Adversarial Knowledge Distillation for a Compact Generator

Hideki Tsunashima[1,2], Hirokatsu Kataoka[2], Junji Yamato[3], Qiu Chen[3], Shigeo Morishima[4]
1) Waseda University, 2) AIST, 3) Kogakuin University, 4) Waseda Research Institute for Science and Engineering

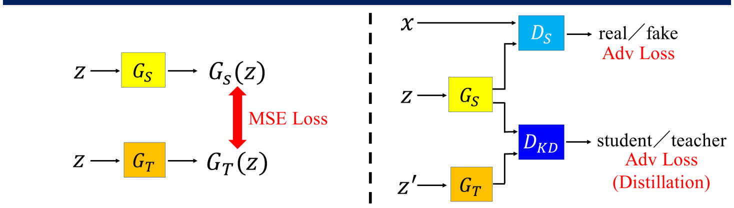
Introduction

◆ Training SOTA GANs needs large amount computation resources (e.g. StyleGAN needs 7 days on 8 Tesla V100).

◆ Conventional compression methods for GANs.
- ◻ LIT←only applied to ResNet architecture.
- ◻ MSE-method←aggravate image quality from original GANs.

◆ Our contributions
- ◻ Close image quality of a student to a teacher.
- ◻ Ours isn't constrained by network architecture.

Adversarial Knowledge Distillation for Generative models (AKDG)



(a) MSE-method (Aguinaldo et al.) (b) **AKDG (ours)**

◆ AKDG
The key concept is very simple but effective.

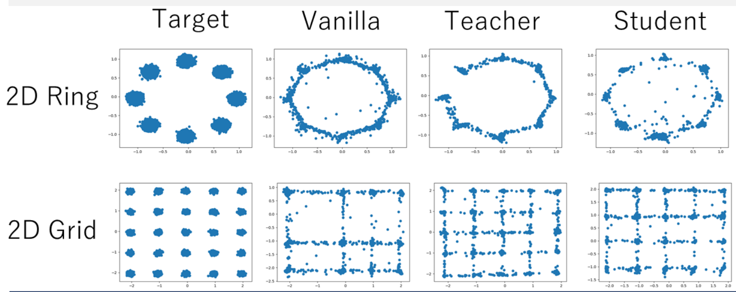
The key of AKDG is the additional discriminator which judges a teacher or student.

Results

◆ Confirming the effectiveness of AKDG on a toy dataset

◻ 2D Ring
The connection between modes is sparse in the student trained with AKDG.

◻ 2D Grid
The student does not miss the modes in comparison with the vanilla model.



◆ Compressing GANs on the real datasets.

◻ CIFAR-10

(IS: Inception Score FID: Fréchet Inception Distance)
Quality and diversity of the images generated by the student trained with AKDG is the best.
Moreover, AKDG is robust in terms of the difference of architecture.

Methods	# layers	# params	# mult-add	IS ↑	FID ↓
Teacher (SNGAN [4])	8	4.1M	793M	8.49(±0.13)	16.50
DCGAN vanilla [40]	5	2.6M(37.6% ↓)	229M(71.1% ↓)	6.86(±0.07)	31.58
DCGAN-MSE [21]	5	2.6M(37.6% ↓)	229M(71.1% ↓)	6.45(±0.07)	40.06
DCGAN-AKDG (ours)	5	2.6M(37.6% ↓)	229M(71.1% ↓)	**7.05**(±0.07)	**29.52**
MLP vanilla [40]	4	2.2M(46.3% ↓)	2.2M(99.7% ↓)	4.66(±0.02)	61.30
MLP-MSE [21]	4	2.2M(46.3% ↓)	2.2M(99.7% ↓)	3.97(±0.05)	108.92
MLP-AKDG (ours)	4	2.2M(46.3% ↓)	2.2M(99.7% ↓)	**5.14**(±0.05)	**53.73**

Teacher (Miyato et al.) Vanilla (Mescheder et al.) **AKDG (ours)**

◻ LSUN bedroom

Quality of the images generated by the student with AKDG is the best.

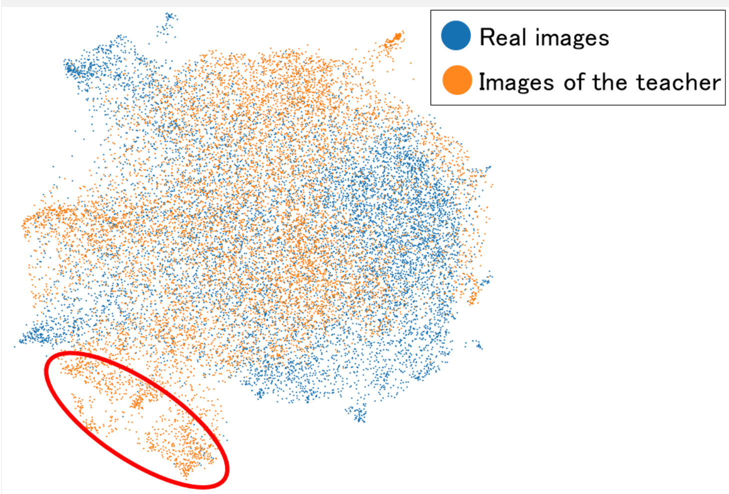
Methods	# layers	# params	# mult-add	FID ↓
Teacher (PGGAN [2])	15	18.3M	8.9B	21.22
DCGAN vanilla (batch64) [40]	8	3.7M(79.8% ↓)	785M(91.2% ↓)	44.72
DCGAN vanilla (batch2048) [40]	8	3.7M(79.8% ↓)	785M(91.2% ↓)	37.89
DCGAN-MSE [21]	8	3.7M(79.8% ↓)	785M(91.2% ↓)	98.96
DCGAN-AKDG (ours)	8	3.7M(79.8% ↓)	785M(91.2% ↓)	**27.86**

Teacher (Karras et al.) Vanilla (Mescheder et al.) **AKDG (ours)**

◆ Why does not AKDG work in case of the small batchsize?

Visualize the real images and the images of the teacher using UMAP. Red circle indicates images of the teacher which does not overlap the real images.

The distribution gap incurs the collapse of training.



Conclusion and Future Work

Conclusion
We propose the novel compression method for unconditional GANs.
In all experiments, our AKDG outperforms the conventional compression method.

Future Work
- ◆ We will fix the problem of large batchsize training.
- ◆ We will incorporate other network compression methods, such as pruning, quantization and tensor factorization.