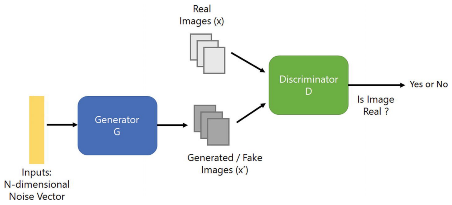
GANs: Advanced Topics

5 марта 2020 г.

Contents

Previous Episodes

Reminder: vanilla GAN



- game between Generator and Discriminator;
- uses backprop for turn-by-turn optimisation;
- optimises Jensen-Shannon-like divergence.

vanilla GAN: main problems

- Stability of training:
  - each step needs optimal discriminator;
  - imbalances in generator/discriminator (in architecture and instance performance).
- Mode collapse:
  - connected to optimisation (JS and others);
  - connected to manifold character.
- Vanishing gradients:
  - JS divergence specific problems.

f-GAN

- The framework of distribution learning $\min_Q D(Q, P)$ includes many examples. In particular the specific form:

$$\min_{\theta} \max_{T} \mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q_\theta}[f(T(x))]$$

  includes examples such as the original GAN, f-divergence GAN, MMD-GAN. Each model has its tradeoff.
- Using different f-divergence leads to very different learning dynamics.
- f-divergence generalizes learning using information theoretic divergence.

Wassertein GAN

- WGAN allowed to use really deep models for image generation. Some tricks still need to be done:
  - Gradient penalty.
  - Spectral normalisation.
- Some things are still missing, like unbiased gradients.

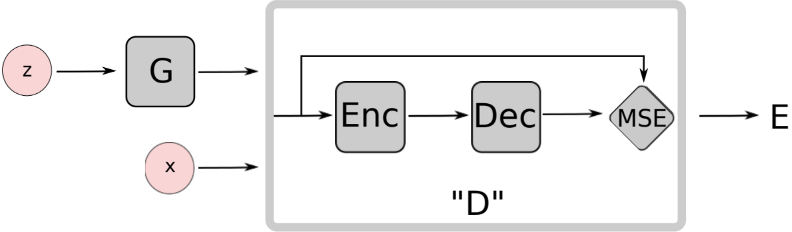
Energy Based GAN

Alternative training

- What else can be done?
- Should G and D have the same architecture?
- We can try a different style D?

Autoencoder Discriminator

- We use Autoencoder Discriminator.



- Thus, discriminator changes to $D(x) = ||Dec(Enc(x)) - x||$, MSE term.
- This does not provide probability, but an ”energy”.

From: J. Zhao et al., Energy-based Generative Adversarial Networks

Objective function: energy based

- For $[.]^+ = \max(0, .)$:

$$\mathcal{L}_D(x,z) = D(x) + [m - D(G(z))]^+;$$
$$\mathcal{L}_G(x,z) = D(G(z)),$$

- a autoencoder: reconstruction cost $D(x)$ for real images is low;
- a critic: to penalize the discriminator if the reconstruction error for generated images drops below a value m;
- first training steps are used for $D(x)$ only (due to likelihood normalisation);
- problem: the process stops being probabilistic, rather energy based.

Repelling regularizer

- A known problem of autoencoder - possibility to have identity transformation;
- we can address it using pulling-away term:

$$f_{PT} = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \frac{S_i^T S_j}{||S_i||||S_j||}.$$

where $S \in \mathbb{R}^{s \times N}$ a batch of sample of size N representations taken from the encoder output layer.

EBGAN:results

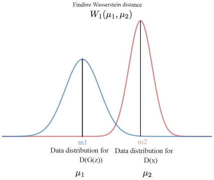
Figure 4: Generation from the grid search on MNIST. Left(a): Best GAN model; Middle(b): Best EBGAN model. Right(c): Best EBGAN-PT model.

Boundary equilibrium GAN

Wasserstein Distance lower bound

- Wasserstein distance for two distributions μ_i:

$$W(\mu_1, \mu_2) = \inf_{\gamma \in \Pi} \mathbb{E}_{(x,y)\sim\gamma} ||x - y||,$$



- it can be bounded (via Jensen inequality):

$$\inf_{\gamma \in \Pi} \mathbb{E}_{(x,y)\sim\gamma} ||x - y|| \geq \inf_{\gamma \in \Pi} |\mathbb{E}_{(x,y)\sim\gamma} |x - y|| = |m_1 - m_2|,$$

where m_i are the mean of μ_i.

From: Berthelot et al. BEGAN: Boundary Equilibrium Generative Adversarial Networks   Figure: J. Hui blog

Wasserstein Discriminator

- D needs to separate two losses as much as possible.
- We still have D as AE:

$$\mathcal{D}(x) = ||Dec(Enc(x)) - x||;$$

- We can use $\mathcal{D}$ in minibatch instead of mean:

$$\mathcal{L}_D = \mathcal{D}(x) - \mathcal{D}(G(z)).$$

- We thus optimise W between losses.
- The loss function for the generator remains the same:

$$\mathcal{L}_G = \mathcal{D}(G(z)).$$

No need for K-Lipshitz, since no Kantorovich-Rubinstein duality is used.

Equilibrium term

- we need to maintain balance between G and D:

$$\mathbb{E}(\mathcal{D}(x)) = \mathbb{E}(\mathcal{D}(G(z)))$$

- we thus can use a parameter to balance the impact:

$$\gamma = \frac{\mathbb{E}(\mathcal{D}(x))}{\mathbb{E}(\mathcal{D}(G(z)))}$$

- γ can be chosen to sharpen the image.

BEGAN formulation

We thus can write full optimisation for BEGAN:

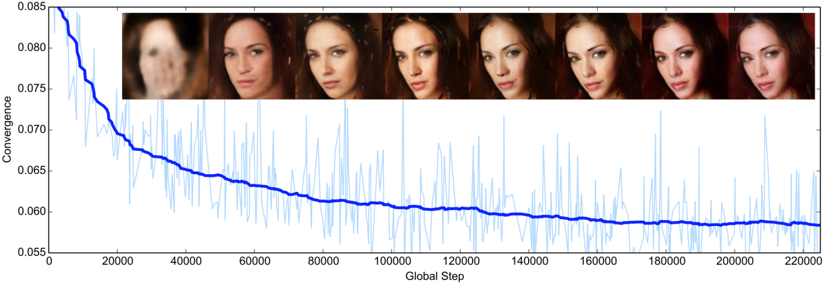
$$\begin{aligned} \mathcal{L}_D &= \mathcal{D}(x) - k_t \mathcal{D}(G(z)); \\ \mathcal{L}_G &= \mathcal{D}(G(z)); \\ k_{t+1} &= k_t + \lambda_k(\gamma \mathcal{D}(x) - \mathcal{D}(G(z))). \end{aligned}$$

When we drop γ, the image quality improves but the mode starts collapsing too.

Convergence measure

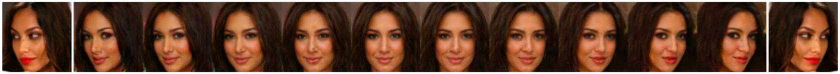
In usual GAN, the objective cannot be used to monitor convergence, here $\mathcal{D}(x)$ instead can. We can construct

$$\mathcal{M}_{global} = \mathcal{D}(x) + (\gamma \mathcal{D}(x) - \mathcal{D}(G(z)))$$



BEGAN:results

(c) Our results (128x128 with 128 filters)



(d) Mirror interpolations (our results 128x128 with 128 filters)

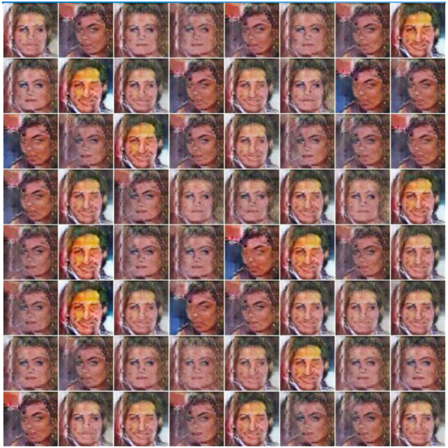
Figure 4: Interpolations of real images in latent space

Wrap Up

- We can change the architecture of discriminator.
- If we use autoencoder as discriminator we have access to the energy instead of probability.
- We can optimise Wasserstein distance not only for datasets, but also for functions.
- BEGAN output looks quite good (probably, better than WGAN). But not perfect: look in the test.

VAE-GAN

Idea: improving GANs



Main problems with GAN:

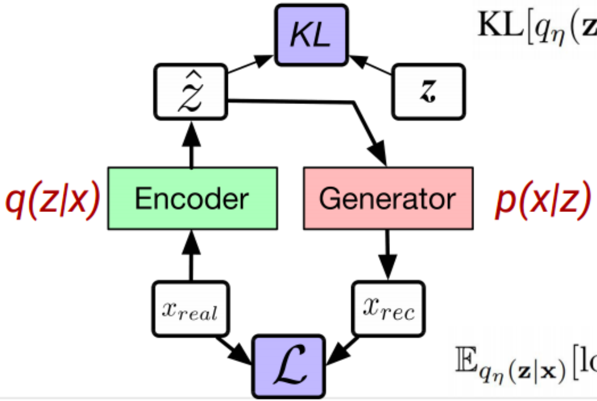
- mode collapse;
- intractable likelihood.

Idea:

- use likelihood based model.
  - have easier inference;
  - diversify sampling.

Figure and some contents in next slides: M. Rosca Tutorial at CVPR'18

Reminder: VAE structure



$$\mathrm{KL}[q_\eta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \qquad \mathcal{L}_{\text{prior}}$$

$$\mathbb{E}_{q_\eta(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] \qquad -\mathcal{L}_{\text{llike}}^{\text{pixel}}$$

VAE loss:

$$\mathcal{L}_{VAE} = \mathcal{L}_{\text{llike}}^{\text{pixel}} + \mathcal{L}_{\text{prior}},$$

Modifying VAE

- The main problem of VAE is blurry output.
- This is overcome in GAN due to the use of discriminator.
- Let's introduce discriminator loss:

$$\mathcal{L}_{GAN} = \mathbb{E}_{x\sim\mathrm{data}} \log(D(x)) - \mathbb{E}_{z\sim\mathrm{gen}} \log(1 - D(G(z)));$$

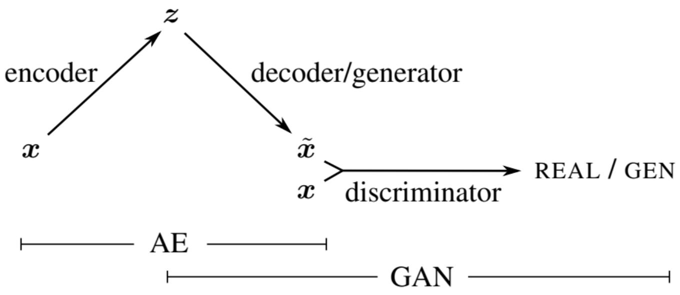
- We also introduce a reconstruction error of lth layer of discriminator D:

$$p(D(x)|z) = \mathcal{N}(D(x); \tilde{D}(x), I),$$

in order to give additional freedom.

A. Larsen, Autoencoding beyond pixels using a learned similarity metric

VAE-GAN



We thus construct a VAE with error term changed to GAN and layer error:

$$\mathcal{L}_{VAE+GAN} = \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{llike}}^{\text{D}_\text{l}} + \mathcal{L}_{GAN}$$

VAE-GAN results

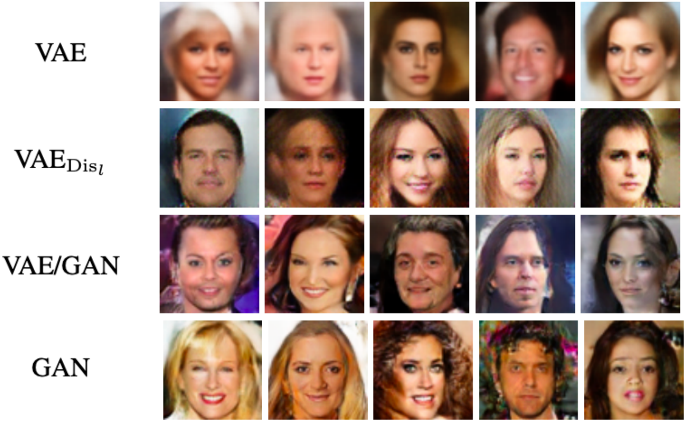
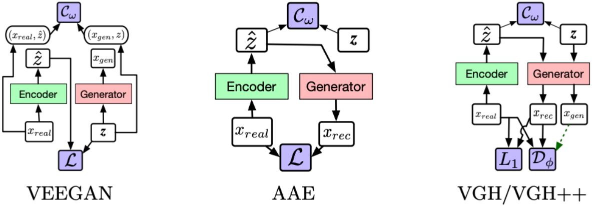


Figure 3. Samples from different generative models.

The training appears to be very tricky, however, it produces results.

KL $\longrightarrow$ Classifier

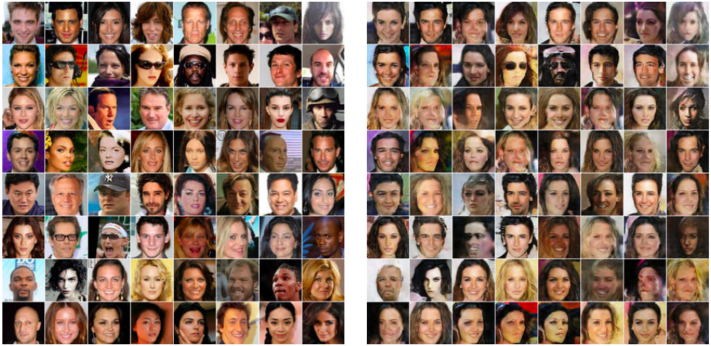


We can:

- AAE: minimize KL using adversarial training;
- VGH: Marginal matching and implicit distributions using GANs both in latent and visible space;
- VEEGAN: Directly match in joint space.

M. Rosca et al., Distribution Matching in Variational Inference

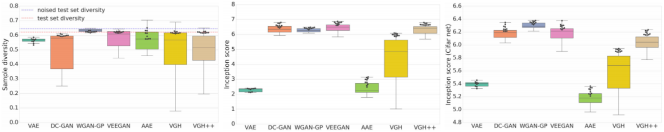
Some results



(c) VGH++

Figure 25: Training reconstructions obtained using a standard VAE, Adversarial Autoencoders and VGH++ on CelebA. Left is the data and right are reconstructions.

Aim evaluation



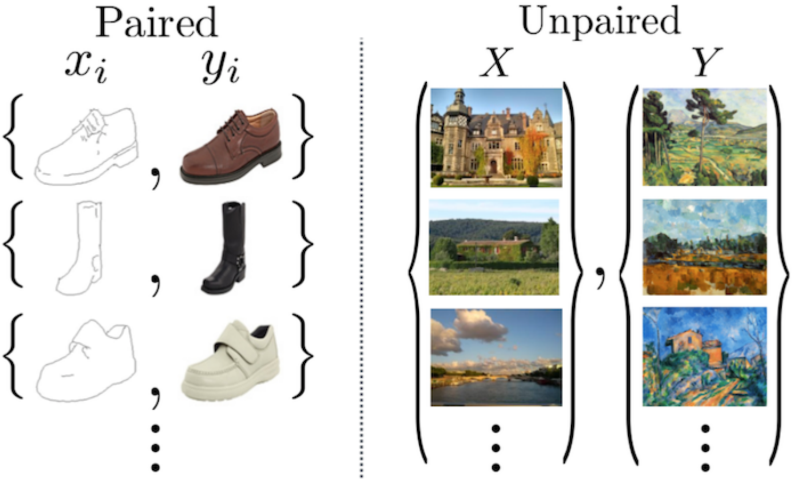
(a) Diversity score (CelebA) (b) Inception score (ImageNet) (c) Inception score (CIFAR)

Figure 8: (Left) Sample diversity on CelebA, and is viewed relative to test set: too much diversity shows failure to capture the data distribution, too little is indicative of mode collapse. We also report the diversity obtained on a noised-version of the test set, which has a higher diversity than the test set. (Middle) Inception scores on CIFAR-10. (Right) Inception scores computed using a VGG-style network on CIFAR-10. For inception scores, higher values are better. For test data, diversity score: 0.621, inception score: 11.25, inception score (using CIFAR-10 trained net): 9.18. Best results are shown with black dots, and box plots show the hyperparamter sensitivity.

- Our aim was to improve GAN's mode collapse and VAE's blurriness.
- The results are yet to be improved.
- Look also at VQ-VAE2:A. Razavi et al., Generating Diverse High-Fidelity Images with VQ-VAE-2

Image-to-image translation

Type of translations



S. Wolf blog

Paired translation

- The most straightforward way: create a regression.
- From the point of view of GANs:
  - Conditional to a picture.
  - Need to have L1 loss in encoder to capture low frequency details.
  - Need to create a discriminator that tries to recognise small details.

pix2pix GAN

- Generator with additional L1 skip connections:

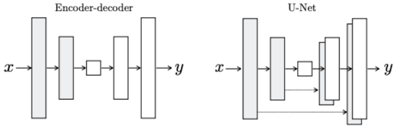


Figure 3: Two choices for the architecture of the generator. The "U-Net" [50] is an encoder-decoder with skip connections between mirrored layers in the encoder and decoder stacks.

- Discriminator: PatchGAN
  - checks each $N \times N$ patch in an image is real or fake;
  - relies on L1 for low-frequency problems.

P. Isola et al., Image-to-Image Translation with Conditional Adversarial Networks

pix2pix: results

Figure 4: Different losses induce different quality of results. Each column shows results trained under a different loss. Please see `https://phillipi.github.io/pix2pix/` for additional examples.

More results here.

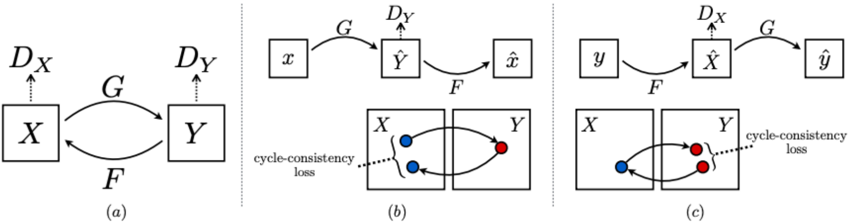
Unpaired translation

- Paring images require significant efforts.
- Can we relax the condition?
- CycleGAN:
  - take two-directional GANs;
  - introduce a cycle consistency loss.

J.-Y. Zhu et al., Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

CycleGAN Architecture

We have usual mappings $G : X \to Y$ and $F : Y \to X$:



However, we need to introduce additional cycle consistency loss to have matched points:

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[||F(G(x)) - x||] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[||G(F(y)) - y||].$$

Cycle Consistency Loss Effect

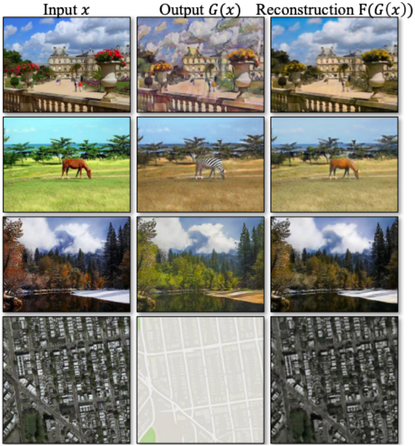


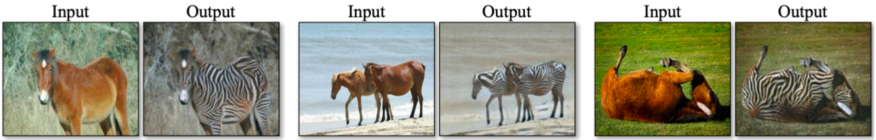
Figure 4: The input images x, output images $G(x)$ and the reconstructed images $F(G(x))$ from various experiments. From top to bottom: photo↔Cezanne, horses↔zebras, winter→summer Yosemite, aerial photos↔Google maps.

We observe the effect of loss:

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[||F(G(x)) - x|$$
$$+ \mathbb{E}_{y\sim p_{\text{data}}(y)}[||G(F(y)) - y||$$

Looking at various translations.

CycleGAN:results



horse → zebra


zebra → horse

CycleGAN:Fails

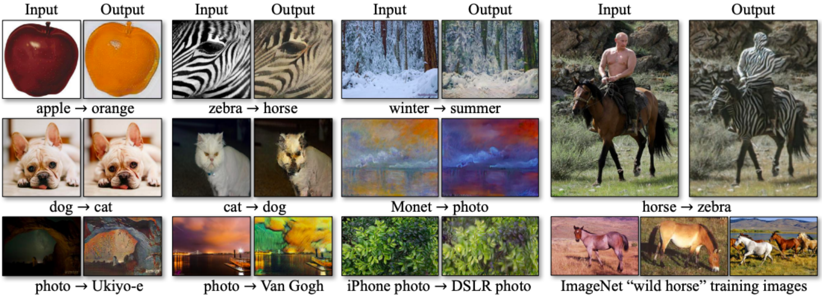


Figure 17: Typical failure cases of our method. Left: in the task of dog→cat transfiguration, CycleGAN can only make minimal changes to the input. Right: CycleGAN also fails in this horse → zebra example as our model has not seen images of horseback riding during training. Please see our website for more comprehensive results.

More results here.

Outlook

GAN-story so far

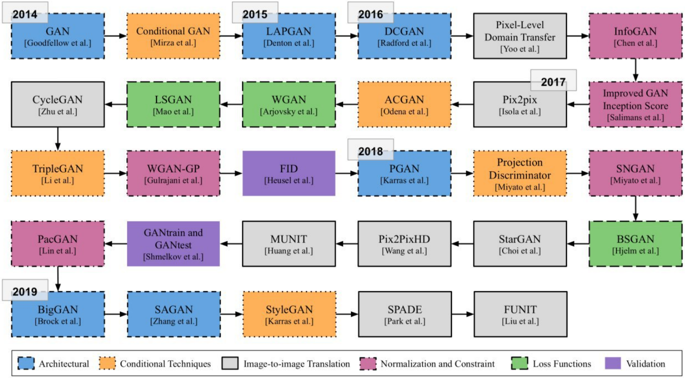


Fig. 1: Timeline of the GANs covered in this paper. Just like our text, we split it in six fronts (architectural, conditional techniques, normalization and constraint, loss functions, image-to-image translation and validation metrics), each represented by a different color and a different line/border style.

From: A. Bissoto The Six Fronts of the Generative Adversarial Networks

Are GANs created equal?

- GANs quality depends a lot on the amount of GPU one has.
- One needs to keep in mind that the ”bad” model can outperform the ”good” model given the lack of resources.

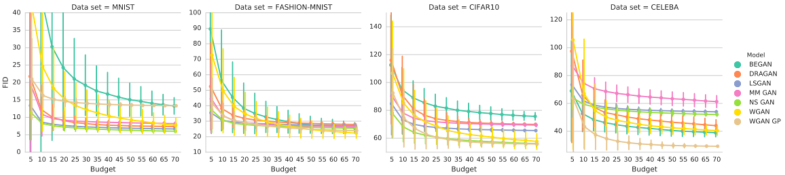


Figure 3: How does the minimum FID behave as a function of the budget? The plot shows the distribution of the minimum FID achievable for a fixed budget along with one standard deviation interval. For each budget, we estimate the mean and variance using 5000 bootstrap resamples out of 100 runs. We observe that, given a relatively low budget, all models achieve a similar minimum FID. Furthermore, for a fixed FID, “bad” models can outperform “good” models given enough computational budget. We argue that the computational budget to search over hyperparameters is an important aspect of the comparison between algorithms.

From: M. Lucic, Are GANs Created Equal? A Large-Scale Study

Conclusion

- GANs are cool to produce a single image.
- They still are not able to produce inference.
- Mode collapse is still a problem.