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

NeurIPS 2021

Antoine Labatie

Dominic Masters

Zach Eaton-Rosen

Carlo Luschi

GRAPHCORE

CONTEXT OF THIS WORK

- Normalization is a critical component of deep neural networks to reach optimal performance for a given model size

- In ConvNets, the go-to normalization is Batch Norm:
  - ✓ Batch Norm works very well when the batch size is large enough
  - ✗ Batch Norm's batch dependence leads to degraded performance at small batch size

- Our goal in this work is to design a normalization approach that:
  - ✓ Matches or exceeds the performance of Batch Norm
  - ✓ Is fully batch-independent and thus does not incur degraded performance at small batch size

COMMON NORMALIZATION TECHNIQUES IN CONVNETS

	Norm	Act
Batch Norm	$\mathbf{y}^l = \frac{\mathbf{x}^l - \mu_c(\mathbf{x}^l)}{\sigma_c(\mathbf{x}^l)}$	$\mathbf{z}^l = \phi(\gamma^l \mathbf{y}^l + \beta^l)$
Layer Norm	$\mathbf{y}^l = \frac{\mathbf{x}^l - \mu_x(\mathbf{x}^l)}{\sigma_x(\mathbf{x}^l)}$	
Instance Norm	$\mathbf{y}^l = \frac{\mathbf{x}^l - \mu_{x,c}(\mathbf{x}^l)}{\sigma_{x,c}(\mathbf{x}^l)}$	
Group Norm	$\mathbf{y}^l = \frac{\mathbf{x}^l - \mu_{x,c\in G}(\mathbf{x}^l)}{\sigma_{x,c\in G}(\mathbf{x}^l)}$	

COMMON NORMALIZATION TECHNIQUES IN CONVNETS: PROS AND CONS

		Channel-wise normalization of $\mathbf{y}^l$	Approximate preservation of the network's expressivity	Batch independence
Batch Norm	Act	✓	✓	✗
Layer Norm		✗	✓	✓
Instance Norm		✓	✗	✓
Group Norm		~	~	✓

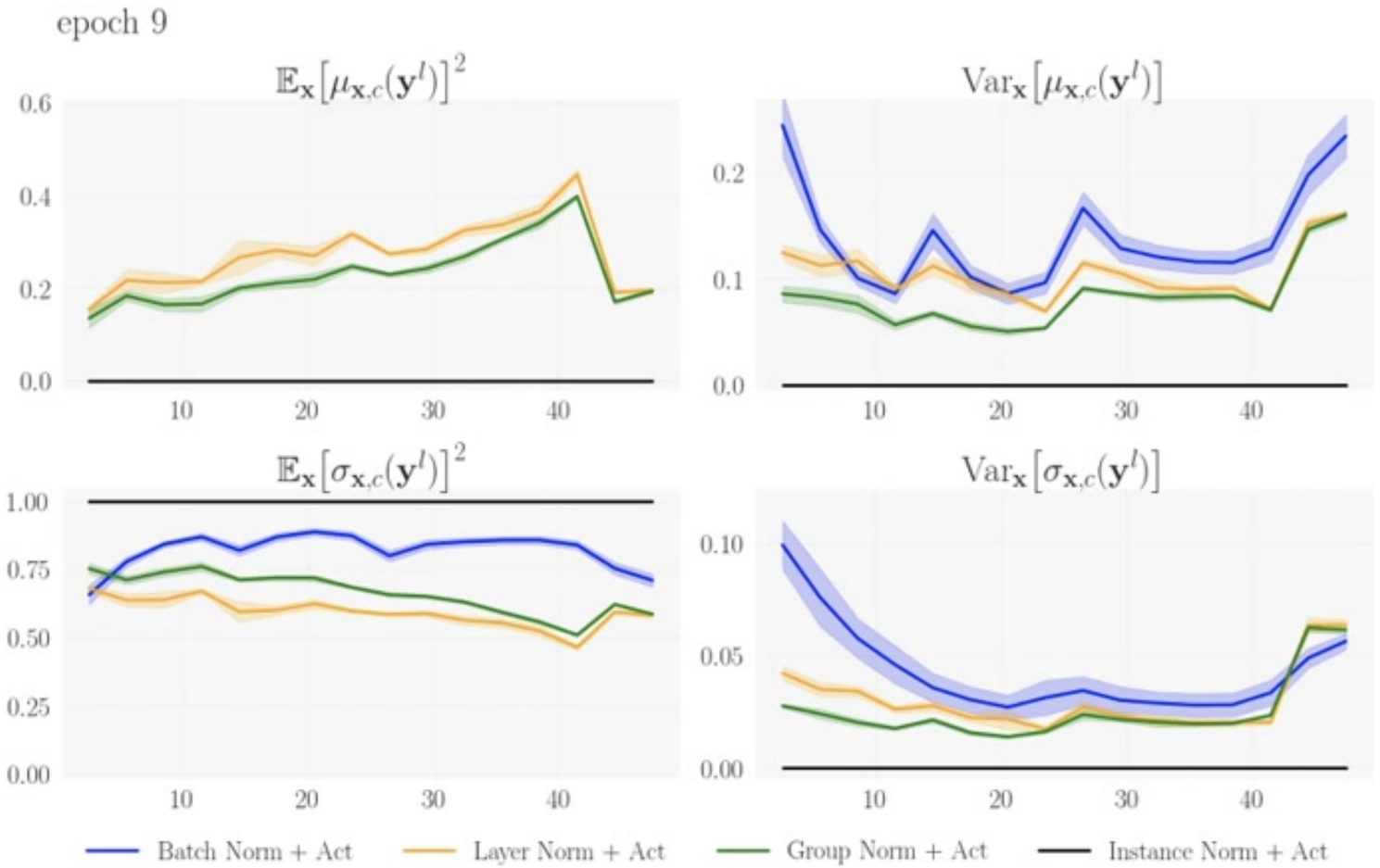
COMMON NORMALIZATION TECHNIQUES IN CONVNETS: VISUALIZING PROS AND CONS

Novel methodology

- Look at expectation and variance over the input $\mathbf{x}$ of $\mu_{\mathbf{x},c}(\mathbf{y}^l)$ and $\sigma_{\mathbf{x},c}(\mathbf{y}^l)$
- The sum of these four terms equals 1

Results

- Layer Norm: 1st term is dominant in deep layers
- Instance Norm: 2nd and 4th terms are constrained to be 0
- Group Norm: middle ground between those 2 issues



MAINTAINING CHANNEL-WISE NORMALIZATION WITH A PROXY-NORMALIZING ACTIVATION STEP(I)

What causes channel-wise denormalization with Layer Norm?

- Layer Norm is not an active cause of aggravation, but not a cause of alleviation either
- The culprits are the affine transform and the activation function ϕ

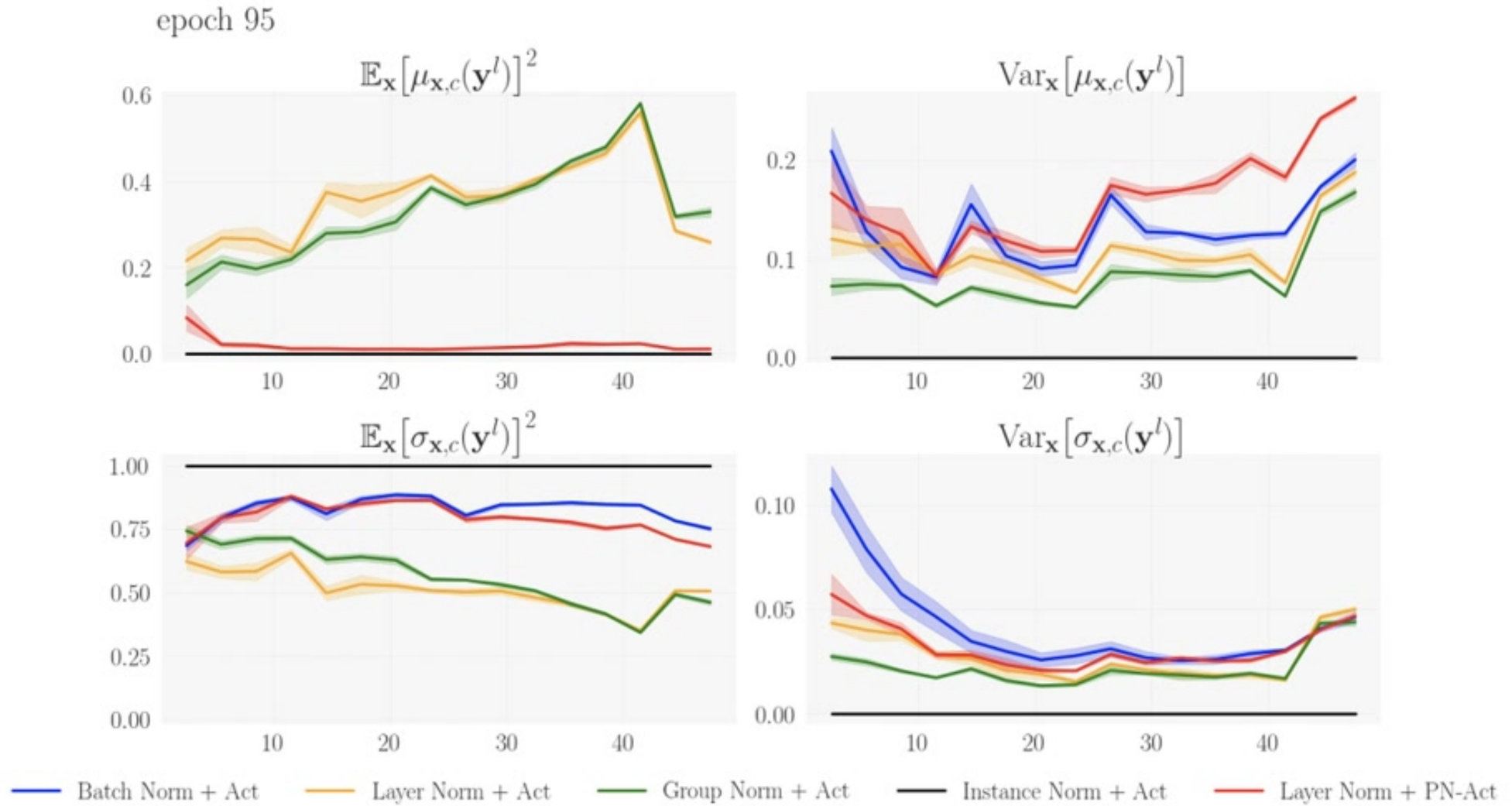
Idea: cancel the effect of the affine transform and ϕ on channel-wise denormalization

- Assimilate $\mathbf{y}^l$ to a proxy variable $\mathbf{Y}^l \approx \mathcal{N}(0,1)$
- Replace the activation step by a proxy-normalized activation step:

✗ $\mathbf{z}^l = \mathrm{Act}(\mathbf{y}^l) = \phi(\gamma^l \mathbf{y}^l + \beta^l)$

✓ $\tilde{\mathbf{z}}^l = \mathrm{PN\text{-}Act}(\mathbf{y}^l) = \dfrac{\phi(\gamma^l \mathbf{y}^l + \beta^l) - \mathbb{E}_{\mathbf{Y}^l}[\phi(\gamma^l \mathbf{Y}^l + \beta^l)]}{\sqrt{\mathrm{Var}_{\mathbf{Y}^l}[\phi(\gamma^l \mathbf{Y}^l + \beta^l)]}}$

MAINTAINING CHANNEL-WISE NORMALIZATION WITH A PROXY-NORMALIZING ACTIVATION STEP (II)



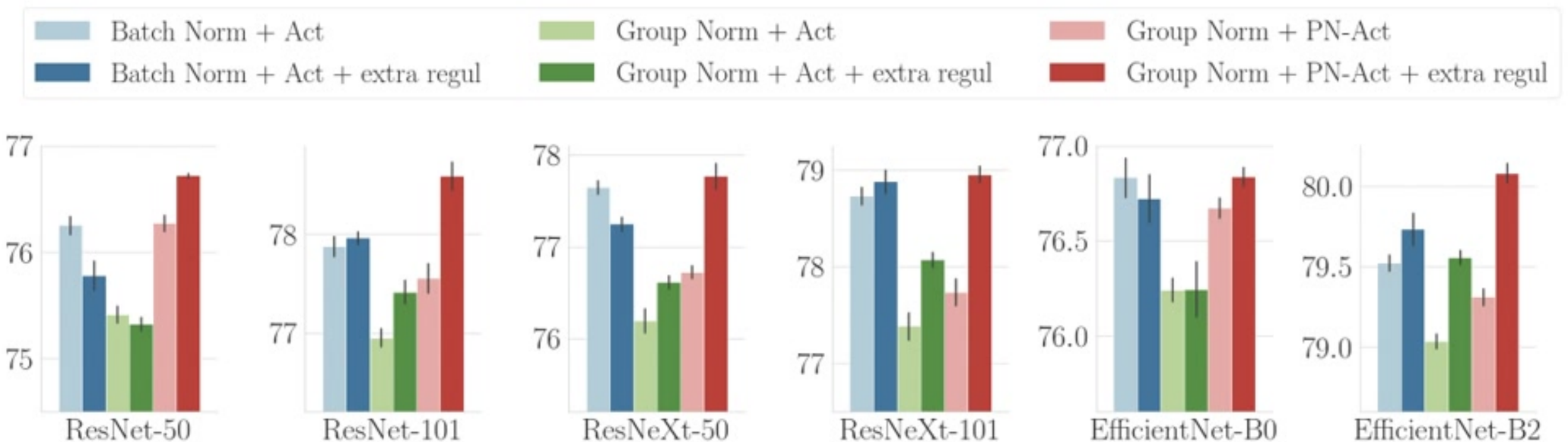
NORMALIZATION APPROACHES IN CONVNETS: PROS AND CONS

		Channel-wise normalization of $\mathbf{y}^l$	Approximate preservation of the network's expressivity	Batch independence
Layer Norm	PN-Act	✓	✓	✓
Group Norm (w/ small number of groups)		✓	✓	✓

PERFORMANCE OF OUR BATCH-INDEPENDENT APPROACH

In our experiments on ImageNet, good task performance is tied to the combination of an **efficient normalization** and an **efficient regularization**.

On larger datasets, regularization would likely be less beneficial and good task performance would likely be tied mainly to an **efficient normalization**.



SUMMING UP

- The incompatibility of Batch Norm with small batch sizes will become more and more problematic in the future.

- With approaches based on the combination of Layer Norm / Instance Norm / Group Norm with the activation step Act:
  - ✗ Either channel-wise normalization is not maintained
  - ✗ Or the network's expressivity is strongly altered
  - ✗ Batch Norm's performance is not matched

- With our batch-independent approach based on the combination of Layer Norm or Group Norm (w/ small number of groups) with the proxy-normalized activation step $\mathrm{PN\text{-}Act}$:
  - ✓ Channel-wise normalization is maintained
  - ✓ The network's expressivity is approximately preserved
  - ✓ Batch Norm's performance is matched

GOING FURTHER

Any question regarding this work?

- Feel free to reach out to us!
- And do not hesitate to come and meet us at the poster session!
- We'll be very happy to interact