Improve Wasserstein GAN-GP with Group Normalization

Nan Wu, Qianyu Cheng, Zhiming Guo

Center for Data Science, NYU

BACKGROUND

- **Generative Adversarial Network**, estimated generative models by using an adversarial process
- **WGAN** further developed the model by introducing Wasserstein distance as a metric to optimize discriminator
- As the model still appears slow to converge and hard to train, improved **WGAN-GP** with a gradient penalty in loss to achieve better learning result

$$L = \underbrace{\underset{\tilde{\boldsymbol{x}} \sim \mathbb{P}_g}{\mathbb{E}}[D(\tilde{\boldsymbol{x}})] - \underset{\boldsymbol{x} \sim \mathbb{P}_r}{\mathbb{E}}[D(\boldsymbol{x})]}_{\text{Original critic loss}} + \underbrace{\lambda \underset{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}}{\mathbb{E}}\left[(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2\right]}_{\text{Our gradient penalty}}.$$

MOTIVATION & INTRODUCTION

- **Issue with batch normalization (BN) in WGAN-GP**
  - WGAN-GP penalizes the norm of the gradient to individual data input, not the entire batch of data
  - Batch normalization, however, maps whole batch of inputs for a batch of outputs
- **Layer normalization (LN)and instance normalization (IN)**
  - Both don't suffer from BN's issues but didn't perform well as expected
- **Group normalization (GN)**
  - GN divides the channels into groups and within each group computes the mean and variance and therefore overcomes the BN's issues
- **Goal:** we embedded GN in WGAN-GP and DAGAN. Quality of image generation is evaluated by inception score

NORMALIZATION

- **General Normalization Formula**

$$\hat{x}_i = \frac{1}{\sigma_i}(x_i - \mu_i). \qquad i = (i_N, i_C, i_H, i_W)$$

where N is the batch axis, C is the channel axis, and H and W are the spatial height and width axes.

- **Batch Normalization**

$$\mathcal{S}_i = \{k \mid k_C = i_C\}$$

where i_c (and k_c) denotes the sub-index of i (and k) along the C axis

- **Layer Normalization**

$$\mathcal{S}_i = \{k \mid k_N = i_N\}$$

- **Group Normalization**

$$\mathcal{S}_i = \{k \mid k_N = i_N, \lfloor \frac{k_C}{C/G} \rfloor = \lfloor \frac{i_C}{C/G} \rfloor\}$$

where G is the number of groups, C/G is the number of channels per group

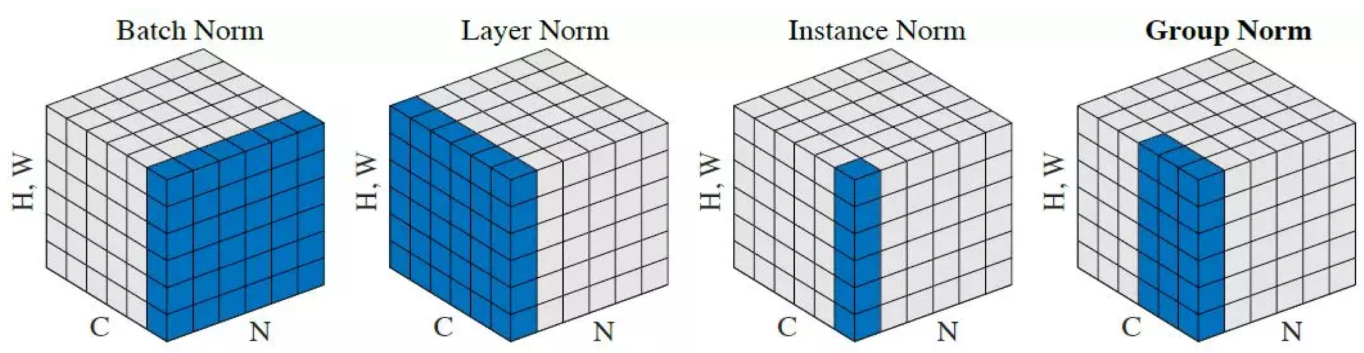

Figure 1: Different Types of Normalization

ARCHITECTURE & METHODOLOGY

- **Design of architecture**
  - RESNET is used for generator and discriminator.

Generator G(z)			
	Kernel Size	Resample	Output Shape
z	-	-	128
Linear	-	-	$128 \times 4 \times 4$
Residual Block	$[3 \times 3] \times 2$	Up	$128 \times 8 \times 8$
Residual Block	$[3 \times 3] \times 2$	Up	$128 \times 16 \times 16$
Residual Block	$[3 \times 3] \times 2$	Up	$128 \times 32 \times 32$
Conv, tanh	3×3	-	$3 \times 32 \times 32$

Discriminator D(x)			
	Kernel Size	Resample	Output Shape
Residual Block	$[3 \times 3] \times 2$	Down	$128 \times 16 \times 16$
Residual Block	$[3 \times 3] \times 2$	Down	$128 \times 8 \times 8$
Residual Block	$[3 \times 3] \times 2$	-	$128 \times 8 \times 8$
Residual Block	$[3 \times 3] \times 2$	-	$128 \times 8 \times 8$
ReLU, mean pool	-	-	128
Linear	-	-	1

Figure 2: ResNet Architecture used in WGAN-GP for CIFAR-10

- **Design of discriminator**
  - There are 6 normalization operation in discriminator: the discriminator has 3 residual blocks; each residual block includes 2 convolutional layers; each convolutional layer has an operation

EXPERIMENT RESULTS ON CIFAR-10

- **Dataset: CIFAR-10**
  - 32x32 color images in 10 classes
  - 50,000 training and 10,000 test

Figure 3: Generated samples from ResNet with Group Norm

The images in figure 3 can be roughly classified to a class. However, due to the low resolution of CIFAR 10 makes difficulty to tell if the generated images are shaper.

Model Settings	IS(Unsupervised)
Simple-CNN	3.261
Simple-CNN (Group norm)	3.324
ResNet (Layer norm)	3.632
ResNet (Group norm)	3.841

Figure 4: Inception Score

Group Normalization improves inception score a little bit for both Simple-CNN and ResNet over baseline mode.

DATA AUGMENTATION WITH WGAN

We explored the ability of DAGAN in augmenting patch level mammograms and improved its performance by introducing group normalization in the architecture. Patches with a resolution of 256 * 256 are sampled from a clinically realistic mammography data set which is provided by Medical School of New York University.

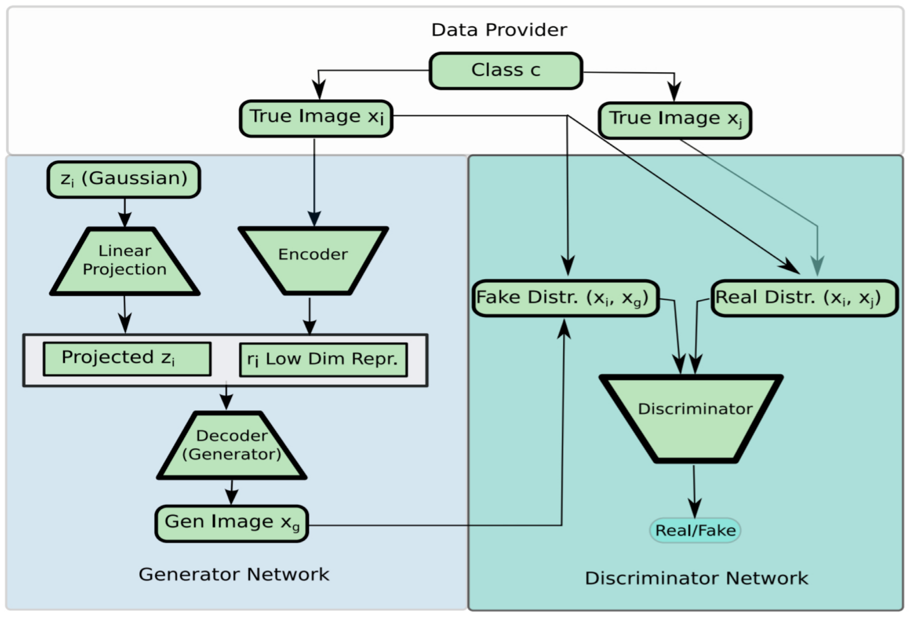

Figure 5: DAGAN Architecture

Figure 6: Original DAGAN

Figure 7: DAGAN with GN

DISCUSSION & FUTURE WORK

- We visualized samples generated from ResNet with layer normalization and group normalization, (here we took the best setting of number of groups in each layer according to our experiment)
- Due to the low resolution of images in CIFAR 10, it is hard to tell the classes of images. However, model with group normalization does give raise sharper and more realistic samples
- Group Normalization did show better result for WGAN-GP, we will try out more combination of groups for group normalization layers in future work

REFERENCES

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.

Yuxin Wu and Kaiming He. Group normalization. *arXiv preprint arXiv:1803.08494*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.