Proxy-Normalizing Activations to Match Batch Normalization while Removing Batch Dependence

Antoine Labatie Dominic Masters Zach Eaton-Rosen Carlo Luschi
Graphcore Research
antoine.labatie@centraliens.net {dominicm, zacher, carlo}@graphcore.ai

Context

Normalization is crucial in deep learning to successfully scale to large and deep models.

Batch Norm [1] is the most well-established normalization technique in vision tasks:

- Batch Norm performs very well when the batch size is *large enough*.
- Batch Norm performs poorly when the batch size is *too small*. That is because Batch Norm's batch dependence entails an excessive regularization when the batch size is too small.

Though robust to the use of small batch sizes, *batch-independent alternatives* to Batch Norm tend to incur a *performance degradation* compared to Batch Norm.

Contributions

1. We introduce a novel framework to characterize the neural network properties affected by the choice of normalization.
2. Using this framework, we show that Batch Norm's beneficial properties are not retained when solely using the prototypical batch-independent norms [2, 3, 4], but are retained when combining some of these norms with *Proxy Norm*, a novel batch-independent norm that we hereby introduce.
3. We demonstrate experimentally that our novel batch-independent approach, that is based on this combination of norms, consistently matches or exceeds Batch Norm's performance.

Batch Norm

Batch Norm

$$\mathbf{y} = \frac{\mathbf{x} - \hat{\mu}_c(\mathbf{x})}{\hat{\sigma}_c(\mathbf{x})}$$
$$\mathbf{z} = \underbrace{\phi(\boldsymbol{\gamma}\mathbf{y} + \boldsymbol{\beta})}_{\mathrm{Act}(\mathbf{y})}$$

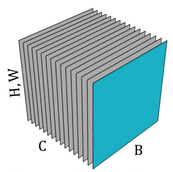


Figure of Batch Norm, adapted from [5]

1st Beneficial Property

✓ Channel-wise normalization:

- Effective use of *width* with equal power in all channels:

$$\mathcal{P}_c(\mathbf{y}) = \mathbb{E}_{b,h,w}[\mathbf{y}^2_{b,h,w,c}] = 1$$

- Effective use of *depth* with ϕ effectively channel-wise nonlinear. That means the channel-wise linear best-fit $\tilde{\mathbf{y}}$ of the post-activations $\mathbf{z}$ using $\mathbf{y}$ is such that

$$\mathcal{P}_c(\mathbf{z} - \tilde{\mathbf{y}}) \ll \mathcal{P}_c(\mathbf{z})$$

2nd Beneficial Property

✓ Preservation of expressivity:

- Suppose mini-batch statistics $\hat{\mu}_c(\mathbf{x})$, $\hat{\sigma}_c(\mathbf{x})$ approximate well full-batch statistics $\mu_c(\mathbf{x})$, $\sigma_c(\mathbf{x})$
- Then, Batch Norm's effect can be "undone" into the succeeding affine transform by choosing $\boldsymbol{\gamma}_c = \sigma_c(\mathbf{x})$ and $\boldsymbol{\beta}_c = \mu_c(\mathbf{x})$
- Conversely, Batch Norm's effect can be "absorbed" into any preceding convolution with bias

Detrimental Property

✗ Batch dependence:

- The dependence of mini-batch statistics $\hat{\mu}_c(\mathbf{x})$, $\hat{\sigma}_c(\mathbf{x})$ on the random choice of inputs in the mini-batch leads to a stochasticity of $\mathbf{y}$
- This stochasticity leads to an *inherent regularization* of Batch Norm
- When the batch size is too small, this regularization becomes excessive and the performance becomes degraded

Prototypical Batch-Independent Norms

Layer Norm

$$\mathbf{y} = \frac{\mathbf{x} - \mu_b(\mathbf{x})}{\sigma_b(\mathbf{x})}$$
$$\mathbf{z} = \phi(\boldsymbol{\gamma}\mathbf{y} + \boldsymbol{\beta})$$

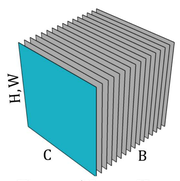


Figure of Layer Norm, adapted from [5]

✗ Strong channel-wise denormalization
✓ Very weak alteration of expressivity
✓ No batch dependence

Instance Norm

$$\mathbf{y} = \frac{\mathbf{x} - \mu_{b,c}(\mathbf{x})}{\sigma_{b,c}(\mathbf{x})}$$
$$\mathbf{z} = \phi(\boldsymbol{\gamma}\mathbf{y} + \boldsymbol{\beta})$$

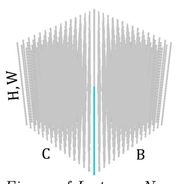


Figure of Instance Norm, adapted from [5]

✓ No channel-wise denormalization
✗ Strong alteration of expressivity
✓ No batch dependence

Group Norm

$$\mathbf{y} = \frac{\mathbf{x} - \mu_{b,c \bmod G}(\mathbf{x})}{\sigma_{b,c \bmod G}(\mathbf{x})}$$
$$\mathbf{z} = \phi(\boldsymbol{\gamma}\mathbf{y} + \boldsymbol{\beta})$$

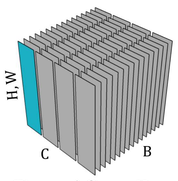


Figure of Group Norm, adapted from [5]

~ Weak channel-wise denormalization
~ Weak alteration of expressivity
✓ No batch dependence

Proxy Norm

Layer Norm + Proxy Norm

$$\mathbf{y} = \frac{\mathbf{x} - \mu_b(\mathbf{x})}{\sigma_b(\mathbf{x})}$$
$$\tilde{\mathbf{z}} = \underbrace{\frac{\phi(\boldsymbol{\gamma}\mathbf{y} + \boldsymbol{\beta}) - \mathbb{E}_{\mathbf{Y}\sim\mathcal{N}(0,1)}[\phi(\boldsymbol{\gamma}\mathbf{Y} + \boldsymbol{\beta})]}{\sqrt{\mathrm{Var}_{\mathbf{Y}\sim\mathcal{N}(0,1)}[\phi(\boldsymbol{\gamma}\mathbf{Y} + \boldsymbol{\beta})] + \epsilon}}}_{\text{PN-Act}(\mathbf{y})}$$

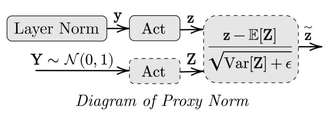


Diagram of Proxy Norm

Rationale for Proxy Norm

Layer Norm does not actively create channel-wise denormalization, it just does not counteract it.

What actively creates channel-wise denormalization is Act, i.e.

1. The affine transform
2. The activation function ϕ

Idea of Proxy Norm: assimilate $\mathbf{y}$ with a Gaussian proxy variable $\mathbf{Y} \sim \mathcal{N}(0, 1)$ and counteract the effect of Act on channel-wise denormalization.

Our Batch-Independent Approach

We adopt a novel approach that *combines Layer Norm, or Group Norm with few groups, and Proxy Norm*. With this approach, we have:

✓ Only a very weak channel-wise denormalization
✓ Only a very weak alteration of expressivity
✓ No batch dependence

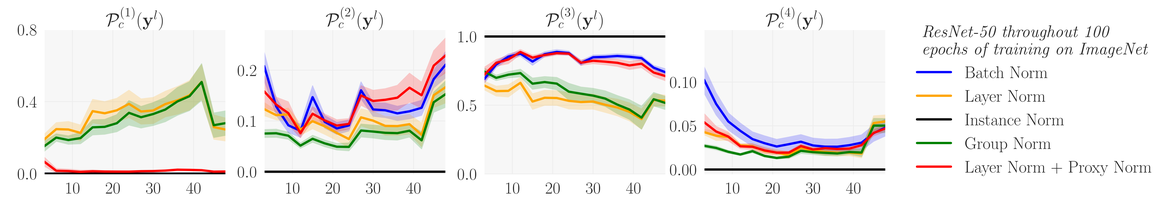
Strengths of this novel batch-independent approach:

1. Channel-wise normalization maintained *throughout training* and not just *at initialization*
2. Wide applicability
3. Ease of implementation

Novel Framework of Characterization of Neural Network Properties

At each layer l, we *decompose* the channel-wise powers $\mathcal{P}_c(\mathbf{y}^l)$ of $\mathbf{y}^l$ into 4 terms which depend on the instance statistics $\mu_{b,c}(\mathbf{y}^l)$ and $\sigma_{b,c}(\mathbf{y}^l)$:

$$\mathcal{P}_c(\mathbf{y}^l) = \underbrace{\mathbb{E}_b[\mu_{b,c}(\mathbf{y}^l)]^2}_{\mathcal{P}_c^{(1)}(\mathbf{y}^l)} + \underbrace{\mathrm{Var}_b[\mu_{b,c}(\mathbf{y}^l)]}_{\mathcal{P}_c^{(2)}(\mathbf{y}^l)} + \underbrace{\mathbb{E}_b[\sigma_{b,c}(\mathbf{y}^l)]^2}_{\mathcal{P}_c^{(3)}(\mathbf{y}^l)} + \underbrace{\mathrm{Var}_b[\sigma_{b,c}(\mathbf{y}^l)]}_{\mathcal{P}_c^{(4)}(\mathbf{y}^l)}.$$



Layer Norm

- Strong channel-wise denormalization characterized by $\mathcal{P}_c(\mathbf{y}^l) - \mathcal{P}_c^{(1)}(\mathbf{y}^l)$ getting smaller and smaller with depth
- Very weak alteration of expressivity characterized by $\mathcal{P}_c^{(2)}(\mathbf{y}^l)$, $\mathcal{P}_c^{(4)}(\mathbf{y}^l)$ having similar values as with Batch Norm

Instance Norm

- No channel-wise denormalization characterized by $\mathcal{P}_c^{(1)}(\mathbf{y}^l)$ equal to 0 at all depths
- Strong alteration of expressivity characterized by $\mathcal{P}_c^{(2)}(\mathbf{y}^l)$, $\mathcal{P}_c^{(4)}(\mathbf{y}^l)$ equal to 0

Group Norm

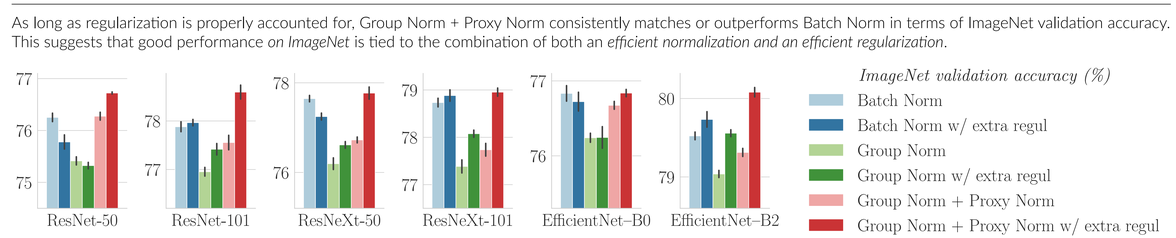
- Weak channel-wise denormalization characterized by $\mathcal{P}_c(\mathbf{y}^l) - \mathcal{P}_c^{(1)}(\mathbf{y}^l)$ getting smaller and smaller with depth
- Weak alteration of expressivity characterized by $\mathcal{P}_c^{(2)}(\mathbf{y}^l)$, $\mathcal{P}_c^{(4)}(\mathbf{y}^l)$ having low values

Our Batch-Independent Approach

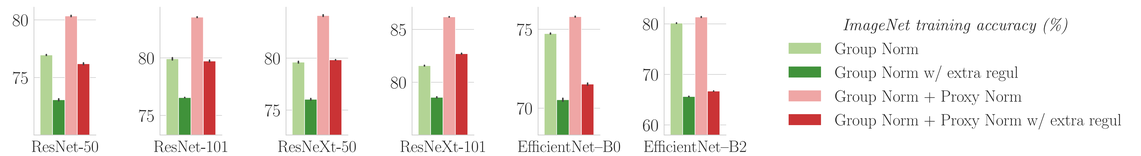
- Very weak channel-wise denormalization characterized by $\mathcal{P}_c^{(1)}(\mathbf{y}^l)$ being close to 0 at all depths
- Very weak alteration of expressivity characterized by $\mathcal{P}_c^{(2)}(\mathbf{y}^l)$, $\mathcal{P}_c^{(4)}(\mathbf{y}^l)$ having similar values as with Batch Norm

Practical Performance of our Batch-Independent Approach

As long as regularization is properly accounted for, Group Norm + Proxy Norm consistently matches or outperforms Batch Norm in terms of ImageNet validation accuracy. This suggests that good performance *on ImageNet* is tied to the combination of both an *efficient normalization and an efficient regularization*.



Group Norm + Proxy Norm consistently outperforms alternative batch-independent approaches in terms of ImageNet training accuracy. This suggests that good performance *on datasets larger than ImageNet* would be tied to an *efficient normalization alone*.



Conclusion

The prototypical batch-independent norms do not retain Batch Norm's beneficial properties:
✗ Layer Norm leads to a strong channel-wise denormalization
✗ Instance Norm leads to a strong alteration of expressivity

In contrast, our novel batch-independent approach does approximately:
✓ Retain channel-wise normalization
✓ Preserve expressivity

Our approach can be used to retain Batch Norm's beneficial properties while at the same time:
1. Alleviating the burden of large activation memory
2. Avoiding Batch Norm's regularization when this regularization is detrimental

References

[1] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ICML 2015*.

[2] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey Everest Hinton. Layer Normalization. *arXiv 2016*.

[3] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv 2016*.

[4] Yuxin Wu and Kaiming He. Group Normalization. *ECCV 2018*.

[5] Neofytos Dimitriou and Ognjen Arandjelovic. A New Look at Ghost Normalization. *arXiv 2020*.