NICE: NoIse-modulated Consistency rEgularization for Data-Efficient GANs

Yao Ni[†] Piotr Koniusz[§,†]

[†]The Australian National University
[§]Data61♥CSIRO

NeurIPS 2023

Background: Challenges in training GANs on limited data

- Discriminator overfitting on limited training data.
- Training instability.

Goal: To improve the generalization of GAN.

Methods: Generalization error of GAN

n: dataset size. $\mathcal{H}/\mathcal{G}$: discriminator/generator sets. $\forall h \in \mathcal{H}$, $\|h\|_\infty \leq \Delta$. μ/ν: measures on real/fake data. $\hat{\mu}_n/\nu_n$: empirical measures. Assume $d_{\mathcal{H}}(\hat{\mu}_n, \nu_n) - \inf_{\nu \in \mathcal{G}} d_{\mathcal{H}}(\hat{\mu}_n, \nu) \leq \epsilon$.

$$\underbrace{d_{\mathcal{H}}(\mu, \nu_n) - \inf_{\nu \in \mathcal{G}} d_{\mathcal{H}}(\mu, \nu)}_{\text{How far the fake data is from the real unseen data.}} \leq \underbrace{2 \sup_{h \in \mathcal{H}} \left| \mathbb{E}_\mu[h] - \mathbb{E}_{\hat{\mu}_n}[h] \right|}_{\text{Discrepancy between seen and unseen real data.}} + \epsilon$$

$$\leq \underbrace{2R_n^{(\mu)}(\mathcal{H})}_{\text{Rademacher complexity of the discriminator.}} + 2\Delta \sqrt{\frac{2 \log(1/\delta)}{n}} + \epsilon$$

Lower Rademacher complexity of discriminator → better generalization ☺

Methods: Rademacher complexity of a neural network

For $\forall i \in \{1, ..., n\}$, $\|\boldsymbol{x}^{(i)}\|_2 \leq q$ and a t-layer fully-connected network parameterized from set $\mathcal{V} = \{v_{\boldsymbol{\theta}} : \|\boldsymbol{W}_i\|_{\text{lip}} \leq k_i, \|\boldsymbol{W}_i^T\|_{2,1} \leq b_i\}$:

$$R_n^{(\mu)}(\mathcal{V}) \leq \frac{q}{\sqrt{n}} \cdot \left(\prod_{i=1}^{t} k_i\right) \cdot \left(\sum_{i=1}^{t} \frac{\overbrace{b_i^{2/3}}^{\text{Weight norm.}}}{k_i^{2/3}}\right)^{3/2}$$

Smaller weight norms → lower complexity → better generalization ☺

Methods: Regularization through multiplicative noise

$\boldsymbol{w}_k$: the k-th column vector of the second layer weight $\boldsymbol{W}_2$. $\hat{a}_k$: mean feature norm ≥ 0. β^2: variance of noise. $\boldsymbol{y}$: label. Multiplicative noise modulation $\boldsymbol{z}$ on the latent feature $\boldsymbol{a}^{(i)}$ in a two-layer net induces weight regularization.

$$
\begin{aligned}
\hat{L}_{\text{noise}}(w) : &= \hat{\mathbb{E}}_i \mathbb{E}_{\boldsymbol{z}} \left[\|\boldsymbol{y}^{(i)} - \boldsymbol{W}_2 (\overbrace{\boldsymbol{z} \odot \boldsymbol{a}^{(i)}}^{\text{Noise modulation with latent feature.}}) \|_2^2 \right] \\
&= \hat{\mathbb{E}}_i \left[\|\boldsymbol{y}^{(i)} - \boldsymbol{W}_2 \boldsymbol{a}^{(i)} \|_2^2 \right] + \underbrace{\beta^2 \sum_k \overbrace{\hat{a}_k}^{\text{Mean feature norm}\geq 0.} \|\boldsymbol{w}_k\|_2^2}_{\text{Implicit regularization on } \|\boldsymbol{w}_k\|_2.}
\end{aligned}
$$

Noise modulation $\rightarrow$ smaller weight norms $\rightarrow$ better generalization ☺

Methods: Noise incurs gradient issue

Noise modulation has the potential to amplify gradient

$$\min_{\boldsymbol{\theta}_d} L_D^{\mathrm{AN}} := \mathbb{E}_{\tilde{\boldsymbol{a}}}\mathbb{E}_{\boldsymbol{z}}\big[h(\boldsymbol{z}\odot\tilde{\boldsymbol{a}})\big] - \mathbb{E}_{\boldsymbol{a}}\mathbb{E}_{\boldsymbol{z}}\big[h(\boldsymbol{z}\odot\boldsymbol{a})\big]$$

$$\approx \mathbb{E}_{\tilde{\boldsymbol{a}}}\big[h(\tilde{\boldsymbol{a}})\big] - \mathbb{E}_{\boldsymbol{a}}\big[h(\boldsymbol{a})\big] + \frac{\beta^2}{2}\Big(\mathbb{E}_{\tilde{\boldsymbol{a}}}\big[\textstyle\sum_k \tilde{a}_k^2 H_{kk}^{(h)}(\tilde{\boldsymbol{a}})\big] \boxed{-\mathbb{E}_{\boldsymbol{a}}\big[\textstyle\sum_k a_k^2 H_{kk}^{(h)}(\boldsymbol{a})\big]\Big)}$$

$$\min_{\boldsymbol{\theta}_g} L_G^{\mathrm{AN}} := -\mathbb{E}_{\boldsymbol{z}}\mathbb{E}_{\tilde{\boldsymbol{a}}}\big[h(\boldsymbol{z}\odot\tilde{\boldsymbol{a}})\big] \approx -\mathbb{E}_{\tilde{\boldsymbol{a}}}\big[h(\tilde{\boldsymbol{a}})\big] \boxed{-\frac{\beta^2}{2}\mathbb{E}_{\tilde{\boldsymbol{a}}}\big[\textstyle\sum_k \tilde{a}_k^2 H_{kk}^{(h)}(\tilde{\boldsymbol{a}})\big]}$$

$\boldsymbol{a}$: real feature, $\tilde{\boldsymbol{a}}$: fake feature.
$H^{(h)}(\boldsymbol{a})$: Hessian matrix of discriminator h evaluated at $\boldsymbol{a}$.

Noise modulation $\rightarrow$ greater gradient norms $\rightarrow$ unstable training ☹

Methods: Consistency regularization

Enforces the discriminator to be consistent for same input under different noises.

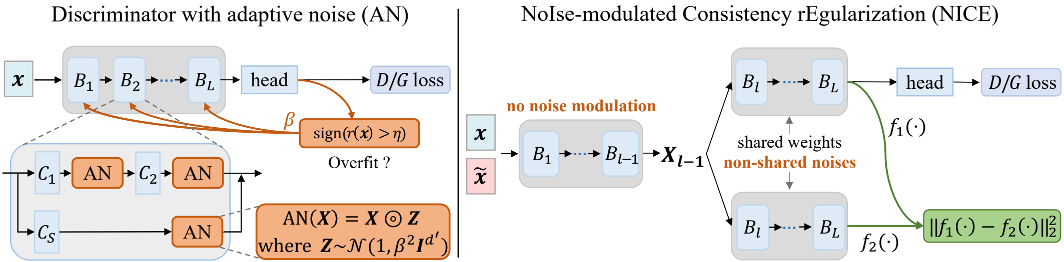
$$\begin{aligned}\ell^{\text{NICE}}(\boldsymbol{a}) &:= \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\Big[\big(f(\boldsymbol{z}_1 \odot \boldsymbol{a}) - f(\boldsymbol{z}_2 \odot \boldsymbol{a})\big)^2\Big] \\ &\approx \boxed{2\beta^2 \textstyle\sum_k a_k^2 \nabla_k^2 f(\boldsymbol{a}) + \beta^4 \sum_{j,k} a_j^2 a_k^2 (H_{jk}^{(f)}(\boldsymbol{a}))^2}\end{aligned}$$

$\nabla f(\boldsymbol{a})$, $H^{(f)}(\boldsymbol{a})$: gradient and Hessian matrix of feature extractor f evaluated at $\boldsymbol{a}$.

NICE $\approx$ Gradient penalization $\rightarrow$ smaller gradient norms ☺

NICE: weight regularization $\rightarrow$ smaller weight norms $\rightarrow$ better generalization
NICE: gradient penalization $\rightarrow$ smaller gradient norms $\rightarrow$ stable training

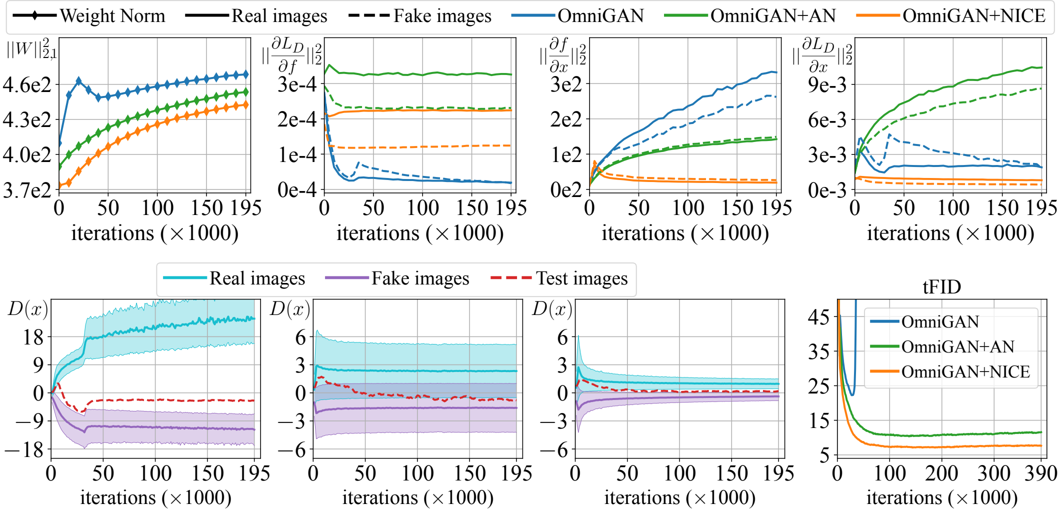
Pipeline



d': feature dim. $\circledcirc$: expands $\boldsymbol{Z}$ to $d' \times d^H \times d^W$ then performs element-wise product. B_l:l-th block. C_S: Conv. in skip branch. f: feat. extractor. $\boldsymbol{x}/\tilde{\boldsymbol{x}}$: real/fake image. η: a threshold.

Update β: control the variance of noise by monitoring $r(\boldsymbol{x}) = \mathbb{E}[\text{sign}(D(\boldsymbol{x}))]$.
Update $\beta_{t+1} = \beta_t + \Delta_\beta \cdot \text{sign}\left(r(\boldsymbol{x}) > \eta\right)$.

Experiments: Analysis



(a) OmniGAN (b) OmniGAN+AN (c) OmniGAN+NICE (d) tFID curves

Experiments: Results

Method	CIFAR-10			CIFAR-100		
	100% data	20% data	10% data	100% data	20% data	10% data
	IS↑/tFID↓	IS↑/tFID↓	IS↑/tFID↓	IS↑/tFID↓	IS↑/tFID↓	IS↑/tFID↓
BigGAN	9.21/5.48	8.74/16.20	8.24/31.45	11.02/7.86	9.94/25.83	7.58/50.79
+NICE	**9.50/4.19**	**8.96/8.51**	**8.73/13.65**	10.99/**6.31**	**10.32**/**13.17**	**8.96/19.53**
LeCam+DA	9.45/4.32	9.01/8.53	8.81/12.64	11.25/6.45	10.12/15.96	9.17/22.75
+NICE	**9.52/3.72**	**9.12/6.92**	**8.99/9.86**	**11.28/5.72**	**10.54/10.02**	**9.35/14.95**
OmniGAN+ADA	10.24/4.95	9.41/27.04	7.86/40.05	13.07/6.12	12.07/13.54	8.95/44.65
+NICE	**10.38/2.25**	**10.18/4.39**	**10.08/5.49**	**13.82/3.78**	**12.75/6.28**	**12.04/9.32**

Method	FID ↓ on ImageNet		
	10%	5%	2.5%
BigGAN	38.30	91.16	133.80
ADA	31.89	43.21	56.83
DA	32.82	56.75	63.49
MaskedGAN	26.51	35.70	38.62
KDDLGAN	20.32	22.35	28.79
NICE	21.44	24.72	31.45
ADA+NICE	**18.29**	**20.07**	**24.41**

Method (FID↓)	Obama	GrumpyCat	Panda	AnimalCat	AnimalDog
StyleGAN2	80.20	48.90	34.27	71.71	131.90
StyleGAN2+NICE	**24.56**	**18.78**	**8.92**	**25.25**	**46.56**
ADA	45.69	26.62	12.90	40.77	56.83
LeCam+KDDLGAN	29.38	19.65	8.41	31.89	50.22
ADA+NICE	**20.09**	**15.63**	**8.18**	**22.70**	**28.65**

Method (FID↓ on FFHQ)	100	$1K$	$2K$	$5K$
StyleGAN2	179	100.16	54	49.68
ADA	85.8	21.29	15.39	10.96
ADA-Linear	82	19.86	13.01	9.39
InsGen	45.75	18.21	11.47	7.83
FakeCLR	42.56	15.92	9.90	7.25
ADA+NICE	**38.42**	**14.57**	**8.85**	**6.48**

Conclusions

- The noise modulation **regularizes the weight norm** $\rightarrow$ improved generalization.
- The consistency regularization **penalizes the gradient norm** $\rightarrow$ stable GAN training.