Group Normalization

ECCV 2018, Munich

Yuxin Wu, Kaiming He
Facebook AI Research (FAIR)
Batch Normalization – a Milestone
What’s Batch Norm

What’s Batch Norm

- Batch ...

- Normalization!

\[ \hat{x} = \frac{x - \mu}{\sigma} \]

Batch: also source of drawbacks

- Small batch
  - large models
  - detection / segmentation / video classification ...
Batch: also source of drawbacks

- Small batch
- Varying batch
Our Method: **Group Normalization**

- GN is **batch-independent**
- **Small batch**
- **Varying batch**
“Group Norm” in Retrospective: HOG/SIFT

Dalal, Navneet, and Bill Triggs. “Histograms of oriented gradients for human detection.” Computer Vision and Pattern Recognition, 2005
“Group Norm” in Retrospective: HOG/SIFT

- Group-wise features

Dalal, Navneet, and Bill Triggs. “Histograms of oriented gradients for human detection.” Computer Vision and Pattern Recognition, 2005
“Group Norm” in Retrospective: HOG/SIFT

- Group-wise features
- Group-wise normalized

Dalal, Navneet, and Bill Triggs. “Histograms of oriented gradients for human detection.” Computer Vision and Pattern Recognition, 2005
What’s Group Norm

$\hat{x} = \frac{x - \mu}{\sigma}$

$H, W$

$C, N$

$\mu, \sigma$
Layer Norm

Group Norm

Instance Norm

Batch Norm

all channels in one group

one channel per group

one image per batch

one channel


Experiments: ImageNet Classification

Standard Batch Size

val error (%), batch=32

epochs

BN: 23.6
GN: 24.1
val error (%), batch=32
<table>
<thead>
<tr>
<th>Number of Groups</th>
<th>Best G=32 val error (%)</th>
<th>LN val error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments: ImageNet Classification

Small Batch Size
val error

epochs

Batch Norm 😞

batch=4
batch=8
batch=16
batch=32
The graphs show the validation error over epochs for different batch sizes. The batch sizes are 32, 16, 8, 4, and 2. The Batch Norm curves match poorly, indicated by the sad face symbol. In contrast, the Group Norm curves match well, as indicated by the happy face symbol.
Batch: also source of drawbacks

- Small batch
- Varying batch
  - train vs. test
  - pre-train vs. fine-tune
  - backbone vs. head
Experiments: Object Detection

What’s a Batch?
What's a Batch?

Mask R-CNN

Mask R-CNN on COCO

Used by 2018 COCO challenge winner
GN Enables Training Mask R-CNN From Scratch

![Bar chart showing comparison of GN with and without pre-training on Res50 and Res101.](chart.png)

- Res50: 40.8 with pre-train, 42.3 without pre-train
- Res101: 42.3 with pre-train
GN Enables Training Mask R-CNN From Scratch

**preliminary results**

<table>
<thead>
<tr>
<th></th>
<th>w/ pre-train</th>
<th>w/o pre-train</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res50</td>
<td>40.8</td>
<td>39.5</td>
</tr>
<tr>
<td>Res101</td>
<td>42.3</td>
<td>41.0</td>
</tr>
</tbody>
</table>

More to come: 50.9 AP w/o pre-training
Conclusion

- normalization matters
- “batch” is not always ideal
- channels can be grouped, and have substructures

Code: https://github.com/facebookresearch/Detectron/tree/master/projects/GN