# Pairwise Augmented GANs with Adversarial Reconstruction Loss

Aibek Alanov[1,2,3,*] , Max Kochurov[1,3,*] , Daniil Yashkov[5], Dmitry Vetrov[2,3,4]

[1]Samsung AI Center in Moscow
[2]National Research University Higher School of Economics
[3]Skolkovo Institute of Science and Technology
[4]Joint Samsung-HSE lab
[5]FRC "Informatics and Management" of the Russian Academy of Sciences

December 5, 2018

# Contents

# Generative Adversarial Networks (GANs)

**Input:** $x_1, \dots, x_n$ - real samples from $p^*(x)$

**GAN:**

- generator $G_\theta : z \to x,\ z \sim p(z)$ - samples objects from a noise
- discriminator $D_\psi : x \to [0, 1]$ - classifies real objects from generated ones

**Goal:** match the generator's distribution $p_\theta(x)$ to $p^*(x)$

**Discriminator's objective:**

$$\mathbb{E}_{p^*(x)} \log D_\psi(x) + \mathbb{E}_{p(z)} \log(1 - D_\psi(G_\theta(z))) \quad \to \quad \max_\psi$$

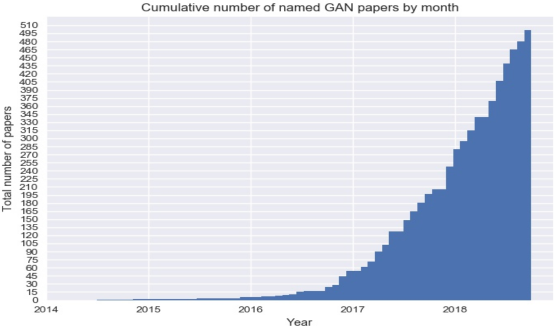
**Generator's objective:**

$$\mathbb{E}_{p(z)} \log D_\psi(G_\theta(z)) \quad \to \quad \max_\theta$$

# GAN Advantages

The idea of adversarial learning is **very fruitful**:

- To date, there are more than 500 different GAN models[1]
- Many applications in computer vision
- Around 6000 cites to the original paper of Goodfellow et al.



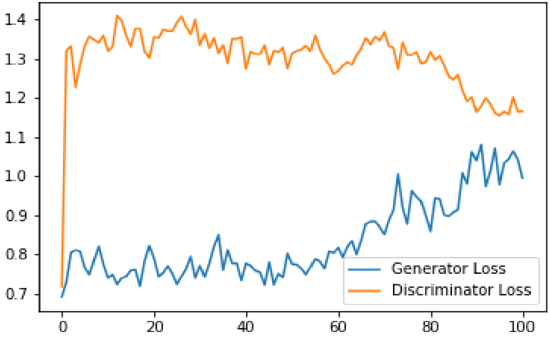

[1] https://github.com/hindupuravinash/the-gan-zoo

# GAN Advantages

GAN generates **high quality** images

It is **hard** to train:

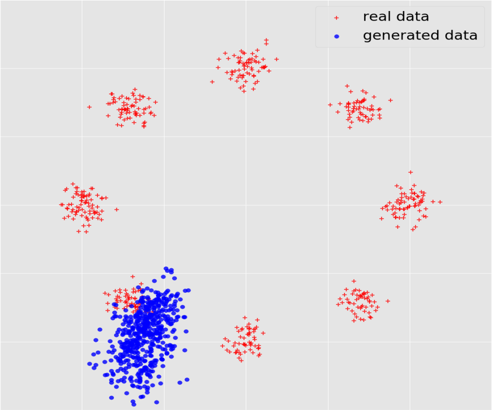
- training process can be unstable
- there is no stopping criteria except for a visual judgement



# GAN Drawbacks

**Mode collapsing** problem:

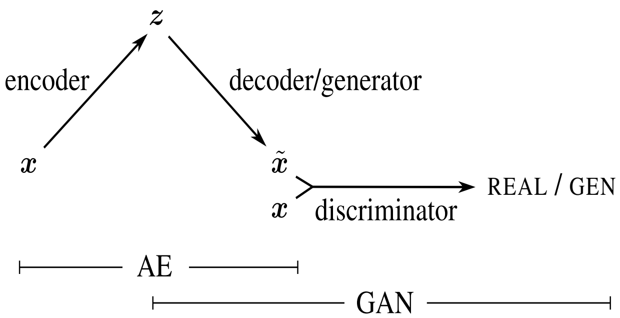
- generator samples only a small subset of training dataset



## GAN Drawbacks

There is no **inverse mapping**:

- there is no **encoder** which maps the generated image to the corresponding noise vector
- such auto-encoding property has many applications, e.g. image editing, image inpainting, etc.



# Introducing Encoder Part

Encoder $E_\varphi : x \to z$ maps input image to the corresponding latent vector.

**Objective** for the encoder: to have **good reconstructions**, i.e.,

$$G_\theta(E_\varphi(x)) \approx x$$

# Reconstruction Loss

Standard reconstruction losses:

- $\|x - y\|_2^2$ - $L_2$ loss;
- $\|x - y\|_1^2$ - $L_1$ loss;
- $\|\Phi(x) - \Phi(y)\|_2^2$ - perceptual loss where $\Phi(\cdot)$ is the output of intermediate layers of a pretrained network (e.g. VGG)

Many bidirectional GANs use them:

- AGE,
- $\alpha$-GANs,
- Cycle-GANs,
- ALICE,
- MINE,
- SVAE

# Drawbacks of Standard Losses

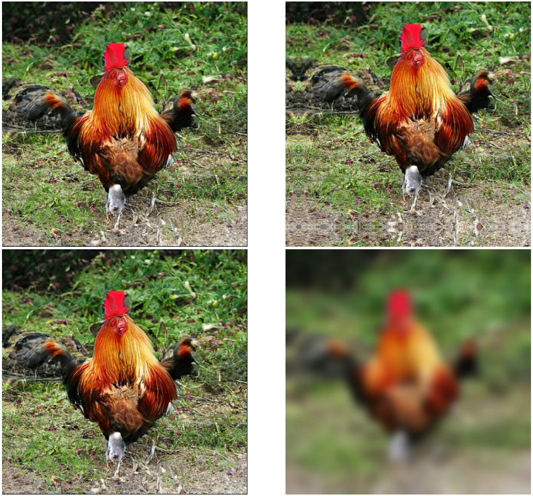


Figure: First column is original, second is augmentation

# Drawbacks of Standard Losses

| Loss | Blur | Pad + crop |
|---|---|---|
| $L_1$ | 0.21 | 0.4 |
| $L_2$ | 0.074 | 0.26 |
| Perceptual-123 | 2.24 | 3.52 |
| Perceptual-345 | 9.02 | 13.79 |

# Drawbacks of $L_1$ and $L_2$

- The space of pixels is very noisy and does not capture the perceptual similarity of images
- $L_1$ and $L_2$ encourage the exact coincidence of images rather than a content-wise similarity
- $L_1$ and $L_2$ enforce auto-encoding model to recover too many unnecessary details of the source object

# Drawbacks of Perceptual Loss

- The choice of intermediate layers and their weights is heuristic
- First layers have the same problems as $L_1$ and $L_2$, deep layers lose local details of the image
- Necessity of an additional pretrained network

# Augmentation Function

An augmentation function $a(\cdot) : x \to y$ is a stochastic transformation of input image

Examples:

- Gaussian blur;
- contrast;
- combination of padding and random crop

Figure: Original, Blur, Contrast, Pad+Crop

# Conditional Distributions

Mappings $G_\theta(z)$, $E_\varphi(x)$ and $a(x)$ induce the following conditional distributions:

- $p_\theta(x|z)$ over outputs of the generator $G_\theta(z)$ given $z$;
- $q_\varphi(z|x)$ over outputs of the encoder $E_\varphi(x)$ given $x$;
- $r(y|x)$ over the augmentations $a(x)$ given a source object $x$.

Two classes of pairs:

- **real** class: $(x, y)$ from $p^*(x)r(y|x)$, i.e., $x$ is real, $y = a(x)$ is its augmentation;
- **fake** class: $(x, y)$ from $p^*(x)p_{\theta,\varphi}(y|x) = p^*(x)\int p_\theta(y|z)q_\varphi(z|x)dz$, i.e., $x$ is real, $y = G_\theta(E_\varphi(x))$ is its reconstruction

Figure: Left - real pair, right - fake pair

# Discriminator on Pairs

Discriminator $D_\tau(x, y)$ classifies mentioned two classes of pairs.

**Discriminator's objective:**

$$\mathbb{E}_{p^*(x)r(y|x)} \log D_\tau(x, y) + \mathbb{E}_{p^*(x)p_{\theta,\varphi}(y|x)} \log(1 - D_\tau(x, y)) \to \max_\tau$$

**Generator's objective:**

$$\mathbb{E}_{p^*(x)p_{\theta,\varphi}(y|x)} \log D_\tau(x, y) \quad \to \quad \max_\theta$$

**Encoder's objective:**

$$\mathbb{E}_{p^*(x)p_{\theta,\varphi}(y|x)} \log D_\tau(x, y) \quad \to \quad \max_\varphi$$

It is crucial to use augmentation pairs!

# Matching Encoder to Prior

- Outputs of $E_\varphi(x)$ for real images can be very far from the prior distribution $p(z)$.
- $G_\theta$ should generate good images both for samples from the prior $p(z)$ and for outputs of $E_\varphi$.
- As a result, it will lead to unstable training of $G_\theta$

Therefore we introduce the third discriminator $D_\zeta(z)$ for matching $E_\varphi$ to the prior $p(z)$.

**Discriminator's objective:**

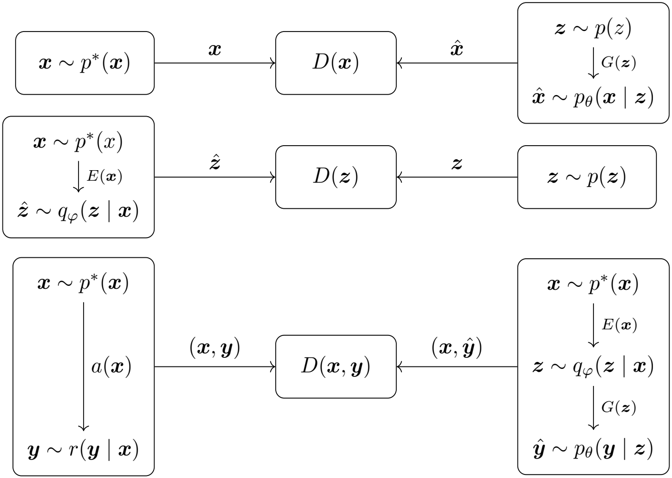
$$\mathbb{E}_{p(z)} \log D_\zeta(z) + \mathbb{E}_{p^*(x)} \log(1 - D_\zeta(E_\varphi(x))) \quad \to \quad \max_\zeta$$

**Encoder's objective:**

$$\mathbb{E}_{p^*(x)} \log D_\zeta(E_\varphi(x))) \quad \to \quad \max_\varphi$$

# PAGAN Diagram

The diagram of Pairwise Augmented GAN (PAGAN) model:



# PAGAN Algorithm

**Algorithm 1** The PAGAN training algorithm.

$\theta, \varphi, \psi, \zeta, \tau \leftarrow$ initialize network parameters

**repeat**

$\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)} \sim p^*(\boldsymbol{x})$ ▷ Draw $N$ samples from the dataset and the prior

$\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(N)} \sim p(\boldsymbol{z})$

$\hat{\boldsymbol{z}}^{(i)} \sim q_\varphi(\boldsymbol{z} \mid \boldsymbol{x} = \boldsymbol{x}^{(i)}), \quad i = 1, \dots, N$ ▷ Sample from the conditionals

$\boldsymbol{x}_{pr}^{(j)} \sim p_\theta(\boldsymbol{x} \mid \boldsymbol{z} = \boldsymbol{z}^{(j)}), \quad j = 1, \dots, N$

$\boldsymbol{x}_{rec}^{(i)} \sim p_\theta(\boldsymbol{x} \mid \boldsymbol{z} = \hat{\boldsymbol{z}}^{(i)}), \quad j = 1, \dots, N$

$\boldsymbol{x}_{aug}^{(i)} \sim r(\boldsymbol{y} \mid \boldsymbol{x} = \boldsymbol{x}^{(i)}), \quad j = 1, \dots, N$

$\mathcal{L}_d^x \leftarrow -\frac{1}{N}\sum_{i=1}^N \log D(\boldsymbol{x}^{(i)}) - \frac{1}{N}\sum_{j=1}^N \; log\left(1 - D(\boldsymbol{x}_{pr}^{(j)})\right)$ ▷ Compute discriminator loss

$\mathcal{L}_d^z \leftarrow -\frac{1}{N}\sum_{i=1}^N \log D(\boldsymbol{z}^{(i)}) - \frac{1}{N}\sum_{j=1}^N \; log\left(1 - D(\hat{\boldsymbol{z}}^{(j)})\right)$

$\mathcal{L}_d^{xx} \leftarrow -\frac{1}{N}\sum_{i=1}^N \log D(\boldsymbol{x}^{(i)}, \boldsymbol{x}_{aug}^{(i)}) - \frac{1}{N}\sum_{j=1}^N \; log\left(1 - D(\boldsymbol{x}^{(j)}, \boldsymbol{x}_{rec}^{(j)})\right)$

$\mathcal{L}_g \leftarrow -\frac{1}{N}\sum_{i=1}^N \log D(\boldsymbol{x}_{pr}^{(i)}) - \frac{1}{N}\sum_{j=1}^N \log D(\boldsymbol{x}^{(j)}, \boldsymbol{x}_{rec}^{(j)})$ ▷ Compute generator loss

$\mathcal{L}_e \leftarrow -\frac{1}{N}\sum_{i=1}^N \log D(\hat{\boldsymbol{z}}^{(i)}) - \frac{1}{N}\sum_{j=1}^N \log D(\boldsymbol{x}^{(j)}, \boldsymbol{x}_{rec}^{(j)})$ ▷ Compute encoder loss

$\psi \leftarrow \psi - \nabla_\psi \mathcal{L}_d^x, \; \zeta \leftarrow \zeta - \nabla_\zeta \mathcal{L}_d^z$ ▷ Gradient update on discriminator networks

$\tau \leftarrow \tau - \nabla_\tau \mathcal{L}_d^{xx}$

$\theta \leftarrow \theta - \nabla_\theta \mathcal{L}_g, \; \varphi \leftarrow \varphi - \nabla_\varphi \mathcal{L}_e$ ▷ Gradient update on generator-encoder networks

**until** convergence

# Samples and Reconstructions

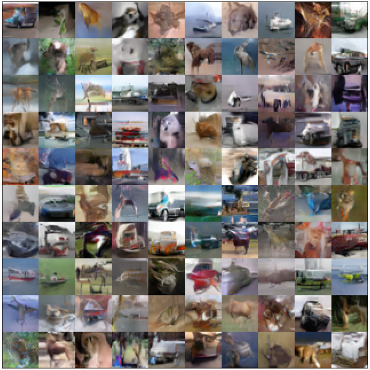


Figure: Samples and reconstructions of PAGAN model for CIFAR10 dataset.

# Inception Score, Fréchet Inception Distance (FID)

| Model | FID | Inception Score |
|---|---|---|
| WAE-GAN | 87.7 | $4.18 \pm 0.04$ |
| ALI | | $5.34 \pm 0.04$ |
| AGE | 39.51 | $5.9 \pm 0.04$ |
| ALICE | | $6.02 \pm 0.03$ |
| S-VAE | | 6.055 |
| $\alpha$-GANs | | 6.2 |
| AS-VAE | | 6.3 |
| PD-WGAN | 33.0 | $\mathbf{6.70} \pm \mathbf{0.09}$ |
| PAGAN (ours) | **32.84** | $6.56 \pm 0.06$ |

# Reconstruction Inception Dissimilarity

- As we showed, standard reconstruction losses are not good metric for evaluating reconstruction quality
- We introduced a novel metric Reconstruction Inception Dissimilarity (RID) which is based on a pre-trained classification network:

$$RID = \exp\left\{\mathbb{E}_{x \sim \mathcal{D}} D_{\mathrm{KL}}(p(y|x) \| p(y|G(E(x))))\right\}$$

where $p(y|x)$ is a pre-trained classifier that estimates the label distribution given an image.

# RID Results

| Model | RMSE | RID |
|---|---|---|
| AUG | 8.89 | $1.57 \pm 0.02$ |
| VAE | 5.85 | $44.33 \pm 2.27$ |
| SVAE | 8.59 | $38.13 \pm 1.92$ |
| AGE | 6.675 | $19.02 \pm 0.84$ |
| PAGANs | 8.12 | $\mathbf{13.01} \pm \mathbf{0.82}$ |

# Ablation Study

| Model | FID | IS | RID |
|---|---|---|---|
| PAGAN | **32.84** | $\mathbf{6.56 \pm 0.06}$ | $\mathbf{13.01 \pm 0.82}$ |
| PAGAN-L1 | 76.73 | $4.46 \pm 0.03$ | $30.94 \pm 1.58$ |
| PAGAN-NOAUG | 111.151 | $4.23 \pm 0.06$ | $50.15 \pm 2.71$ |

# Choice of Augmentation

| Augmentation | | IS | FID | RID |
|---|---|---|---|---|
| crop+padding | 0 | 3.35±0.03 | 108.81 | |
| | 0.05 | 5.62±0.01 | 45.60 | 14.70±1.08 |
| | 0.1 | **6.56±0.09** | **37.20** | 12.75±0.75 |
| | 0.15 | 6.16±0.03 | 39.38 | **12.25±0.71** |
| | 0.2 | 6.16±0.19 | 39.18 | 13.86±0.72 |
| Blur | | 2.15±0.01 | 200.66 | 32.92±1.46 |
| Contrast | | 4.18±0.01 | 101.27 | 50.02±2.10 |

# Conclusion

- We propose a novel auto-encoding generative model
- We introduce an augmented adversarial loss based on the discriminator on pairs
- We propose Reconstruction Inception Dissimilarity as an alternative metric for evaluating reconstruction quality
- Our model shows good results on sampling from the prior and on encoding real images