It penalizes absolute value of weights
  - It can make some weights to zero. So useful for model compression.
  - λ is a regularization hyper parameter. Controls the relative weight.
- L2 regularizer: Cost function = Loss + $\lambda \sum ||w||^2$
  - It penalizes second norm of weights.
  - It is also termed as weight decay as it pushes the weights near to zero. But it does not make exactly zero always.

Data Augmentation

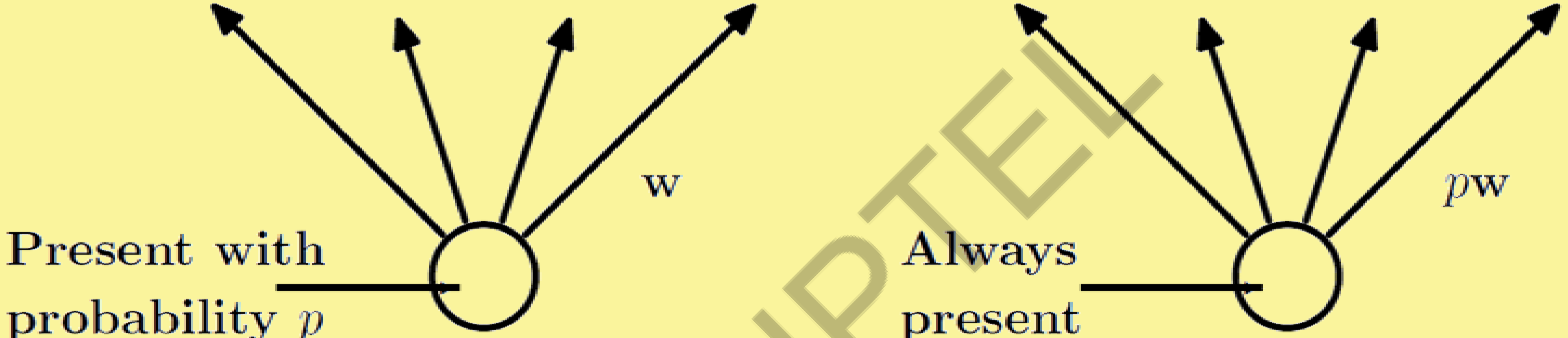
- Increasing the size of training data is a way to prevent overfitting.
- It is difficult and costly to increase the training data.
- Data augmentation is a way to create a different image from one image while keeping the context same.
- There are a few ways of augmenting training data– rotating, flipping, scaling, shifting, contrast enhancement, brightness control, etc.

Dropout



Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15.1 (2014)

Dropout

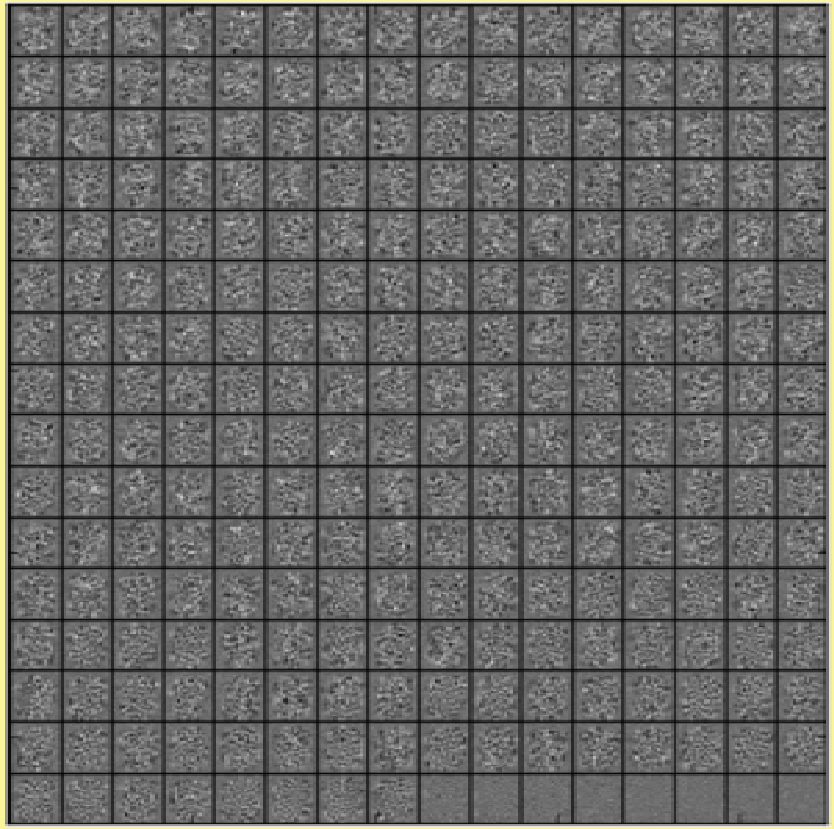


During Training

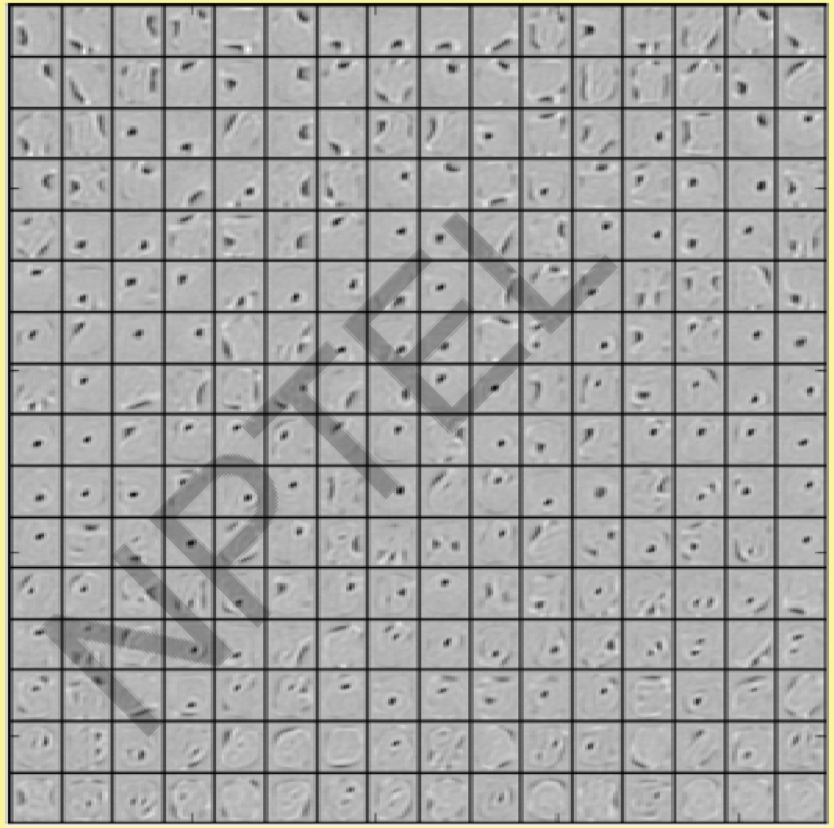
During Testing

Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15.1 (2014)

Dropout: Effect on learned features



Without dropout

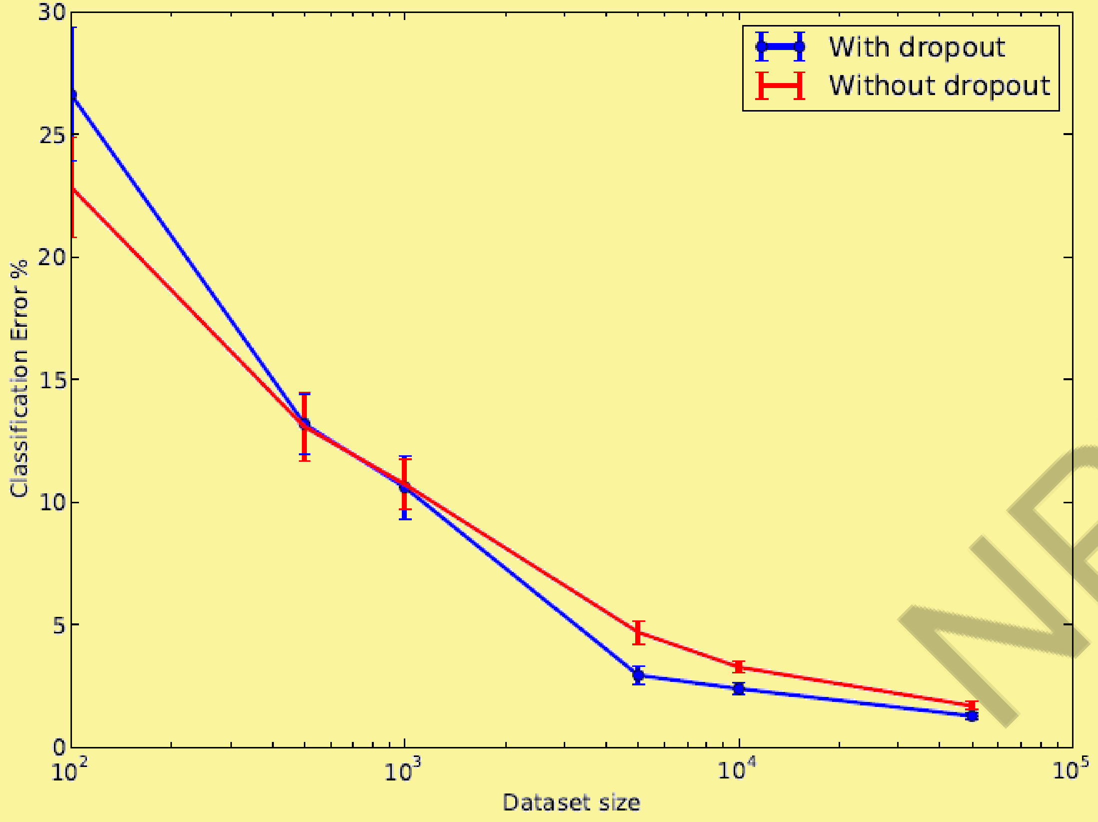


With dropout

Features learned by an autoencoder on MNIST with a single hidden layer of 256 rectified linear units with/ without dropout.

Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15.1 (2014)

Dropout: Effect on Data Size



- While model complexity is fixed, dropout does not generalize the model for very small amount of data
- As the size of the data set is increased, the gain from doing dropout increases up to a point and then declines.
- There is a sweet spot where amount of data is large enough.

Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15.1 (2014)

Early Stopping

- Hyperparameters need to be tuned for good performance while training neural networks.
- Number of iteration is a hyperparameter to be tuned. Lesser iteration may lead to underfit and more iteration may lead to overfit.
- Early stopping attempts to remove the need of manually setting this value.
- It can also be considered a type of regularization method.

Image Source: Internet

Early Stopping

- Hyperparameters need to be tuned for good performance while training neural networks.
- Number of iteration is a hyperparameter to be tuned. Lesser iteration may lead to underfit and more iteration may lead to overfit.
- Early stopping attempts to remove the need of manually setting this value.
- It can also be considered a type of regularization method.

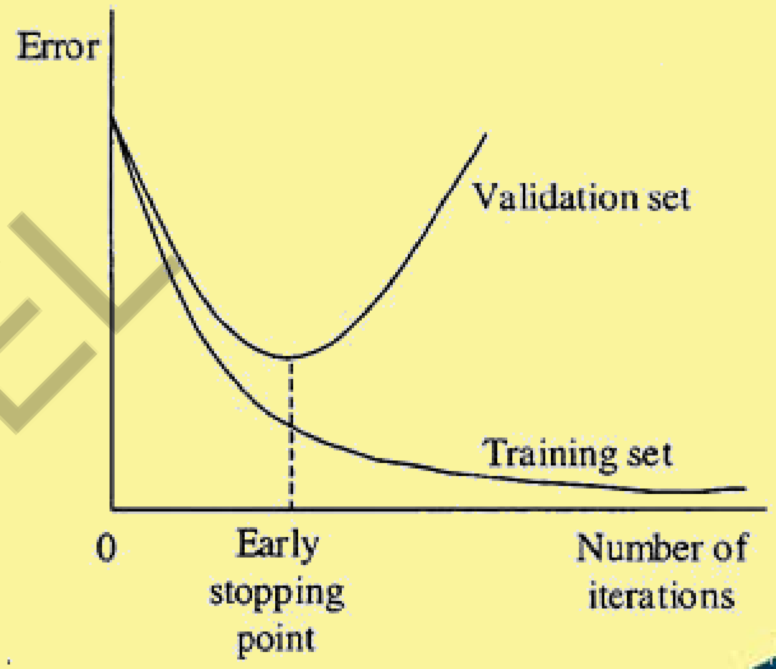


Image Source: Internet

Early Stopping

Early stopping algorithm is as follows:

- ❑ Split data into train, validation and test set
- ❑ After each training epoch:
  - ❑ Evaluate the model performance using validation data
  - ❑ Save the best model evaluated on validation data
- ❑ Use final model that has the best validation performance for testing.

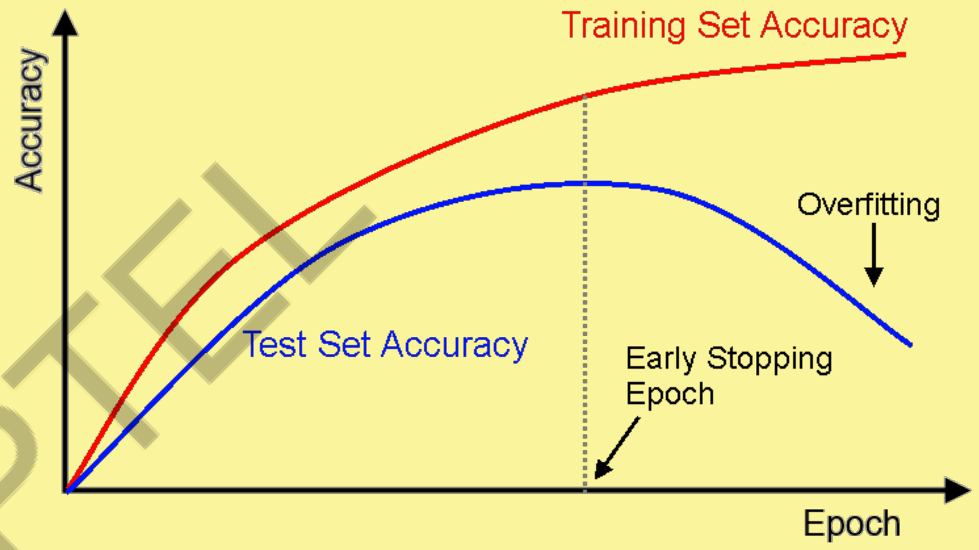


Image Source: Internet

NPTEL ONLINE CERTIFICATION COURSES

Thank you