
An Empirical Study of Batch Normalization and Group Normalization in Conditional Computation

Vincent Michalski^{1,5}, Vikram Voleti^{1,5}, Samira Ebrahimi Kahou^{2,5}, Anthony Ortiz³,
Pascal Vincent^{1,5}, Chris Pal^{4,5,6}, and Doina Precup^{2,5}

¹Université de Montréal ²McGill University ³University of Texas - El Paso
⁴Polytechnique Montréal ⁵Quebec Artificial Intelligence Institute (Mila) ⁶Element AI

Abstract

Batch normalization has been widely used to improve optimization in deep neural networks. While the uncertainty in batch statistics can act as a regularizer, using these dataset statistics specific to the training set impairs generalization in certain tasks. Recently, alternative methods for normalizing feature activations in neural networks have been proposed. Among them, group normalization has been shown to yield similar, in some domains even superior performance to batch normalization. All these methods utilize a learned affine transformation after the normalization operation to increase representational power. Methods used in conditional computation define the parameters of these transformations as learnable functions of conditioning information. In this work, we study whether and where the conditional formulation of group normalization can improve generalization compared to conditional batch normalization. We evaluate performances on the tasks of visual question answering, few-shot learning, and conditional image generation.

1 Introduction

In machine learning, the parameters of a model are typically optimized using a fixed training set. The model is then evaluated on a separate partition of the data to estimate its generalization capability. In practice, even under the i.i.d. assumption¹, the distribution of these two finite sets can *appear* quite different to the learning algorithm, making it challenging to achieve strong and robust generalization. This difference is often the result of the fact that a training set of limited size cannot adequately cover the cross-product of all relevant factors of variation. This issue can be addressed by making strong assumptions that simplify discovering a family of patterns from limited data. Bahdanau et al. [1], for example, show that their proposed synthetic relational reasoning task can be solved by a Neural Module Network (NMN) [2] with fixed tree structure, while models without this structural prior fail.

Recent studies propose different benchmarks for evaluating task specific models for their generalization capacity [3, 4, 1]. While in this paper, we focus on visual question answering (VQA), few-shot learning and generative models, any improvement in this direction can also benefit other domains such as reinforcement learning. Some of the best-performing models for each of these tasks are deep neural networks that employ Conditional Batch Normalization (CBN) [5] for modulating normalized activations with contextual information. For Batch Normalization (BN), one usually has to precompute activation statistics over the training set to be used during inference. Since BN [6] (and thus also CBN) relies on dataset statistics, it seems that it may be vulnerable to significant domain shifts between training and test data. A recent study by Galloway et al. [7] indicates that BN is also vulnerable to adversarial examples.

¹All data samples are assumed to be drawn independently from an identical distribution (i.i.d.).

The recently proposed Group Normalization (GN) [8] normalizes across groups of feature maps instead of across batch samples. Here, we explore whether a conditional formulation of GN is a viable alternative for CBN. GN is conceptually simpler than BN, as its function is the same during training and inference. Further, GN can be used with small batch sizes, which may help in applications with particularly large feature maps, such as medical imaging or video processing, in which the available memory can be a constraint.

We compare Conditional Group Normalization (CGN) and CBN in a variety of tasks to see whether there are any significant performance differences. Section 2 reviews some basic concepts that our work builds upon. Section 3 describes setup and results of our experiments. Finally, we draw conclusions and present some directions for future work in Section 4.

2 Background

2.1 Normalization Layers

Several normalization methods have been proposed to stabilize and speed-up the training of deep neural networks [6, 8, 9, 10]. To stabilize the range of variation of network activations x_i , methods such as BN [6] first normalize the activations by subtracting mean μ_i and dividing by standard deviation σ_i :

$$\hat{x}_i = \frac{1}{\sigma_i} (x_i - \mu_i) \quad (1)$$

The distinction between different methods lies in how exactly these statistics are being computed. Wu and He [8] aptly summarize several methods using the following notation. Let $i = (i_N, i_C, i_H, i_W)$ be a four-dimensional vector, whose elements index the features along the batch, channel, height and width axes, respectively. The computation of the statistics can then be written as

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

where the set \mathcal{S}_i of size m is defined differently for each method and ϵ is a small constant for numerical stability. BN, for instance, corresponds to:

$$\text{BN} \implies \mathcal{S}_i = \{k | k_C = i_C\}, \quad (3)$$

i.e. \mathcal{S}_i is the set of all pixels sharing the same channel axis, resulting in μ_i and σ_i being computed along the (N, H, W) axes.

As Lei Ba et al. [9] point out, the performance of BN is highly affected by the batch size hyperparameter. This insight led to the introduction of several alternative normalization schemes, that normalize per sample, i.e. not along batch axis N . Layer Normalization (LN) [9], which normalizes activations within each layer, corresponds to the following set definition:

$$\text{LN} \implies \mathcal{S}_i = \{k | k_N = i_N\}. \quad (4)$$

Ulyanov et al. [10] introduce Instance Normalization (IN) in the context of image stylization. IN normalizes separately for each sample and each channel along the spatial dimensions:

$$\text{IN} \implies \mathcal{S}_i = \{k | k_N = i_N, k_C = i_C\}. \quad (5)$$

Recently, Wu and He [8] introduced GN, which draws inspiration from classical features such as HOG [11]. It normalizes features per sample, separately within each of G groups, along the channel axis:

$$\text{GN} \implies \mathcal{S}_i = \{k | k_N = i_N, \lfloor \frac{k_C}{C/G} \rfloor = \lfloor \frac{i_C}{C/G} \rfloor\} \quad (6)$$

GN can be seen as a way to interpolate between the two extremes of LN (corresponding to $G = 1$, i.e. all channels are in a single group) and IN (corresponding to $G = C$, i.e. each channel is in its own group).

After normalization, all above mentioned methods insert a scaling and shifting operation using learnable per-channel parameters γ and β :

$$y_i = \gamma \hat{x}_i + \beta \quad (7)$$

This “de-normalization” is done to restore the representational power of the normalized network layer [6].

CBN [5, 12] is a conditional variant of BN, in which the learnable parameters γ and β in Equation 7 are replaced by learnable functions

$$\gamma(c_k) = W_\gamma c_k + b_\gamma, \quad \beta(c_k) = W_\beta c_k + b_\beta \quad (8)$$

of some per-sample conditioning input c_k to the network with parameters $W_\gamma, W_\beta, b_\gamma, b_\beta$. In a VQA model, c_k would for instance be an embedding of the question [12]. Dumoulin et al. [13] introduce Conditional Instance Normalization (CIN), a conditional variant of IN similar to CBN, replacing BN with IN. In our experiments, we also explore a conditional variant of GN.

2.2 Visual Question Answering

In VQA [14, 15], the task is to answer a question about an image. This task is usually approached by feeding both image and question to a parametric model, which is trained to predict the correct answer, for instance via classification among all possible answers in the dataset. One recent successful model for VQA is the Feature-wise Linear Modulation (FiLM) architecture [12], which employs CBN to modulate visual features based on an embedding of the question.

2.3 Few-Shot Classification

The task of few-shot classification consists in the challenge of classifying data given only a small set of support samples for each class. In episodic M -way, k -shot classification tasks, meta-learning models [16] learn to adapt a classifier given multiple M -class classification tasks, with k support samples for each class. The meta-learner thus has to solve the problem of generalizing between these tasks given the limited number of training samples. In this work we experiment with the recently proposed Task dependent adaptive metric (TADAM) architecture [17]. It belongs to the family of meta-learners, that employ nearest neighbor classification within a learned embedding space. In the case of TADAM, the network providing this embedding is modulated by a task embedding using CBN.

2.4 Conditional Image Generation

Some of the most successful models for generating images are Generative Adversarial Networks (GANs) [18]. This approach involves training a neural network (Generator) to generate an image, while the only supervisory signal is that from another neural network (Discriminator) which indicates whether the image looks real or not. Several variants of GANs [19, 20] have been proposed to condition the image generation process on a class label. More recently, the generators that work best stack multiple ResNet-style [21] architectural blocks, involving two CBN-ReLU-Conv operations and an upsampling operation. These blocks are followed by a BN-ReLU-Conv operation to transform the last features into the shape of an image.

Such models can be trained as Wasserstein GANs using gradient penalty (WGAN-GP) as proposed by Gulrajani et al. [22], which gives mathematically sound arguments for an optimization framework. We adopt this framework for our experiments. More recently, two of the most noteworthy GAN architectures, Self-Attention GAN (SAGAN) [23] and BigGAN [24], use architectures similar to WGAN-GP, with some important changes. SAGAN inserts a self-attention mechanism [25, 26, 27] to attend over important parts of features during the generation process. In addition, it uses spectral normalization [28] to stabilize training. The architecture of BigGAN is the same as for SAGAN, with the exception of an increase in batch size and channel widths, as well as some architectural changes to improve memory and computational efficiency. Both these models have been successfully used in generating high quality natural images. In our experiments, we compare performance metrics of WGAN-GP networks using two types of normalization.

3 Experiments

3.1 Visual Question Answering

We study whether substituting CGN for CBN in the VQA architecture FiLM [12] yields comparable performance. We run experiments on several recently proposed benchmarks for compositional generalization.

3.1.1 Datasets

CLEVR Compositional Generalization Test (CLEVR-CoGenT) [3] is a variant of the popular Compositional Language and Elementary Visual Reasoning (CLEVR) dataset [3], that tests for compositional generalization. The images consist of rendered three-dimensional scenes containing several shapes (small and large cubes, spheres and cylinders) of differing material properties (*metal* or *rubber*) and colors. Questions involve *queries* for object attributes, *comparisons*, *counting* of sets and combinations thereof. In contrast to the regular CLEVR dataset, the training set of CLEVR-CoGenT explicitly combines some shapes only with different subsets of four out of eight colors, and provides two validation sets: one with the same combinations (*valA*) and one in which the shape-color assignments are swapped (*valB*). To perform well on *valB*, the model has to generalize to unseen combinations of shapes and colors, i.e. it needs to somewhat capture the compositionality of the task. Figure 1a shows an example from this dataset.

Figure Question Answering (FigureQA) [4] is a VQA dataset consisting of mathematical plots with templated yes/no question-answer pairs that address relations between plot elements. The dataset contains plots of five types (vertical/horizontal bar plots, line plots, pie charts and dot-line plots). Each plot has between 2 and 10 elements, each of which has one of 100 colors. Plot elements (e.g. a slice in a pie chart) are identified by their color names in the questions. Questions query for *one-vs-one* or *one-vs-all* attribute relations, e.g. "Is Lime Green less than WebGray?" or "Does Cadet Blue have the minimum area under the curve?". Similar to CLEVR-CoGenT, FigureQA requires compositional generalization. The overall 100 colors are split into two sets *A* and *B*, each containing 50 unique colors. During training, colors of certain plot types are sampled from set *A*, while the remaining plot types use colors from set *B* (*scheme 1*). There are two validation sets, one using the same color scheme, and one for which the plot-type to color assignments are swapped (*scheme 2*). See Figure 1b for a sample from the dataset.

Spatial Queries On Object Pairs (SQOOP) [1] is a recently introduced dataset that tests for systematic generalization. It consists of images containing five randomly chosen and arranged objects (digits and characters). Questions concern the four spatial relations *LEFT OF*, *RIGHT OF*, *ABOVE* and *BELOW* and the queries are all of the format "X R Y?", where X and Y are left-hand and right-hand objects and R is a relationship between them, e.g. "nine LEFT OF a?". To test for systematic generalization, only a limited number of combinations of each left-hand object with different right-hand objects Y are shown during training. In the hardest version of the task (1 rhs/lhs), only a single right-hand side object is combined with each left-hand side object. For instance, the training set of this version may contain images with the query "A RIGHT OF B", but no images with queries about relations of left-hand object A with any other object than B. The test set contains images and questions about all combinations, i.e. it evaluates generalization to relations between novel object combinations. Figure 1c shows an example from the training set.

3.1.2 Model

We experiment with several small variations of the FiLM architecture [12]. The original architecture in Perez et al. [12] consists of an unconditional *stem* network, a core of four ResNet [21] blocks with CBN [5] and a classifier. The stem network is either a sequence of residual blocks trained from scratch or a fixed pre-trained feature extractor followed by a learnable layer of 3×3 convolutions. The scaling and shifting parameters of the core layers are affine transforms of a question embedding provided by a gated recurrent unit (GRU) [29]. The output of the last residual block is fed to the classifier, which consists of a layer of $512 \ 1 \times 1$ convolutions, global max-pooling, followed by a

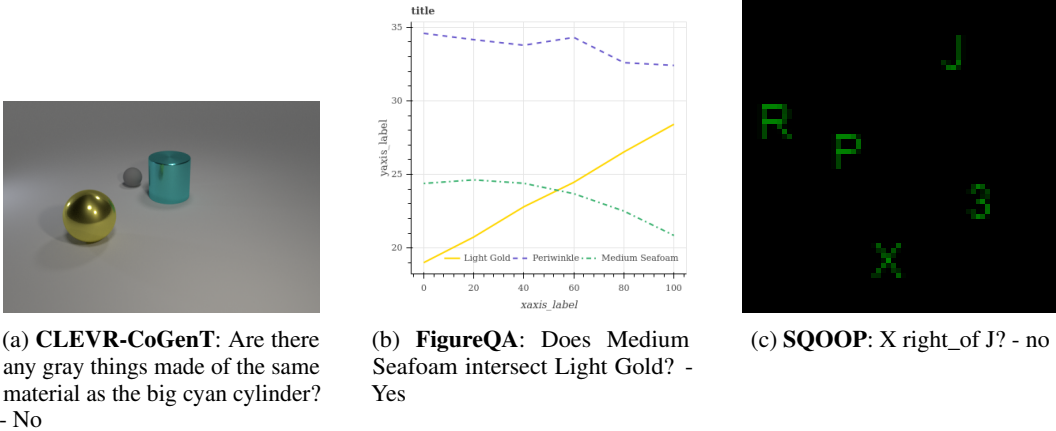


Figure 1: Examples of the VQA datasets used in our experiments.

fully-connected ReLU [30] layer using (unconditional) BN and a softmax layer, which outputs the probability of each possible answer. We train the following three variants that include CGN²:

1. all conditional and regular BN layers are replaced with corresponding conditional or regular GN layers.
2. all CBN layers are replaced with CGN, regular BN layers are left unchanged.
3. all CBN layers are replaced with CGN, regular BN layers are left unchanged, except the fully-connected hidden layer in the classifier, for which we remove normalization.

Besides the described changes in the normalization layers, the architecture and hyperparameters are the same as used in Perez et al. [12] for all experiments, except for SGOOP where they are the same as in Bahdanau et al. [1]. The only difference is that we set the constant ϵ of the Adam optimizer [31] to $1e-5$ to improve training stability³. For SGOOP, the input to the residual network are the raw image pixels. For all other networks, the input is features extracted from layer *conv4* of a ResNet-101 [21], pre-trained on ImageNet [32], following Perez et al. [12].

3.1.3 Results

Tables 1, 2 and 3 show the results of training FiLM with CBN and CGN on the three considered datasets. In the experiments on CLEVR-CoGenT, all three CGN variants of FiLM achieve a slightly higher average accuracy. On FigureQA, CBN outperforms CGN slightly. In the hardest SGOOP variant with only one right-hand side object per left-hand side object (*1 rhs/lhs*), all three variants of CGN achieve a higher performance than CBN. For SGOOP variants whose training sets contain more combinations, CGN did not converge in some cases. Learning curves of models successfully trained on SGOOP seem to follow the same pattern: For a relatively large number of gradient updates there is no significant improvement. Then, at some point, almost instantly the model achieves 100% training accuracy. It is possible that a hyperparameter search or additional regularization is required to guarantee convergence.

3.2 Few-Shot Learning

CBN has also been used in recent methods for few-shot learning [17, 33]. We replicate the experiments of Oreshkin et al. [17] on Mini-ImageNet and Fewshot-CIFAR100 (FC100) using their code for TADAM⁴ and compare the results with a version that uses CGN instead of CBN.

²We always set the number of groups to 4, as the authors of Wu and He [8] showed that this hyperparameter does not have a large influence on the performance. This number was selected using uniform sampling from the set $\{2, 4, 8, 16\}$.

³The authors of Perez et al. [12] confirmed occasional gradient explosions with the original setting of $1e-8$.

⁴<https://github.com/ElementAI/TADAM>

Table 1: Classification accuracy on CLEVR-CoGenT *valB*. Mean and standard deviation of three runs with early stopping on *valA* are reported for the models we trained.

Model	Accuracy (%)
CBN (FiLM [12])	75.600
CBN (FiLM, our results)	75.539 ± 0.671
CGN (all GN)	75.758 ± 0.356
CGN (BN in stem, classifier no norm)	75.703 ± 0.571
CGN (BN in stem and classifier)	75.807 ± 0.511

Table 2: Classification accuracy on FigureQA *validation2*, mean and standard deviation of three runs after early stopping on *validation1*.

Model	Accuracy (%)
CBN (FiLM, our results)	91.618 ± 0.132
CGN (all GN)	91.343 ± 0.436
CGN (BN in stem, classifier no norm)	91.080 ± 0.166
CGN (BN in stem and classifier)	91.317 ± 0.514

Table 3: Test accuracies on several versions of SGOOP. Mean and standard deviation of three runs after early stopping on the validation set are reported for the models we trained.

Dataset	Model	Accuracy (%)
1 rhs/lhs	CBN (FiLM [1])	65.270 ± 4.610
	CBN (FiLM, our results)	72.369 ± 0.529
	CGN (all GN)	74.020 ± 2.814
	CGN (BN in stem, classifier no norm)	73.824 ± 0.334
	CGN (BN in stem and classifier)	74.929 ± 3.888
2 rhs/lhs	CBN (FiLM [1])	80.200 ± 4.320
	CBN (FiLM, our results)	84.966 ± 4.165
	CGN (all GN)	86.689 ± 6.308
	CGN (BN in stem, classifier no norm)	83.109 ± 0.381
	CGN (BN in stem and classifier)	85.859 ± 5.318
4 rhs/lhs	CBN (FiLM [1])	90.420 ± 1.000
	CBN (FiLM, our results)	97.043 ± 1.958
	CGN (all GN)	91.404 ± 0.318
	CGN (BN in stem, classifier no norm)	91.601 ± 1.937
	CGN (BN in stem and classifier)	99.474 ± 0.254
35 rhs/lhs	CBN (FiLM [1])	99.803 ± 0.219
	CBN (FiLM, our results)	99.841 ± 0.043
	CGN (all GN)	99.755 ± 0.025
	CGN (BN in stem, classifier no norm)	99.815 ± 0.122
	CGN (BN in stem and classifier)	99.782 ± 0.155

3.2.1 Datasets

Mini-ImageNet was proposed by Vinyals et al. [34] as a benchmark for few-shot classification. It contains 100 classes, for each of which there are 600 images of resolution 84×84 . To generate five-way five-shot classification tasks five classes and five support samples for each class are sampled uniformly. The remaining images are used to compute the accuracy. Using the proposed split by Ravi and Larochelle [16], we uniformly sample training tasks from a subset of 64 classes. The remaining 36 classes are divided into 16 for meta-validation and 20 for meta-testing.

Fewshot-CIFAR100 [17] is a few-shot classification version of the popular CIFAR100 data set [35]. Similarly to Mini-ImageNet, it contains 100 classes and 600 samples per class. The resolution of the images is 32×32 . The classes are split by superclasses to reduce information overlap between data set partitions, which makes the task more challenging than Mini-ImageNet. The training partition

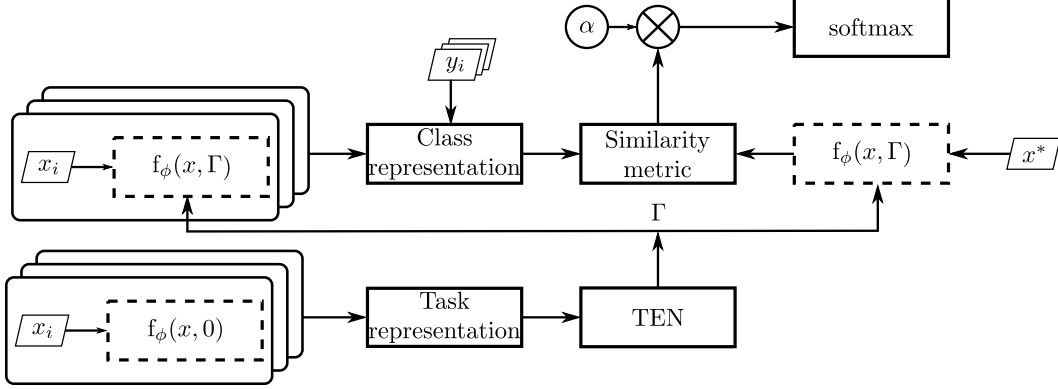


Figure 2: Architecture of TADAM [17]. Boxes with dashed border share parameters. Figure adapted from [17].

contains 60 classes belonging to 12 superclasses. The validation and test partitions contain 20 classes belonging to 5 superclasses each. The tasks are sampled uniformly as in Mini-ImageNet.

3.2.2 Model

TADAM [17] is a metric-based few-shot classifier, i.e. it learns a measure of similarity between query samples and class representations. The metric is based on a learned image embedding $f_\phi(x, c)$ provided by a residual network. Figure 2 shows a diagram of the overall architecture. Each class template is computed as the average embedding of all support samples for the respective class. The Euclidean distances between the embedding of a query sample and each of the class templates, weighted by a learned scaling factor α , is then used to classify the query sample x^* . The embedding network f_ϕ (see the dashed boxes in Figure 2) is modulated using CBN with a conditioning input c . In the computation of the similarity metric, c is fed by a task embedding Γ provided by a task embedding network (TEN), which reads the average embeddings of support samples from all classes of the task. Note that f_ϕ is evaluated without conditioning (i.e. by setting c to a zero vector⁵) in the computation of the task embedding Γ (see bottom of Figure 2). For the GN version we replaced all conditional and regular BN layers with their corresponding conditional or regular GN version (with the number of groups set to 4). For a complete description of the experimental setup, including all other hyperparameters, we refer the reader to Oreshkin et al. [17].

3.2.3 Results

Table 4: Five-way five-shot classification accuracy on Fewshot-CIFAR100 [17] and Mini-Imagenet [34], mean and standard deviation of ten runs.

Dataset	Model	Accuracy (%)
FC100	TADAM (CBN) [17]	52.996 \pm 0.610
	TADAM (CGN)	52.807 \pm 0.509
Mini-Imagenet	TADAM (CBN) [17]	76.414 \pm 0.499
	TADAM (CGN)	74.032 \pm 0.373

We see that using CGN instead of CBN yields only slightly reduced performance on FC100, while there is a considerable 2.4% gap for Mini-ImageNet. Note, that we simply reuse the hyperparameters from Oreshkin et al. [17], which were tuned for CBN.

⁵The conditioning input is implemented as a deviation from the identity transform (unity scaling and zero shift), so setting it to zero does not change the normalized activations.



Figure 3: Samples from models trained with different normalization techniques. The images in each column belong to the same class, ordered as ‘airplane’, ‘automobile’, ‘bird’, ‘cat’, ‘deer’, ‘dog’, ‘frog’, ‘horse’, ‘ship’, ‘truck’. Samples are not cherry-picked.

3.3 Conditional Image Generation

Here we compare CBN and CGN on the task of generating images conditioned on their class label using the WGAN-GP [22] architecture.

3.3.1 Dataset

CIFAR-10 [35] is a data set containing 60000 32×32 images, 6000 for each of 10 classes. The dataset is split into 50000 training and 10000 test samples.

3.3.2 Model

We replicated the WGAN-GP [22] architecture from the original paper, which uses CBN. As in other tasks, we also train the CGN variants, where we substitute conditional and unconditional BN layers with the corresponding conditional or unconditional GN layers, with number of groups set to 4. We use the optimization setup from Gulrajani et al. [22]: a learning rate of $2e^{-4}$ for both generator and discriminator, five discriminator updates per generator update, and we also use the Adam optimizer [31]. We train using a single GPU (NVIDIA P100) and a batch size of 64.

3.3.3 Results

Figure 3 shows samples from WGAN-GP trained using each of the two normalization methods. For both normalization methods, in addition to a qualitative check of the generated samples, we calculate two scores that are widely used in the community to evaluate image generation Inception Score (IS) [36] and Fréchet Inception Distance (FID) [37]. We use publicly available code to calculate IS⁶ and FID⁷. The computed values for real data may differ slightly from the original ones since these use PyTorch [38] implementations, while the original papers use TensorFlow [39]. However, we compare the same implementation of these metrics for true and generated data.

IS is meant to measure the natural-ness of an image by checking the embedding of the generated images on a pre-trained Inception network [40]. Although the suitability of the IS for this purpose has been rightfully put into question [41], it continues to be used frequently. FID measures how similar two sets of images are, by computing the Fréchet distance between two multivariate Gaussians fitted to the embeddings of the images from the two sets. The embeddings are obtained from a pre-trained InceptionV3 network [40]. In this case, we measure the distance between the real CIFAR-10 images, and the generated ones. This is a better metric than IS, since there is no constraint on the images being natural, and it is able to quantify not only their similarity to the real images, but also diversity in the generated images.

We first calculate the IS of the true images of CIFAR-10, for each class separately. Then, during training of a model, we sample images from the generator at regular intervals, and calculate the IS and FID of those images for each class separately. This allows us to see the effect of the different

⁶<https://github.com/sbarratt/inception-score-pytorch>

⁷<https://github.com/mseitzer/pytorch-fid>

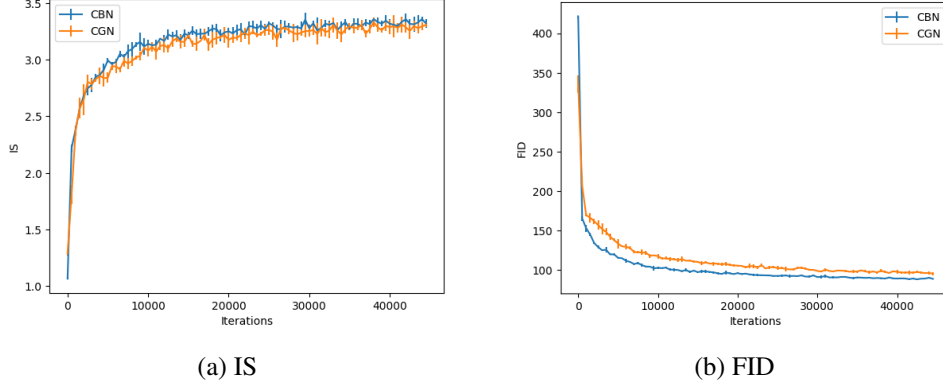


Figure 4: (a) Inception score (IS, higher is better) and (b) FID (lower is better) of samples generated by WGAN-GP model while training on CIFAR-10.

normalization techniques on the conditional generation process. We average our results from four runs with different seeds, shown in Figure 4.

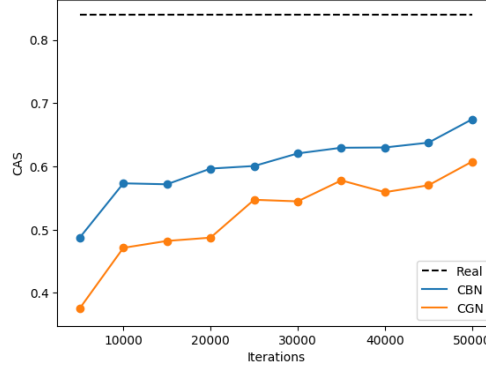


Figure 5: Classification Accuracy Score (CAS) using a ResNet classifier trained on samples generated while training on CIFAR-10 with WGAN-GP using (blue) CBN and (green) CGN, while (black) is the value when trained with true data. All classifiers have been trained with the same hyperparameters.

We also calculate the recently proposed Classification Accuracy Score (CAS) [42] for one instance of training using WGAN-GP with CBN and CGN each, shown in Figure 5. In the computation of this metric, a ResNet [21] classifier is trained on data sampled from the generative model being evaluated. Then the accuracy of this classifier on the true validation data is calculated. Ravuri and Vinyals [42] mention that this could indicate the closeness of the generated data distribution to the true data distribution. All three metrics indicate that CBN is better than CGN in conditional generative models of images such as WGAN-GP.

The WGAN-GP model architecture consists of a series of residual blocks followed by bn-relu-conv layers. Each residual block contains two bn-relu-conv modules. Since the architectures of more recent models such as SAGAN [23] and BigGAN [24] are similar to that of the one we used, it is likely that the conclusions we draw from the WGAN-GP experiments transfer to them.

4 Conclusion

Because the performance of CBN heavily depends on the batch size and on how well training and test statistics match, we investigate the use of CGN as a potential alternative for CBN. We consider a set of experiments for VQA, few-shot learning and image generation tasks in which some of the best models rely on CBN for conditional computation. We experimentally show that the effect of this substitution is task-dependent, with performance increases in some VQA tasks that focus on systematic generalization, but a clear decrease in performance in conditional image generation.

CGN’s simpler implementation, its consistent behaviour during training and inference time, as well as its independence from batch sizes, are all good reasons to explore its adoption instead of CBN in tasks that require systematic generalization. That being said, further analysis is required to be able to confidently suggest one method over the other. For instance, a hyperparameter search for each of the normalization methods would be required to provide a better performance comparison. Also, we would like to characterize the sensitivity of CBN’s performance to the batch size and focus on domains, such as medical imaging or video processing, for which efficient large-batch training becomes nontrivial. Lastly, since some of the success of BN (and consequently also CBN) can be attributed to the regularization effect introduced by noisy batch statistics, it seems worthwhile to explore combinations of CGN with additional regularization as suggested for GN by Wu and He [8]. The latter is also motivated by recent successful attempts at replacing (unconditional) BN with careful network initialization [43], which relies on additional regularization [44] to match generalization performance.

Acknowledgments

We thank Boris Oreshkin, Eugene Belilovsky, Matthew Scicluna, Mahdi Ebrahimi Kahou, Kris Sankaran and Alex Lamb for helpful discussions. This research was enabled in part by support provided by Compute Canada.

References

- [1] Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. Systematic generalization: What is required and can it be learned? *arXiv preprint arXiv:1811.12889*, 2018.
- [2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, 2016.
- [3] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [4] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *Workshop in the International Conference on Learning Representations*, 2017.
- [5] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2017.
- [6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015.
- [7] Angus Galloway, Anna Golubeva, Thomas Tanay, Medhat Moussa, and Graham W Taylor. Batch normalization is a cause of adversarial vulnerability. *arXiv preprint arXiv:1905.02161*, 2019.
- [8] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.
- [9] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [10] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016.
- [11] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR’05)*, volume 1, pages 886–893. IEEE Computer Society, 2005.

- [12] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [13] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *International Conference on Learning Representations (ICLR)*, 2017.
- [14] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014.
- [15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.
- [16] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.
- [17] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731, 2018.
- [18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.
- [19] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.
- [20] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning (ICML)*, 2017.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.
- [22] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, page 5769–5779, 2017.
- [23] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *International Conference on Learning Representations (ICLR)*, 2018.
- [24] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *International Conference on Learning Representations*, 2019.
- [25] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [27] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.
- [28] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *International Conference on Learning Representations (ICLR)*, 2018.
- [29] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

- [30] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
- [31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2014.
- [32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [33] Xiang Jiang, Mohammad Havaei, Farshid Varno, Gabriel Chartrand, Nicolas Chapados, and Stan Matwin. Learning to learn with conditional class dependencies. 2018.
- [34] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.
- [35] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.
- [36] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016.
- [37] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.
- [38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*, 2017.
- [39] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <http://tensorflow.org/>. Software available from tensorflow.org.
- [40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [41] Shane Barratt and Rishi Kant Sharma. A note on the inception score. *CoRR*, abs/1801.01973, 2018.
- [42] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *arXiv preprint arXiv:1905.10887*, 2019.
- [43] Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. In *International Conference on Learning Representations*, 2019.
- [44] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=r1Ddp1-Rb>.