重慶理工大學

An Adversarial Framework for Generating Unseen Images by Activation Maximization

Yang Zhang[*1], **Wang Zhou**[*2†], **Gaoyuan Zhang**[1], **David Cox**[1], **Shiyu Chang**[3]

[1]MIT-IBM Watson AI Lab, Cambridge, MA, USA

[2]Meta AI, New York, NY, USA

[3]Unversity of California at Santa Barbara, USA

AAAI-2022

汇报人：宋小雪

2022.12.3

重慶理工大學

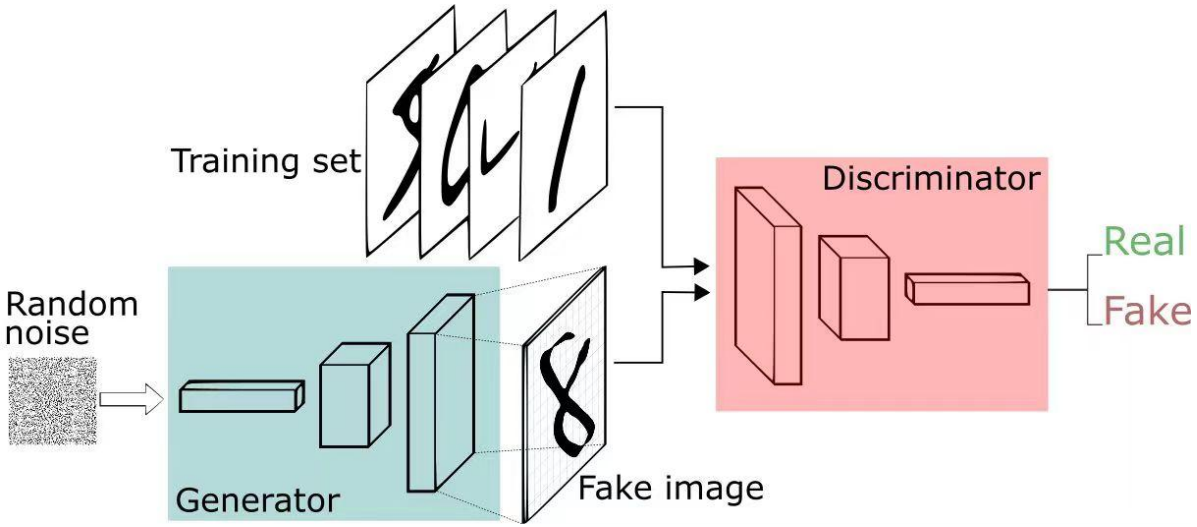
OUTLINE

1 Introduction

2 Method

3 Experience

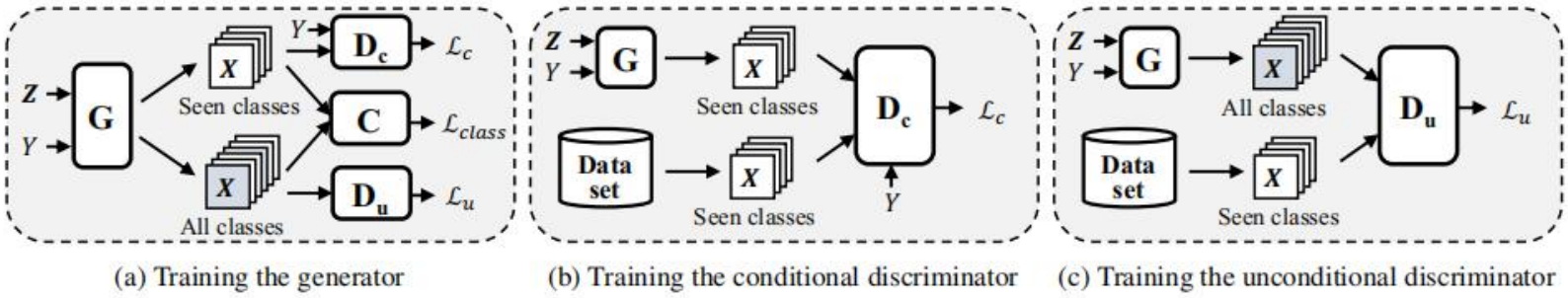
GAN



AM:maximize the activation of a classifier, so that the generated examples conform to the class characteristics as depicted by the classifier.

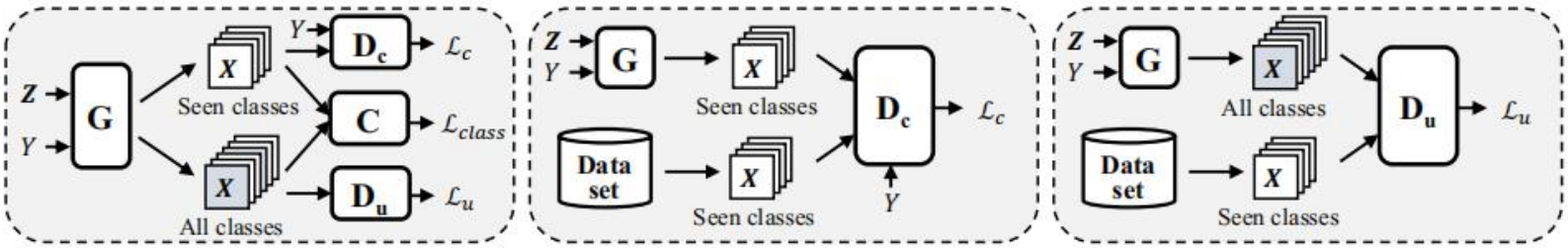
GAN-based AM

PROBEGAN



(a) Training the generator (b) Training the conditional discriminator (c) Training the unconditional discriminator

Figure 1: PROBEGAN framework and the data flow when training different modules. G represents the generator; D_c represents the conditional discriminator; D_u represents the unconditional discriminator; C represents the classifier. Class dilution (mixing the target class into other classes) is performed on the input to the unconditional discriminator.



(a) Training the generator (b) Training the conditional discriminator (c) Training the unconditional discriminator

$$\min_{G(\cdot)} \max_{D_c(\cdot,\cdot),D_u(\cdot)} \sum_{y\in\mathcal{Y}\setminus y^*} \mathcal{L}_c(y) + \mathcal{L}_u + \lambda_g \sum_{y\in\mathcal{Y}} \mathcal{L}_{class}(y), \quad (1)$$

$$\begin{aligned}\mathcal{L}_c(y) =& \mathbb{E}_{\boldsymbol{X}|Y=y}[\min(0, -1 + D_c(X, y))] \\ & - \mathbb{E}_{\boldsymbol{Z}}[\min(0, -1 - D_c(G(\boldsymbol{Z}, y), y))].\end{aligned} \quad (2)$$

$$\mathcal{L}_u = \mathbb{E}_{\boldsymbol{X}|Y\neq y^*}[D_u(X)] - \mathbb{E}_{\boldsymbol{Z},Y}[D_u(G(\boldsymbol{Z}, Y))] + \lambda_{gp}\mathcal{L}_{gp}, \quad (3)$$

$$\mathcal{L}_{class}(y) = -\mathbb{E}_{\boldsymbol{Z}}[\log \hat{p}_{Y|\boldsymbol{X}}(y|G(\boldsymbol{Z}, y))]. \quad (4)$$

Figure 2: Sample generated images of "horse" (left) and "truck" (right) by (a) BIGGAN-AM-regular, (b) BIGGAN-AM-robust, (c) NAIVE-regular, (d) NAIVE-robust, (e) PROBEGAN-regular, and (f) PROBEGAN-robust (marked red).

Dataset	Network	FID ↓	intra-FID ↓									
			plane	auto	bird	cat	deer	dog	frog	horse	ship	truck
CIFAR10	BigGAN	7.99	29.05	13.14	26.73	24.23	16.25	26.25	24.08	14.20	14.64	17.37
	PROBEGAN-oracle	6.67	26.41	12.93	27.12	26.39	15.18	22.42	18.00	13.69	15.95	14.03
w/o horse	BIGGAN-AM-regular	168.4	-	-	-	-	-	-	-	227.3	-	-
	NAIVE-regular	31.70	-	-	-	-	-	-	-	101.7	-	-
	PROBEGAN-regular	8.99	23.38	11.60	23.98	26.45	14.24	23.44	16.69	91.35	13.31	13.56
	BIGGAN-AM-robust	161.6	-	-	-	-	-	-	-	223.3	-	-
	NAIVE-robust	48.92	-	-	-	-	-	-	-	76.02	-	-
	PROBEGAN-robust	8.39	26.13	12.60	24.28	27.08	15.63	24.53	17.69	**45.40**	14.83	14.02
w/o truck	BIGGAN-AM-regular	114.1	-	-	-	-	-	-	-	-	-	179.3
	NAIVE-regular	33.36	-	-	-	-	-	-	-	-	-	118.5
	PROBEGAN-regular	8.71	24.30	12.64	23.89	25.45	13.31	22.30	16.75	13.88	14.12	105.99
	BIGGAN-AM-robust	99.20	-	-	-	-	-	-	-	-	-	161.5
	NAIVE-robust	56.47	-	-	-	-	-	-	-	-	-	84.21
	PROBEGAN-robust	8.80	27.70	14.63	25.62	27.02	14.99	23.27	17.89	15.08	15.54	**68.33**

Table 1: FID results on CIFAR-10. Gray background indicates the unseen class. Results for BigGAN is from our reimplementation, which is better than that is reported in Brock, Donahue, and Simonyan (2019).

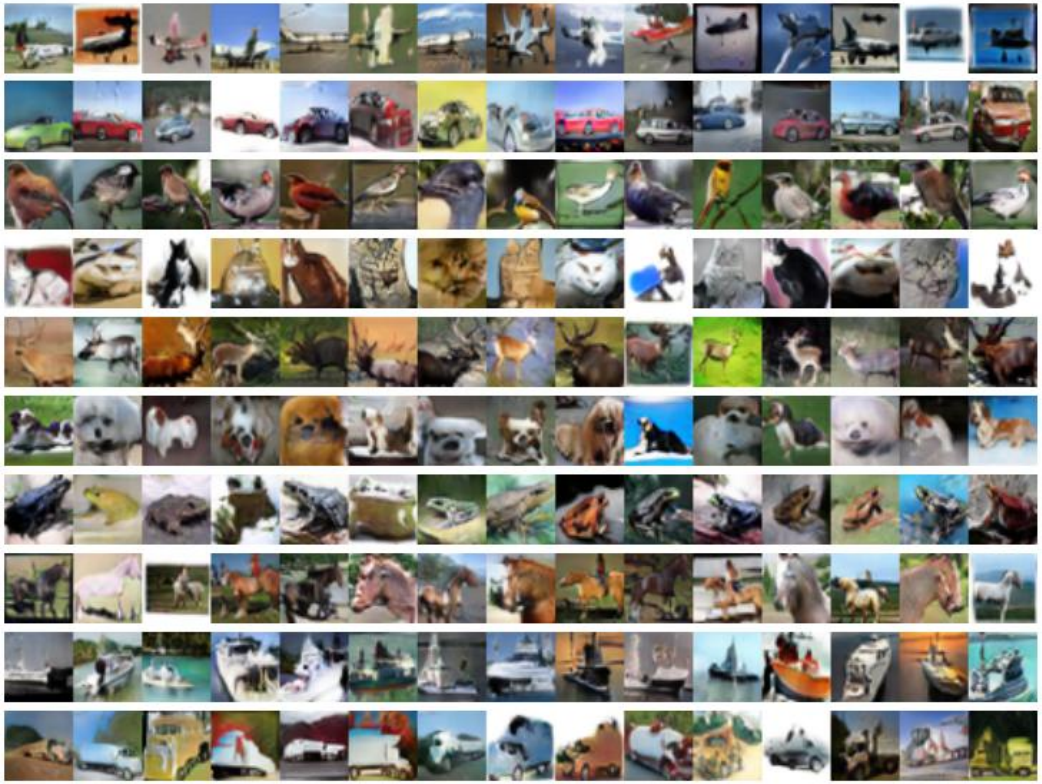


Figure 3: Sample images of CIFAR-10 classes generated by PROBEGAN-robust. Each row corresponds to one unseen class setting, which are, from top to bottom, plane, auto, bird, cat, deer, dog, frog, horse, ship, and truck.

Figure 4: Sample generated images of “truck” when images of “truck” with artificial red blocks are present while other classes remain unchanged.

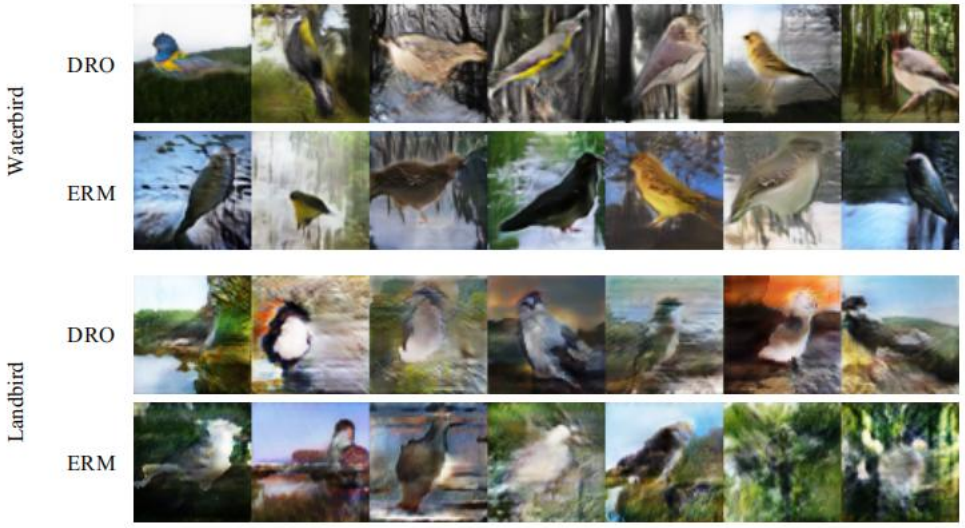


Figure 5: Samples of Waterbird and Landbird with classifier trained using DRO or ERM, respectively. When generating images of waterbirds, only images of landbirds on land background are used to avoid information leak, and vice versa.

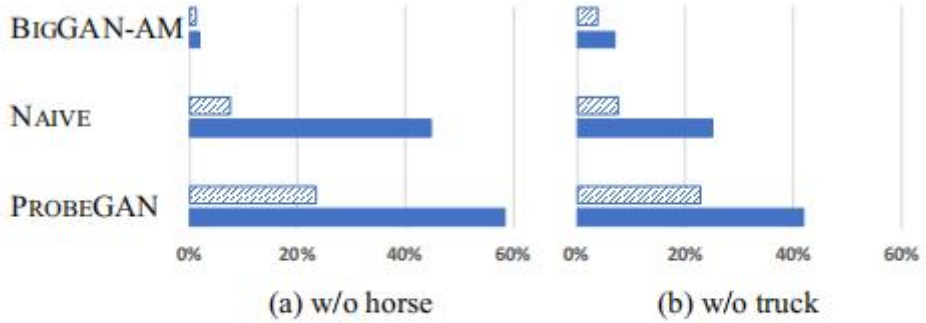


Figure 6: Human recognition rate by MTurk. The dashed bars represent the results using a regular classifier, while the solid bars with a robust classifier.

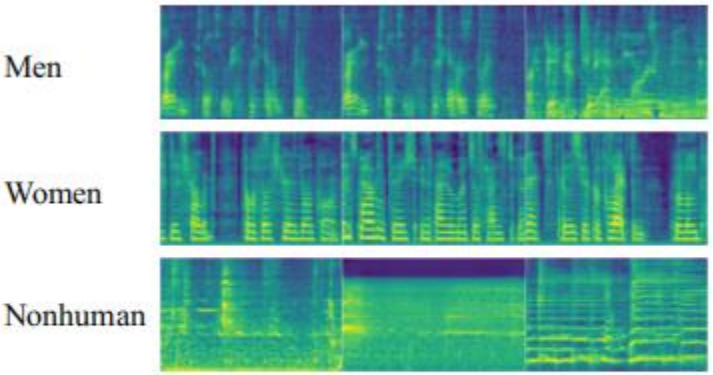


Figure 7: Sample mel-spectrograms of generated audio clips when each of the classes, men, women, and nonhuman, is taken as the "unseen" class.

重慶理工大學

Thank !