Wed Jun 12th 06:30 -- 09:00 PM Room Pacific Ballroom

Poster

The University of Hong Kong The Chinese University of Hong Kong SenseTime Research

Differentiable Dynamic Normalization for Learning Deep Representation

Ping Luo*[1 2]

Zhanglin Peng*[3]

Wenqi Shao[2 3]

Ruimao Zhang[2 3]

Jiamin Ren[3]

Lingyun Wu[3]

[1]The University of Hong Kong [2]The Chinese University of Hong Kong [3]SenseTime Research

What is Dynamic Normalization (DN)?

1. DN adapts to various networks, tasks, and batch sizes.
2. DN can be easily implemented and trained in a **Differentiable** end-to-end manner with merely **small number of parameters**, by replacing the original normalizers.
3. DN has **matrix formulation**, representing a wide range of normalization methods (e.g. GroupNorm with any numbers of groups), shedding light on **analyzing them theoretically**.

What is Dynamic Normalization (DN)?

1. DN adapts to various networks, tasks, and batch sizes.
2. DN can be easily implemented and trained in a **Differentiable** end-to-end manner with merely **small number of parameters**, by replacing the original normalizers.
3. DN has **matrix formulation**, representing a wide range of normalization methods (e.g. GroupNorm with any numbers of groups), shedding light on **analyzing them theoretically**.

What is Dynamic Normalization (DN)?

1. DN adapts to various networks, tasks, and batch sizes.
2. DN can be easily implemented and trained in a **Differentiable** end-to-end manner with merely **small number of parameters**, by replacing the original normalizers.
3. DN has **matrix formulation**, representing a wide range of normalization methods (e.g. GroupNorm with any numbers of groups), shedding light on **analyzing them theoretically**.

What is Dynamic Normalization (DN)?

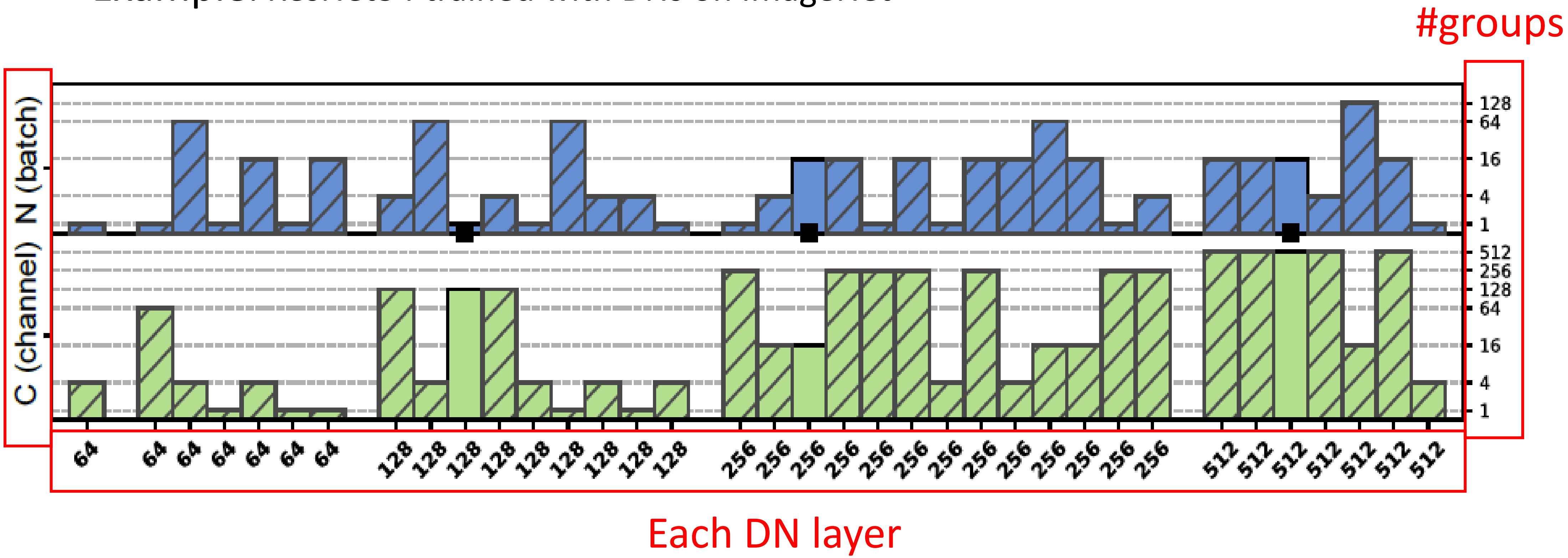
1. DN adapts to various networks, tasks, and batch sizes.

2. DN can be easily implemented and trained in a **Differentiable** end-to-end manner with merely **small number of parameters**, by replacing the original normalizers.

3. DN has **matrix formulation**, representing a wide range of normalization methods (e.g. GroupNorm with any numbers of groups), shedding light on **analyzing them theoretically**.

Dynamic Normalization (DN)

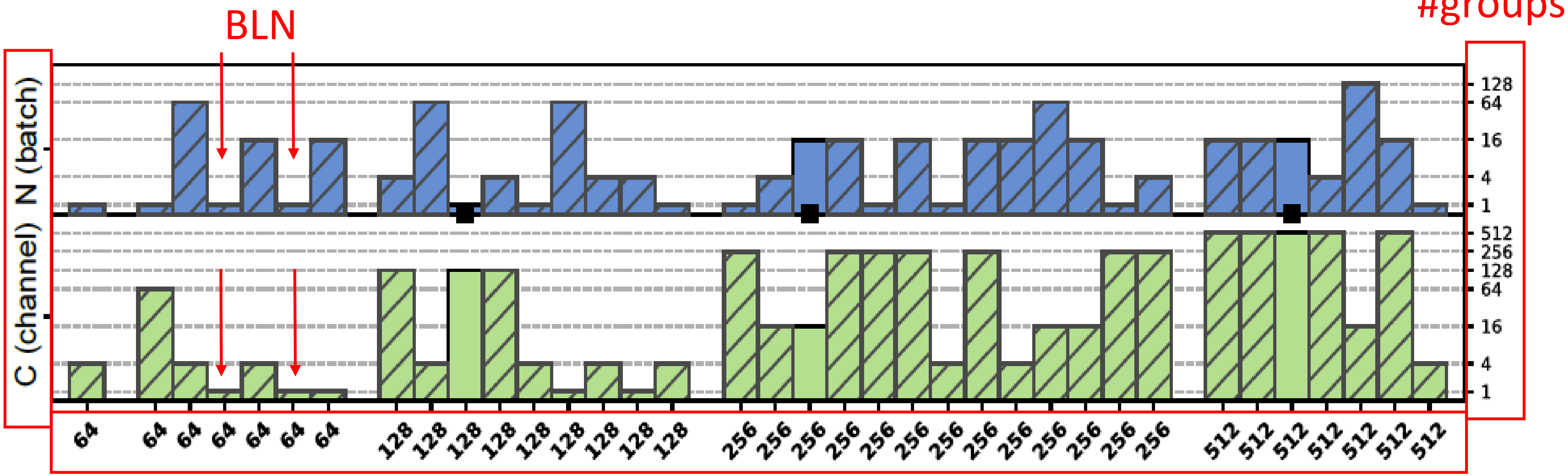
- Example: ResNet34 trained with DNs on ImageNet



The University of Hong Kong
The Chinese University of Hong Kong
SenseTime Research

Dynamic Normalization (DN)

- Example: ResNet34 trained with DNs on ImageNet



Each DN layer

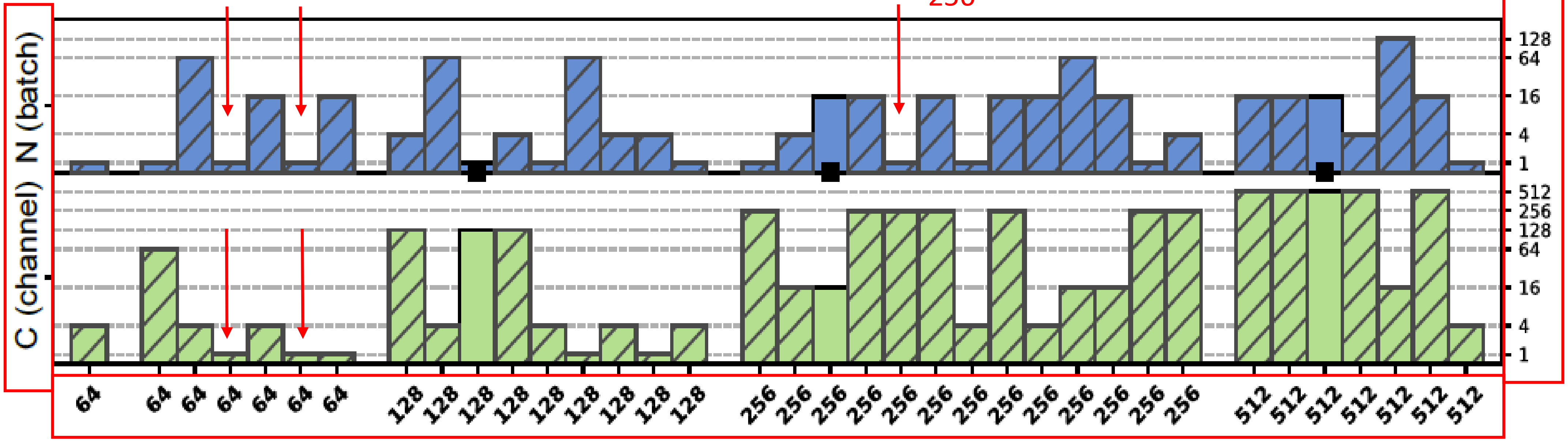
The University of Hong Kong

The Chinese University of Hong Kong

SenseTime Research

Dynamic Normalization (DN)

- Example: ResNet34 trained with DNs on ImageNet



BLN

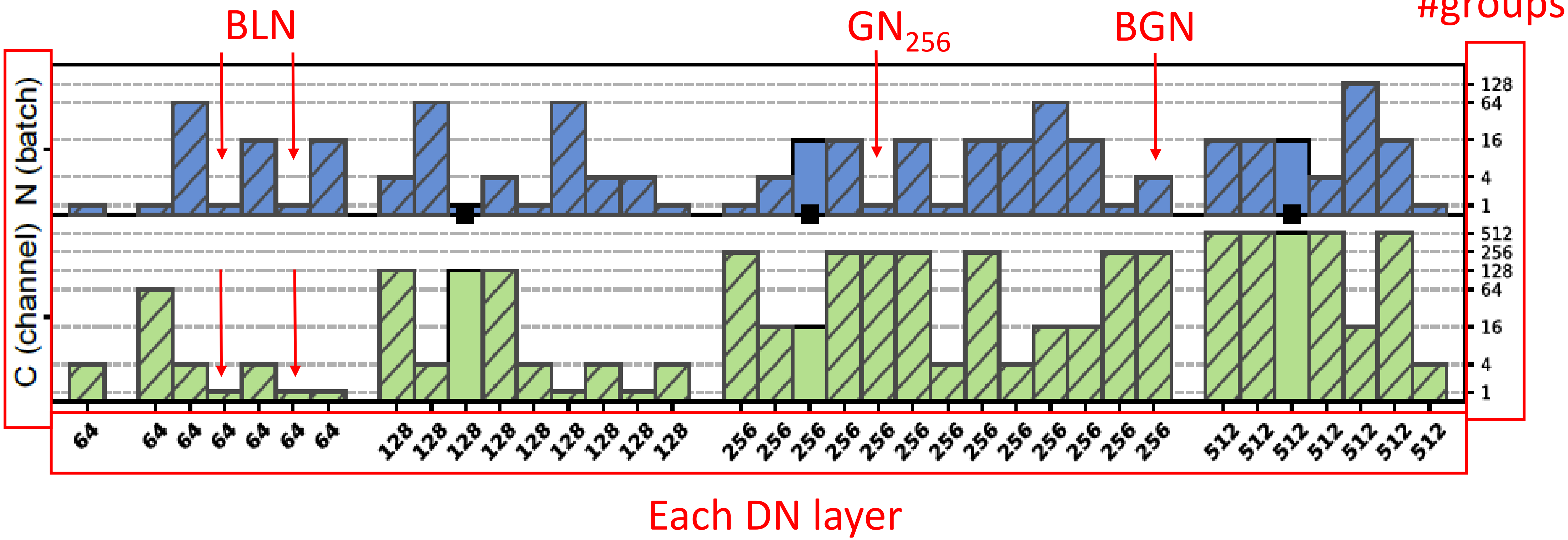
GN_{256}

#groups

Each DN layer

Dynamic Normalization (DN)

- Example: ResNet34 trained with DNs on ImageNet



The University of Hong Kong
The Chinese University of Hong Kong
SenseTime Research

A General Form *vs.* Switchable Normalization (SN)

A General Normalization Form

- Remove means and reduce by variance

normalized feature map

feature map

mean

$$\hat{h} = \frac{h - \mu^k}{\sqrt{(\sigma^k)^2 + \epsilon}}$$

standard deviation

Switchable Normalization: Discrete Learning-to-Normalize

- Learn a linear combination of Batch Norm, Instance Norm, Layer Norm and Group Norm

importance ratio, sum to 1

$$\hat{h} = \frac{h - \sum_{k \in \{\mathrm{BN}, \mathrm{IN}, \mathrm{LN}, \mathrm{GN}, \ldots\}} \lambda^k \mu^k}{\sqrt{\sum_{k \in \{\mathrm{BN}, \mathrm{IN}, \mathrm{LN}, \mathrm{GN}, \ldots\}} \lambda^k (\sigma^k)^2 + \epsilon}}$$

- **Problem: enumerate a large pool of candidate normalizers**

The University of Hong Kong
The Chinese University of Hong Kong
商汤 sensetime
SenseTime Research

A General Form *vs.* Switchable Normalization (SN)

A General Normalization Form

- Remove means and reduce by variance

normalized feature map

feature map

mean

$$\hat{h} = \frac{h - \boxed{\mu^k}}{\sqrt{\boxed{(\sigma^k})^2 + \epsilon}}$$

standard deviation

Switchable Normalization: Discrete Learning-to-Normalize

- Learn a linear combination of Batch Norm, Instance Norm, Layer Norm and Group Norm

importance ratio, sum to 1

$$\hat{h} = \frac{h - \sum_{k\in\{\mathrm{BN},\mathrm{IN},\mathrm{LN},\mathrm{GN},\ldots\}} \lambda^k \mu^k}{\sqrt{\sum_{k\in\{\mathrm{BN},\mathrm{IN},\mathrm{LN},\mathrm{GN},\ldots\}} \lambda^k (\sigma^k)^2 + \epsilon}}$$

- **Problem: enumerate a large pool of candidate normalizers**

The University of Hong Kong

The Chinese University of Hong Kong

商汤 sensetime

SenseTime Research

A General Form *vs.* Switchable Normalization (SN)

A General Normalization Form

- Remove means and reduce by variance

normalized feature map

feature map

mean

$$\boxed{\hat{h}} = \frac{\boxed{h} - \mu^k}{\sqrt{(\sigma^k)^2 + \epsilon}}$$

standard deviation

Switchable Normalization: Discrete Learning-to-Normalize

- Learn a linear combination of Batch Norm, Instance Norm, Layer Norm and Group Norm

importance ratio, sum to 1

$$\hat{h} = \frac{h - \sum_{k \in \{\mathrm{BN}, \mathrm{IN}, \mathrm{LN}, \mathrm{GN}, \ldots\}} \lambda^k \mu^k}{\sqrt{\sum_{k \in \{\mathrm{BN}, \mathrm{IN}, \mathrm{LN}, \mathrm{GN}, \ldots\}} \lambda^k (\sigma^k)^2 + \epsilon}}$$

- **Problem: enumerate a large pool of candidate normalizers**

The University of Hong Kong

The Chinese University of Hong Kong

商汤 sensetime

SenseTime Research

A General Form *vs.* Switchable Normalization (SN)

A General Normalization Form

- Remove means and reduce by variance

normalized feature map

feature map

mean

$$\hat{h} = \frac{h - \mu^k}{\sqrt{(\sigma^k)^2 + \epsilon}}$$

standard deviation

Switchable Normalization: Discrete Learning-to-Normalize

- Learn a linear combination of Batch Norm, Instance Norm, Layer Norm and Group Norm

importance ratio, sum to 1

$$\hat{h} = \frac{h - \sum_{k \in \{\mathrm{BN},\mathrm{IN},\mathrm{LN},\mathrm{GN},\ldots\}} \boxed{\lambda^k} \mu^k}{\sqrt{\sum_{k \in \{\mathrm{BN},\mathrm{IN},\mathrm{LN},\mathrm{GN},\ldots\}} \lambda^k (\sigma^k)^2 + \epsilon}}$$

- **Problem: enumerate a large pool of candidate normalizers**

The University of Hong Kong

The Chinese University of Hong Kong

商汤 sensetime

SenseTime Research

A General Form *vs.* Switchable Normalization (SN)

A General Normalization Form

- Remove means and reduce by variance

normalized feature map

feature map

mean

$$\hat{h} = \frac{h - \mu^k}{\sqrt{(\sigma^k)^2 + \epsilon}}$$

standard deviation

Switchable Normalization: Discrete Learning-to-Normalize

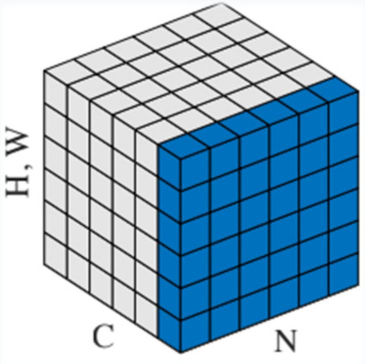
- Learn a linear combination of Batch Norm, Instance Norm, Layer Norm and Group Norm

importance ratio, sum to 1

$$\hat{h} = \frac{h - \sum_{k \in \{\mathrm{BN}, \mathrm{IN}, \mathrm{LN}, \mathrm{GN}, \dots\}} \lambda^k \mu^k}{\sqrt{\sum_{k \in \{\mathrm{BN}, \mathrm{IN}, \mathrm{LN}, \mathrm{GN}, \dots\}} \lambda^k (\sigma^k)^2 + \epsilon}}$$

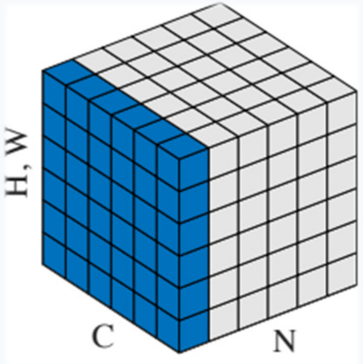
- **Problem: enumerate a large pool of candidate normalizers**

Dynamic Normalization (DN): Continuous Learning-to-Normalize



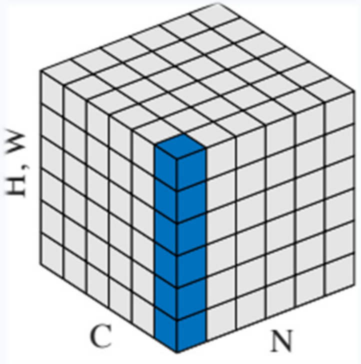
average pixels: {H,W,N}

(a) Batch Norm (BN)



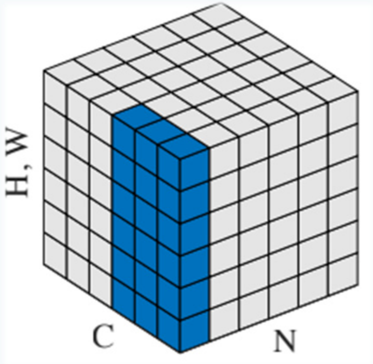
{H,W,C}

(b) Layer Norm (LN)



{H,W}

(c) Instance Norm (IN)



{H,W,C/2}

(d) Group Norm (GN)

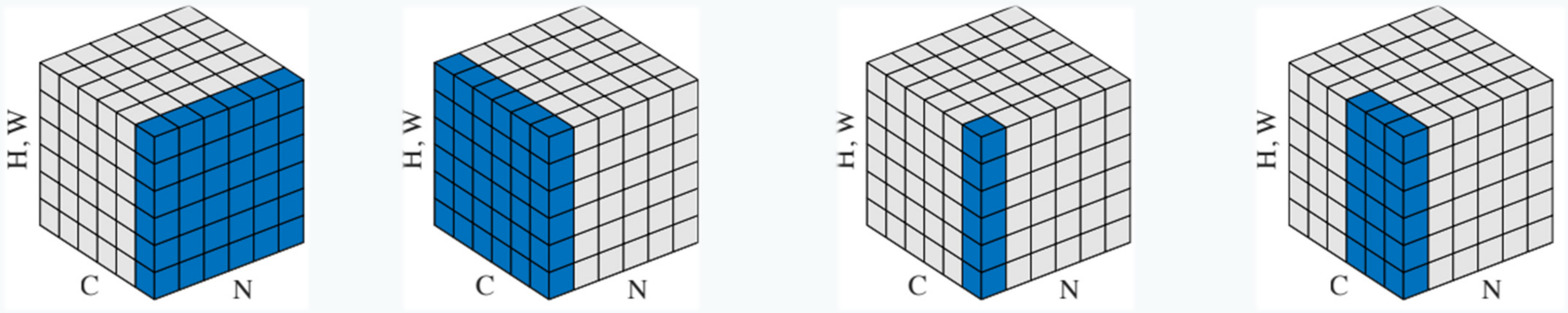
The University of Hong Kong

The Chinese University of Hong Kong

商汤 sensetime

SenseTime Research

Dynamic Normalization (DN): Continuous Learning-to-Normalize



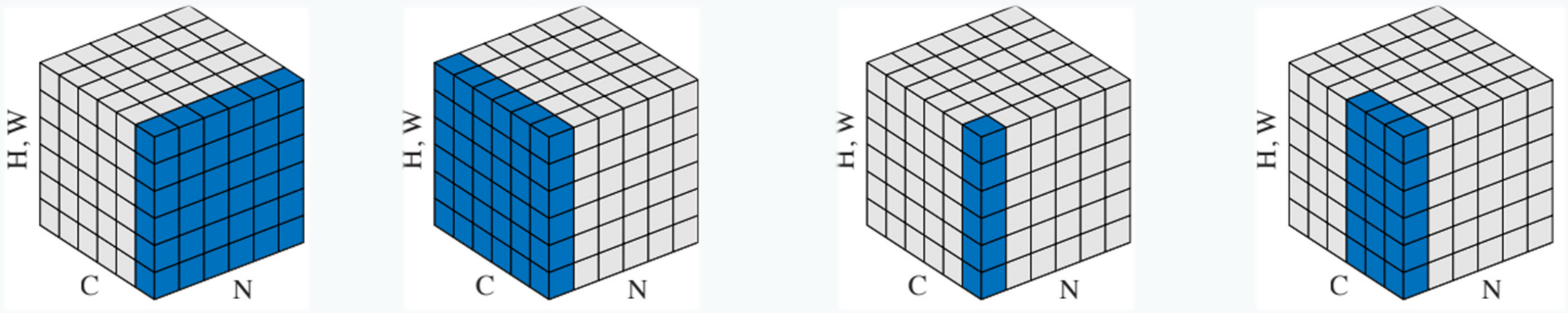
average pixels: {H,W,N} {H,W,C} {H,W} {H,W,C/2}

(a) Batch Norm (BN) **(b) Layer Norm (LN)** **(c) Instance Norm (IN)** **(d) Group Norm (GN)**

Dynamic Normalization: $$\hat{\boldsymbol{h}} = \gamma \frac{\boldsymbol{h} - \boldsymbol{U}\boldsymbol{\mu}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{\sigma}\boldsymbol{V}} + \beta$$

- $\boldsymbol{U} \in \mathbb{R}^{N\times N}, \boldsymbol{V} \in \mathbb{R}^{C\times C}$: two binary diagonal-block matrices
- $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{N\times C}$: means and stds of Instance Normalization (IN), implying that we learn to combine statistics of IN

Dynamic Normalization (DN): Continuous Learning-to-Normalize



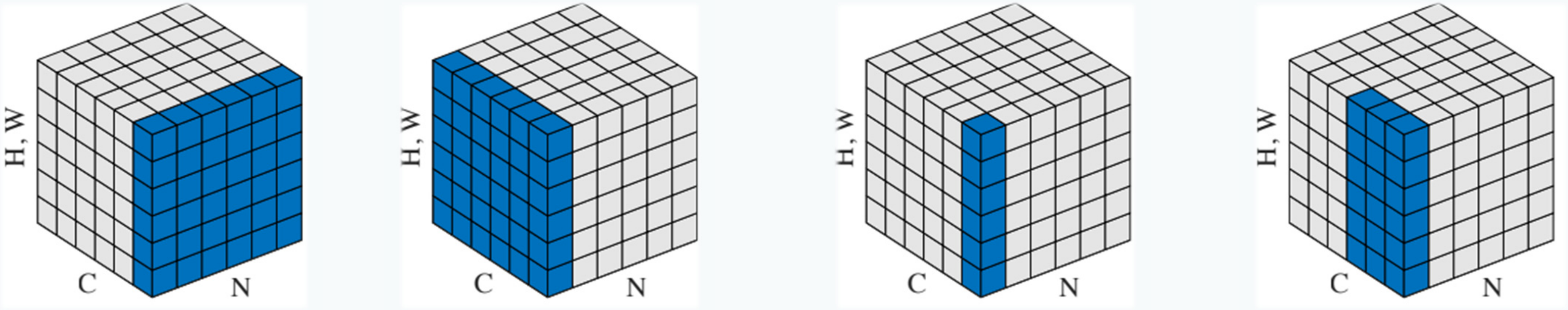
average pixels: {H,W,N} {H,W,C} {H,W} {H,W,C/2}

(a) Batch Norm (BN) (b) Layer Norm (LN) (c) Instance Norm (IN) (d) Group Norm (GN)

Dynamic Normalization: $$\hat{\boldsymbol{h}} = \gamma \frac{\boldsymbol{h} - \boldsymbol{U}\boldsymbol{\mu}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{\sigma}\boldsymbol{V}} + \beta$$

- $\boldsymbol{U} \in \mathbb{R}^{N\times N}, \boldsymbol{V} \in \mathbb{R}^{C\times C}$: two binary diagonal-block matrices
- $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{N\times C}$: means and stds of Instance Normalization (IN), implying that we learn to combine statistics of IN

Dynamic Normalization (DN): Continuous Learning-to-Normalize



average pixels: {H,W,N} {H,W,C} {H,W} {H,W,C/2}

(a) Batch Norm (BN) **(b) Layer Norm (LN)** **(c) Instance Norm (IN)** **(d) Group Norm (GN)**

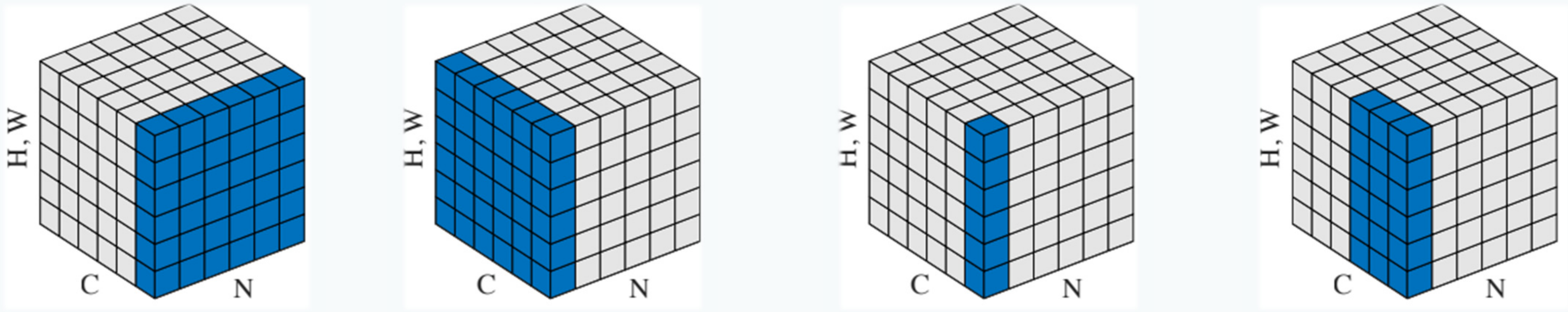
Dynamic Normalization:
$$\widehat{\boldsymbol{h}} = \gamma \frac{\boldsymbol{h} - \boldsymbol{U}\boldsymbol{\mu}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{\sigma}\boldsymbol{V}} + \beta$$

- $\boldsymbol{U} \in \mathbb{R}^{N\times N}, \boldsymbol{V} \in \mathbb{R}^{C\times C}$: two binary diagonal-block matrices
- $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{N\times C}$: means and stds of Instance Normalization (IN), implying that we learn to combine statistics of IN

The University of Hong Kong
The Chinese University of Hong Kong
商汤 sensetime
SenseTime Research

Dynamic Normalization (DN): Continuous Learning-to-Normalize

$\boldsymbol{U}=\mathbf{1}, \boldsymbol{V}=\boldsymbol{I}$ $\quad$ $\boldsymbol{U}=\boldsymbol{I}, \boldsymbol{V}=\mathbf{1}$ $\quad$ $\boldsymbol{U}=\boldsymbol{I}, \boldsymbol{V}=\boldsymbol{I}$ $\quad$ $\boldsymbol{U}=\boldsymbol{I}, \boldsymbol{V}$ is block-diagonal



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

average pixels: {H,W,N} {H,W,C} {H,W} {H,W,C/2}

(a) Batch Norm (BN) **(b) Layer Norm (LN)** **(c) Instance Norm (IN)** **(d) Group Norm (GN)**

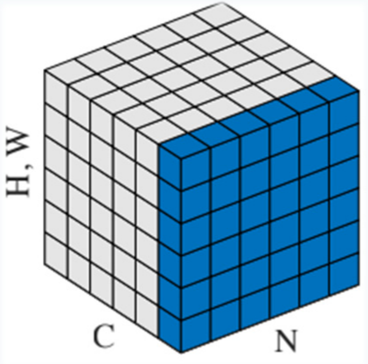
Dynamic Normalization: $$\widehat{\boldsymbol{h}}=\gamma\frac{\boldsymbol{h}-\boldsymbol{U}\boldsymbol{\mu}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{\sigma}\boldsymbol{V}}+\beta$$

- $\boldsymbol{U}\in\mathbb{R}^{N\times N}, \boldsymbol{V}\in\mathbb{R}^{C\times C}$: two binary diagonal-block matrices
- $\boldsymbol{\mu}, \boldsymbol{\sigma}\in\mathbb{R}^{N\times C}$: means and stds of Instance Normalization (IN), implying that we learn to combine statistics of IN

The University of Hong Kong
The Chinese University of Hong Kong
商汤 sensetime
SenseTime Research

Dynamic Normalization (DN): Continuous Learning-to-Normalize

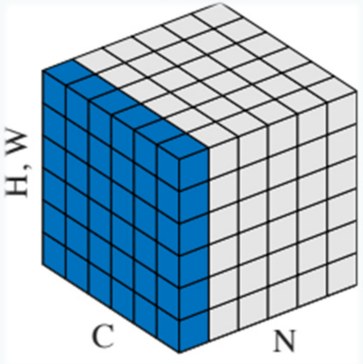
$\boldsymbol{U} = \mathbf{1}, \boldsymbol{V} = \boldsymbol{I}$



average pixels: {H,W,N}

(a) Batch Norm (BN)

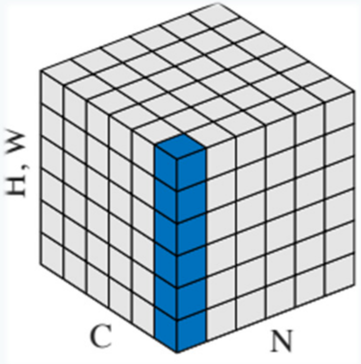
$\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V} = \mathbf{1}$



{H,W,C}

(b) Layer Norm (LN)

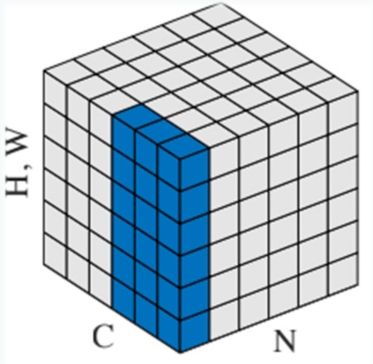
$\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V} = \boldsymbol{I}$



{H,W}

(c) Instance Norm (IN)

$\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V}$ is block-diagonal



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

{H,W,C/2}

(d) Group Norm (GN)

Dynamic Normalization: $$\hat{\boldsymbol{h}} = \gamma \frac{\boldsymbol{h} - \boldsymbol{U\mu V}}{\boldsymbol{U\sigma V}} + \beta$$

- $\boldsymbol{U} \in \mathbb{R}^{N\times N}, \boldsymbol{V} \in \mathbb{R}^{C\times C}$: two binary diagonal-block matrices
- $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{N\times C}$: means and stds of Instance Normalization (IN), implying that we learn to combine statistics of IN

The University of Hong Kong

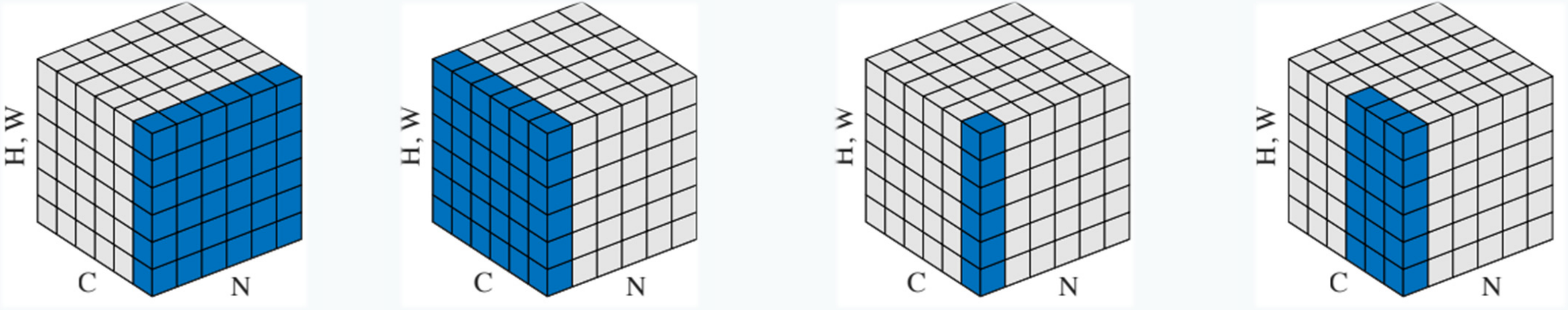
The Chinese University of Hong Kong

商汤 sensetime

SenseTime Research

Dynamic Normalization (DN): Continuous Learning-to-Normalize

$\boldsymbol{U} = \mathbf{1}, \boldsymbol{V} = \boldsymbol{I}$ $\boxed{\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V} = \mathbf{1}}$ $\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V} = \boldsymbol{I}$ $\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V}$ is block-diagonal



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

average pixels: {H,W,N} {H,W,C} {H,W} {H,W,C/2}

(a) Batch Norm (BN) **(b) Layer Norm (LN)** **(c) Instance Norm (IN)** **(d) Group Norm (GN)**

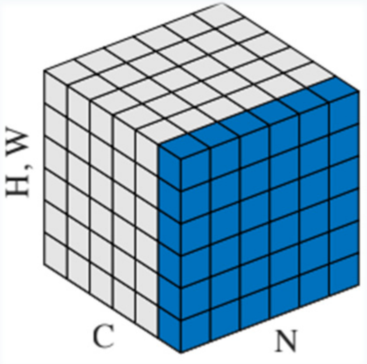
Dynamic Normalization: $$\widehat{\boldsymbol{h}} = \gamma \frac{\boldsymbol{h} - \boldsymbol{U\mu V}}{\boldsymbol{U\sigma V}} + \beta$$

- $\boldsymbol{U} \in \mathbb{R}^{N\times N}, \boldsymbol{V} \in \mathbb{R}^{C\times C}$: two binary diagonal-block matrices
- $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{N\times C}$: means and stds of Instance Normalization (IN), implying that we learn to combine statistics of IN

The University of Hong Kong
The Chinese University of Hong Kong
SenseTime Research

Dynamic Normalization (DN): Continuous Learning-to-Normalize

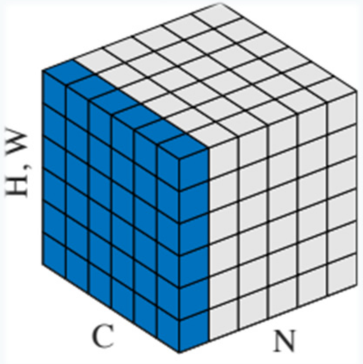
$\boldsymbol{U} = \boldsymbol{1}, \boldsymbol{V} = \boldsymbol{I}$



average pixels: {H,W,N}

(a) Batch Norm (BN)

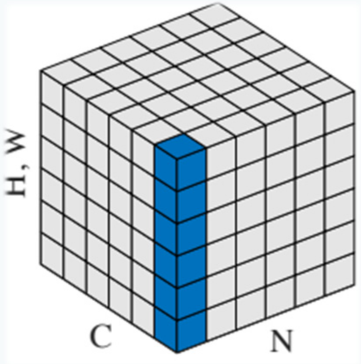
$\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V} = \boldsymbol{1}$



{H,W,C}

(b) Layer Norm (LN)

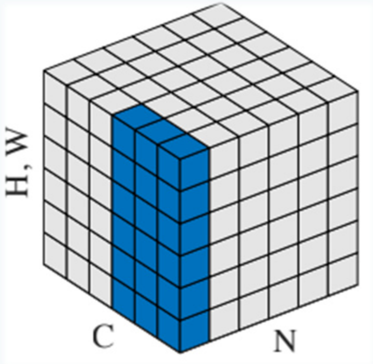
$\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V} = \boldsymbol{I}$



{H,W}

(c) Instance Norm (IN)

$\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V}$ is block-diagonal



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

{H,W,C/2}

(d) Group Norm (GN)

Dynamic Normalization: $$\hat{\boldsymbol{h}} = \gamma \frac{\boldsymbol{h} - \boldsymbol{U}\boldsymbol{\mu}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{\sigma}\boldsymbol{V}} + \beta$$

- $\boldsymbol{U} \in \mathbb{R}^{N\times N}, \boldsymbol{V} \in \mathbb{R}^{C\times C}$: two binary diagonal-block matrices
- $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{N\times C}$: means and stds of Instance Normalization (IN), implying that we learn to combine statistics of IN

The University of Hong Kong

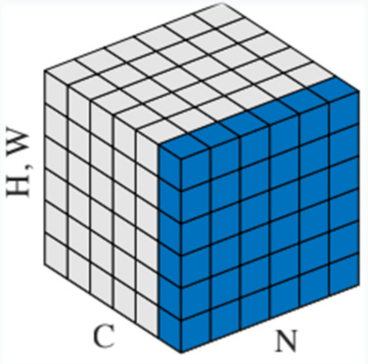
The Chinese University of Hong Kong

商汤 sensetime

SenseTime Research

Dynamic Normalization (DN): Continuous Learning-to-Normalize

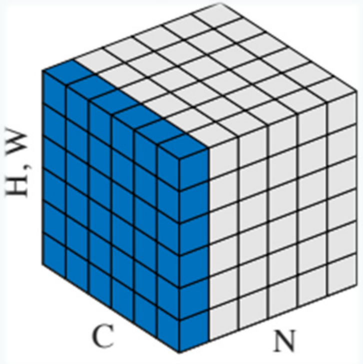
$\boldsymbol{U} = \mathbf{1}, \boldsymbol{V} = \boldsymbol{I}$



average pixels: {H,W,N}

(a) Batch Norm (BN)

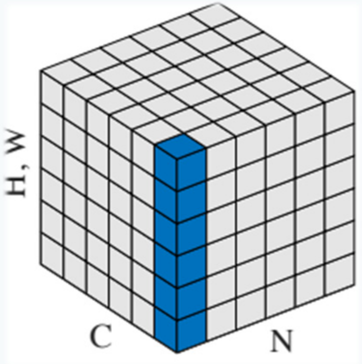
$\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V} = \mathbf{1}$



{H,W,C}

(b) Layer Norm (LN)

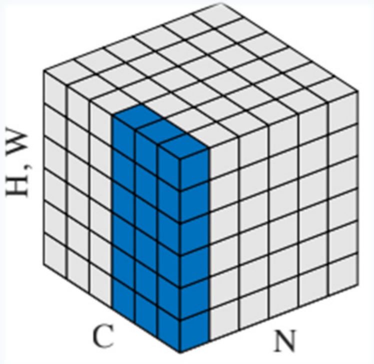
$\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V} = \boldsymbol{I}$



{H,W}

(c) Instance Norm (IN)

$\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V}$ is block-diagonal



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

{H,W,C/2}

(d) Group Norm (GN)

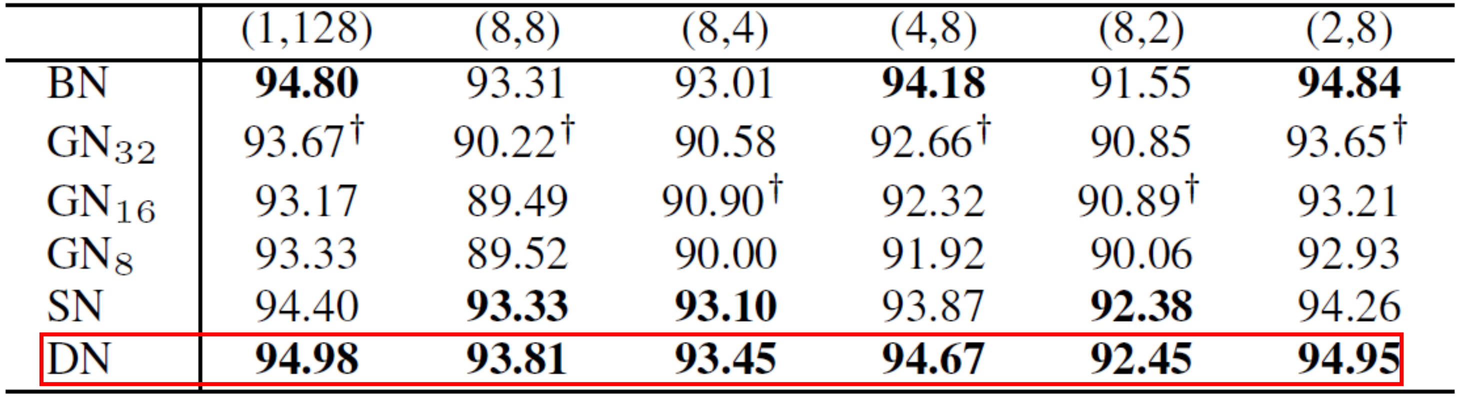
Dynamic Normalization: $$\widehat{\boldsymbol{h}} = \gamma \frac{\boldsymbol{h} - \boldsymbol{U}\boldsymbol{\mu}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{\sigma}\boldsymbol{V}} + \beta$$

- $\boldsymbol{U} \in \mathbb{R}^{N\times N}, \boldsymbol{V} \in \mathbb{R}^{C\times C}$: two binary diagonal-block matrices
- $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{N\times C}$: means and stds of Instance Normalization (IN), implying that we learn to combine statistics of IN

The University of Hong Kong
The Chinese University of Hong Kong
商汤 sensetime
SenseTime Research

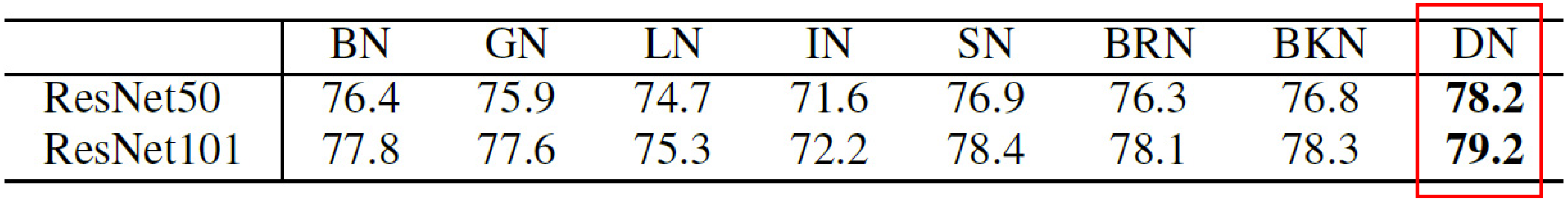
Experimental Results

- ResNet18 on CIFAR10



	(1,128)	(8,8)	(8,4)	(4,8)	(8,2)	(2,8)
BN	**94.80**	93.31	93.01	**94.18**	91.55	**94.84**
GN$_{32}$	93.67†	90.22†	90.58	92.66†	90.85	93.65†
GN$_{16}$	93.17	89.49	90.90†	92.32	90.89†	93.21
GN$_{8}$	93.33	89.52	90.00	91.92	90.06	92.93
SN	94.40	**93.33**	**93.10**	93.87	**92.38**	94.26
DN	**94.98**	**93.81**	**93.45**	**94.67**	**92.45**	**94.95**

- ImageNet



	BN	GN	LN	IN	SN	BRN	BKN	DN
ResNet50	76.4	75.9	74.7	71.6	76.9	76.3	76.8	**78.2**
ResNet101	77.8	77.6	75.3	72.2	78.4	78.1	78.3	**79.2**

The University of Hong Kong

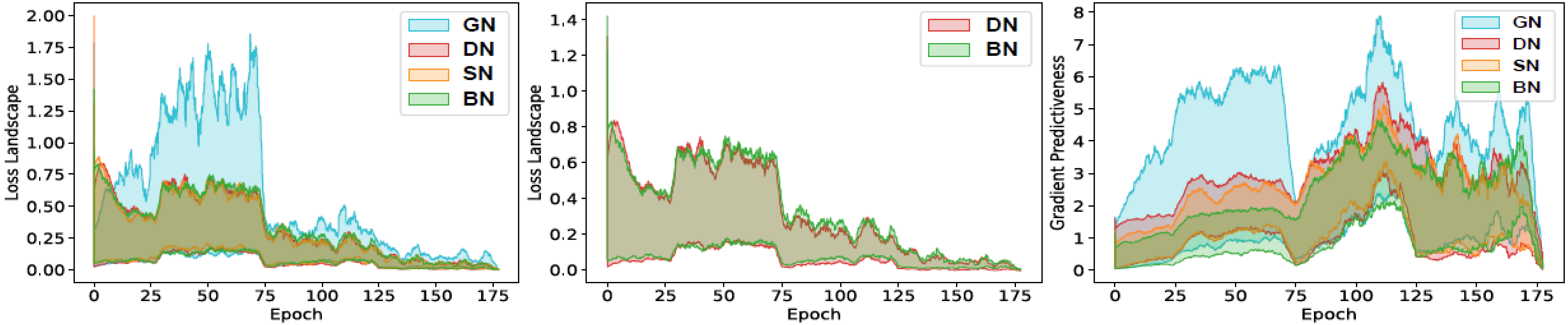
The Chinese University of Hong Kong

商汤 sensetime

SenseTime Research

Comparisons of Loss Landscapes

- ResNet18 on CIFAR10



(a) variations of loss values

(b) variations of losses of DN *v.s.* BN

(c) variations of gradient values

$GN \gg DN \approx SN \approx BN$

$GN \gg DN > SN \approx BN$

Thank You

The University of Hong Kong

The Chinese University of Hong Kong

商汤 sensetime

SenseTime Research

Differentiable Dynamic Normalization for Learning Deep Representation

Wed Jun 12th 06:30 -- 09:00 PM Room Pacific Ballroom

Poster