Comparing Normalization in Conditional Computation Tasks

Vincent Michalski [1,2] Vikram Voleti [1,2] Samira Kahou [1,3] Anthony Ortiz [1,4]
Pascal Vincent [1,2] Chris Pal [1,5] Doina Precup [1,3]

[1] Quebec Artificial Intelligence Institute (Mila) [2] Université de Montréal, Canada
[3] McGill University, Canada [4] University of Texas, El Paso, USA
[5] Polytechnique Montreal, Canada

We study generalization capabilities of **(Conditional) Batch Normalization** and **(Conditional) Group Normalization** in:

1. VQA
2. Few-Shot Learning
3. Conditional Image Generation

We don't find conclusive evidence to say (C)BN is always better than (C)GN.

$$y_i = \gamma \hat{x}_i + \beta, \qquad \hat{x}_i = \frac{1}{\sigma_i}\left(x_i - \mu_i\right),$$

$$\mu_i = \frac{1}{m}\sum_{k \in \mathcal{S}_i} x_k, \qquad \sigma_i = \sqrt{\frac{1}{m}\sum_{k \in \mathcal{S}_i}\left(x_k - \mu_i\right)^2 + \epsilon}$$

Batch Normalization (BN) $\Longrightarrow \mathcal{S}_i = \{k | k_C = i_C\}$

Group Normalization (GN) $\Longrightarrow \mathcal{S}_i = \{k | k_N = i_N, \lfloor \frac{k_C}{C/G} \rfloor = \lfloor \frac{i_C}{C/G} \rfloor\}$

Conditional (C) $\Longrightarrow \gamma$ & β are functions of condition, such as label.

1. Visual Question Answering

1) CLEVR-CoGenT
2) FigureQA
3) SQOOP

Table 1. Classification accuracy on CLEVR-CoGenT *val2* averaged over three runs.

Model	Accuracy (%)
FiLM (Perez et al., 2018)	75.5 ± 0.7
CGN (all GN)	75.7 ± 0.3
CGN (BN in stem, classifier no norm)	75.7 ± 0.6
CGN (BN in stem and classifier)	$\mathbf{75.8 \pm 0.5}$

Table 2. Classification accuracy on FigureQA *validation2* averaged over three runs.

Model	Accuracy (%)
FiLM (Perez et al., 2018)	$\mathbf{91.6 \pm 0.1}$
CGN (all GN)	91.3 ± 0.4
CGN (BN in stem, classifier no norm)	91.1 ± 0.2
CGN (BN in stem and classifier)	91.3 ± 0.5

Table 3. Classification accuracy on SQOOP averaged over three runs.

Dataset	Model	Accuracy (%)
1 rhs/lhs	FiLM (BN)	72.4 ± 0.5
	CGN (all GN)	74.0 ± 2.8
	CGN (BN — no norm)	73.8 ± 0.3
	CGN (BN — classifier)	$\mathbf{74.9 \pm 3.9}$
2 rhs/lhs	FiLM (BN)	85.0 ± 4.2
	CGN (all GN)	$\mathbf{86.7 \pm 6.3}$
	CGN (BN — no norm)	83.1 ± 0.4
	CGN (BN — classifier)	85.9 ± 5.3
4 rhs/lhs	FiLM (BN)	97.0 ± 1.9
	CGN (all GN)	91.4 ± 0.3
	CGN (BN — no norm)	91.6 ± 1.9
	CGN (BN — classifier)	$\mathbf{99.5 \pm 0.2}$
35 rhs/lhs	FiLM (BN)	$\mathbf{99.8 \pm 0.1}$
	CGN (all GN)	99.8 ± 0.1
	CGN (BN — no norm)	99.8 ± 0.1
	CGN (BN — classifier)	99.8 ± 0.2

2. Few-Shot Learning

- Five-shot TADAM on
  - Fewshot-CIFAR100
  - Mini-Imagenet

Table 4. Five-way five-shot classification accuracy on Fewshot-CIFAR100 (Oreshkin et al., 2018) and Mini-Imagenet (Vinyals et al., 2016) averaged over three runs, where TADAM (BN) is from (Oreshkin et al., 2018)

Dataset	Model	Accuracy (%)
FC100	TADAM (BN)	$\mathbf{53.0 \pm 0.6}$
	TADAM (GN)	52.8 ± 0.5
Mini-Imagenet	TADAM (BN)	$\mathbf{76.4 \pm 0.5}$
	TADAM (GN)	74.1 ± 0.4

3. Conditional Image Generation

- WGAN-GP on CIFAR10

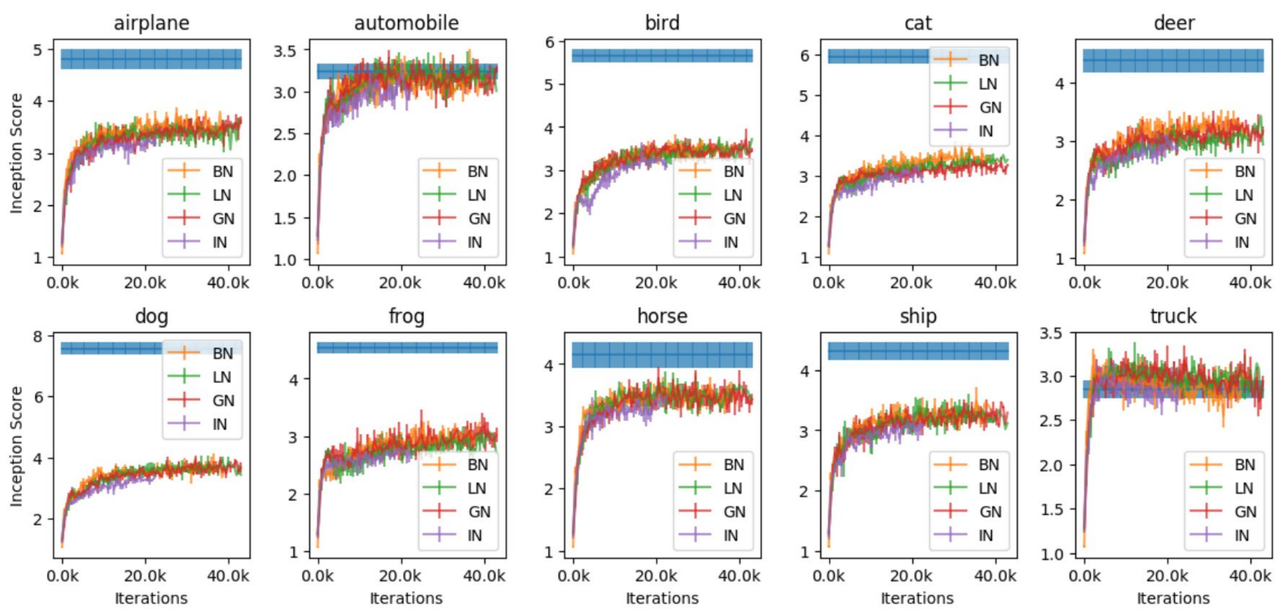

Figure 1. Class-wise Inception Score (IS) of samples generated by WGAN-GP model trained on CIFAR-10. (Blue: IS of true data)