Learning latent space representations, and application to image generation

Aymeric Behaegel, Pierre Cornilleau, Luka Lafaye

Introduction

Introduced in 2014 by Ian Goodfellow et al., Generative Adversarial Networks (GANs) are a framework for training generative models through an adversarial process. This setup consists of two neural networks, a generator G and a discriminator D, that are trained simultaneously in a minimax game.

1 Baseline approach: vanilla GAN

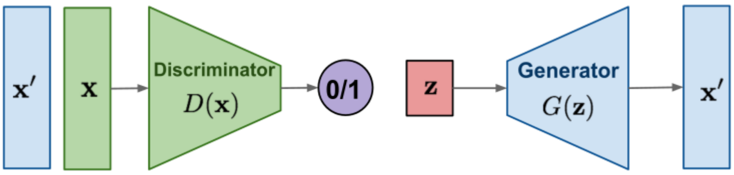
1.1 Introductory settings

Given a real distribution P (of images from database MNIST here), we aim to generate samples x_g obeying a certain distribution P_g as (visually) close as possible to P.

In the original (logistic) approach, the generator G and the discriminator D are trained simultaneously to solve the following min-max problem:

$$\min_G \max_D \mathbb{E}_{x_r \sim P}[\log(D(x))] + \mathbb{E}_{x_g \sim P_g}[\log(1 - D(x))],$$

where $x_g = G(z)$, $G : \mathcal{Z} \to \mathcal{X}$ is the generator function, and $\mathcal{Z}$ the latent space [1]



GAN, image from https://lilianweng.github.io

Figure 1: Generator and discriminator associated to a GAN

1.2 Results

We first used a simple four-layer structure for the Discriminator, three of them being linear followed by a sigmoid layer.

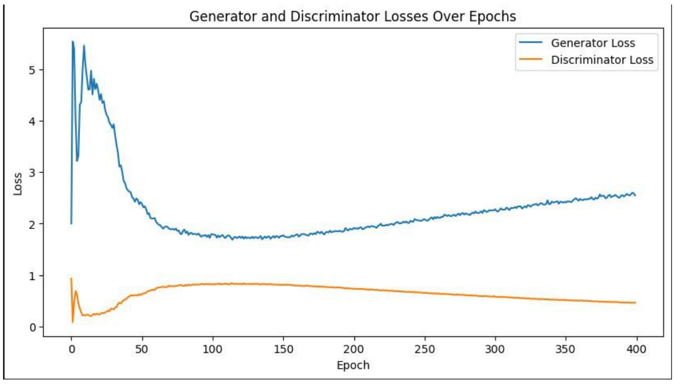


Figure 2: Generator and Discriminator Losses. An overfitting phenomena appears.

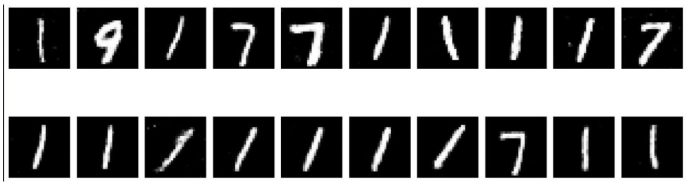


Figure 3: Some characters generated by the vanilla GAN. We notice a lack of diversity.

The metrics obtained are presented below.

[1] As a consequence, $\hat{P}_g$ is the pushforward of a measure on $\mathcal{Z}$.

FID	Precision	Recall
44.12	0.56	0.18

Figure 4: Results of the vanilla GAN method

2 An intermediary model: Wasserstein GAN

2.1 Introduction

The *Earth-Mover* distance or Wasserstein-1 distance is defined

$$W_1\left(P, P_g\right) = \inf_{\gamma \in \Pi(P, P_g)} \mathbb{E}_{(x,y)\sim\gamma}[||x - y||],$$

where $\Pi\left(\mathbb{P}_r, \mathbb{P}_g\right)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_g$. As the Fréchet Inception Distance (FID)/Wasserstein-2 distance, often used as a metric for GAN problems, it shares many properties. Defined as

$$W_2\left(P, P_g\right) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|^2]^{1/2},$$

the FID is also a distance on certain probability measures on $\mathcal{X}$ [2], such that $W_1 \leq W_2$ [2].

Interestingly, the Kantorovich-Rubinstein duality tells us that

$$W_1\left(P, P_g\right) = \sup_{[f]_{\mathrm{Lip}} \leq 1} \left(\mathbb{E}_{x\sim Pr}[f(x)] - \mathbb{E}_{x\sim Pg}[f(x)]\right)$$

where the supremum is over all the 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$. Hence the minimization problem

$$\inf_{P_g} W_1(P, P_g)$$

becomes, similarly to the vanilla GAN, a minimax problem:

$$\inf_{P_g} W_1(P, P_g) = \inf_{P_g} \sup_{[f]_{\mathrm{Lip}} \leq 1} \left(\mathbb{E}_{x\sim Pr}[f(x)] - \mathbb{E}_{x\sim Pg}[f(x)]\right).$$

Apart from this adversarial structure (as $x = G(z)$, the infimum can actually be indexed by the generator G, whereas f is the Discriminator), this last problem also have some *metrisability* property : solving it ensures that the generator's distribution becomes closer to the target distribution in a measurable way.

2.2 Results

Wasserstein GAN Without any modification of the architecture nor other specifications used previously, we only get poor metrics (especially Recall); see table below.

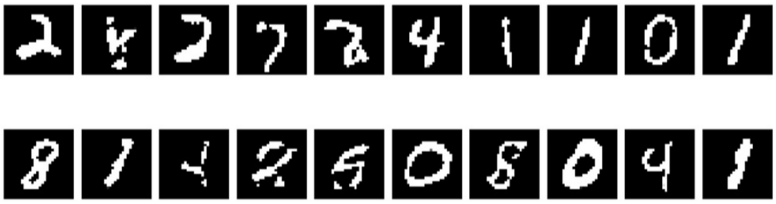


Figure 5: Some characters generated by the Wasserstein GAN

Hinge loss This problem amounts to solve

$$\inf_{P_g} \sup_{[f]_{\mathrm{Lip}} \leq 1} \left(\mathbb{E}_{x\sim P}[\max(0, 1 + f(x))] + \mathbb{E}_{x\sim P_g}[\max(0, 1 - f(x))]\right),$$

meaning that the functions f are cut off inside the expectations. As already presented last week, hinge loss allows us to get better results.

[2] We also have an inequality of the form $W_2 \leq CW_1$ as long as $\mathcal{X}$ is a bounded set of some $\mathbb{R}^{D_x}$.

Method	FID	Precision	Recall
Wasserstein	89.84	0.19	0.00
Hinge Loss	22.08	0.40	0.20

Figure 6: Results of the Wasserstein GAN and hinge–loss GAN methods (results obtained after 150 epochs)

3 Sliced Adversarial Networks

3.1 Introduction

Following [3], we here present a novel approach to enhancing Generative Adversarial Networks (GANs) by introducing the Slicing Adversarial Network (SAN) framework. This framework aims to make GANs more metrizable, meaning that it helps ensure that the generator's distribution becomes closer to the target distribution in a measurable way (as for Wasserstein GAN), without requiring the ideal discriminator.

3.2 Theoretical points

3.2.1 Definitions

Let us denote, for any function f defined on $\mathcal{X}$,

$$d_f(P, P_g) = \mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim P_g}[f(x)];$$

it is actually the Wasserstein GAN loss.

In the following, we introduce some parameter θ to follow the Genrator: μ_θ will correspond to P_g, and μ_0 to the distribution P.

Direction optimality Let $\omega^* \in \mathbb{S}^{D-1}$ and $h : \mathcal{X} \to \mathbb{R}^D$ any functipon. We say that ω^* is an *optimal direction* (related to h and θ) if $\omega^* \in \underset{\omega\in\mathbb{S}^{D-1}}{\arg\max} \quad d_{\langle\omega,h\rangle}(\mu_0, \mu_\theta) =: \hat{d}_h(\mu_0, \mu_\theta)$.

Separability Let us introduce some extra vocabulary.

Definition (Separable). Given μ, ν probability measures on $\mathcal{X}$, let ω be on $\mathbb{S}^{D-1}$, and let $F_\mu^{h,\omega}(\cdot)$ be the cumulative distribution function of $\mathcal{S}^h I_\mu(\cdot, \omega)$. If $\omega^* \in \hat{d}_h(\mu, \nu)$ satisfies $F_\mu^{h,\omega^*}(\xi) \leq F_\nu^{h,\omega^*}(\xi)$ for any $\xi \in \mathbb{R}$, $h \in L^\infty\left(\mathcal{X}, \mathbb{R}^D\right)$ we say that h is *separable* for those probability measures.

We denote the class of all these separable functions for them as $\mathcal{F}_{S(\mu,\nu)}(X)$ or $\mathcal{F}_S$ for notation simplicity.

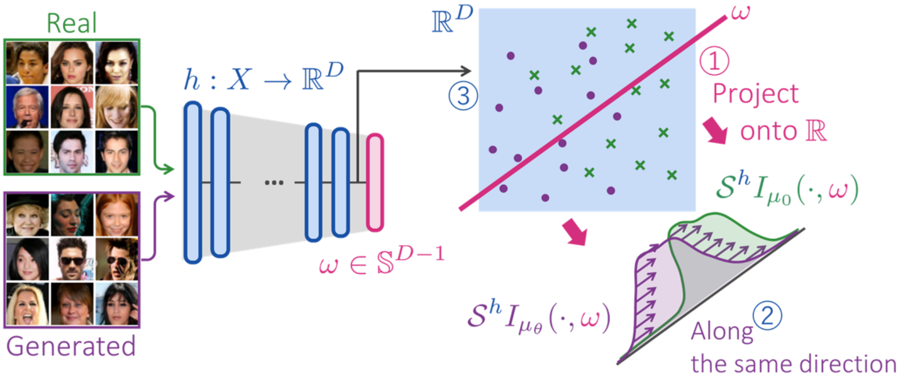

Real
Generated
$h : X \to \mathbb{R}^D$
$\omega \in \mathbb{S}^{D-1}$
$\mathbb{R}^D$
③
ω
① Project
onto $\mathbb{R}$
$\mathcal{S}^h I_{\mu_0}(\cdot, \omega)$
$\mathcal{S}^h I_{\mu_\theta}(\cdot, \omega)$
Along ②
the same direction

Figure 7: Illustration of direction optimality, separability and injectivity properties. Here μ_θ stands for the generated measure and μ_0 for the target measure.

In the sequel, we choose a Discriminator function of the form $f(x) =< h(x), \omega >$, for some $\omega \in \mathbb{S}^{D-1}$ and a function $h : \mathcal{X} \to \mathbb{R}^D$.

3.2.2 Metrisability proprerty

Notations. In the following, we put

$$\mathcal{J}_W(\theta, f) := -\mathbb{E}_{x\sim\mu_\theta}[f(x)],$$

for any function f defined on $\mathcal{X}$. Up to constant (with respect to θ) $\mathbb{E}_{x\sim\mu_0}[f(x)]$, it is the Wasserstein GAN loss.

For any function $h : \mathcal{X} \to \mathbb{R}^D$ and any direction $\omega \in \mathbb{S}^{D-1}$ and any density function I, we introduce

$$\mathcal{S}^h I(\xi, \omega) := \int_{\mathcal{X}} I(x)\delta(\xi - \langle\omega, h(x)\rangle)dx.$$

Now, for P, Q measures that are absolutely continous with respect to the Lebesgue measure, we define

$$max - ASW_h(P, Q) := \max_{\omega \in \mathbb{S}^{D-1}} W_1(S_h I_P(\cdot, \omega), S_h I_Q(\cdot, \omega)),$$

I_P and I_Q being here the densities of P and Q.

Following [3], we also denote $FM_h^*(P, Q) = \|d_h(P, Q)\|_{\mathbb{R}^D}$ for any function $h : \mathcal{X} \to \mathbb{R}^D$.

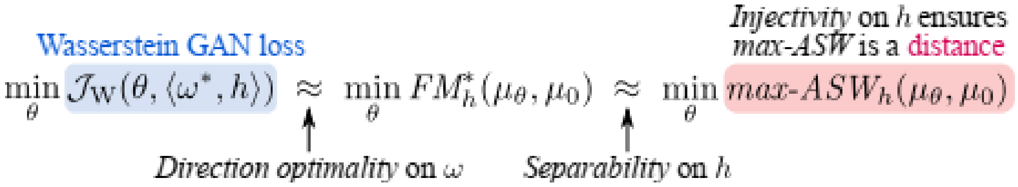

Wasserstein GAN loss
Injectivity on h ensures
max-ASW is a distance
$\min_\theta \mathcal{J}_W(\theta, \langle\omega^*, h\rangle) \approx \min_\theta FM_h^*(\mu_\theta, \mu_0) \approx \min_\theta max\text{-}ASW_h(\mu_\theta, \mu_0)$
Direction optimality on ω
Separability on h

Figure 8: Main Theorem of [3]: Under the previous properties, the Wassertein GAN loss is metrizable.

Adavantage. Contrary to the distance $max - ASW_h(\mu_\theta, \mu_0)$, backpropagation of the Wasserstein GAN loss $\mathcal{J}_W(\theta, < \omega^*, h >)$ can be performed in a satisfying way. And these optimization problems are equivalent under the previous conditions.

3.3 Definition of the loss

For λ a weight parameter and $\tilde{\mu}_0^{r\mathcal{J}\circ f}$, $\tilde{\mu}_\theta^{r\mathcal{J}\circ f}$ some truncated versions of the measures μ_0, μ_θ (see [3, Section6] for more details), we define the loss function as

$$\mathcal{V}^{\text{SAN}}(\omega, h; \mu_\theta) := \underbrace{\mathcal{V}\left(\langle\omega^-, h\rangle ; \mu_\theta\right)}_{\mathcal{L}^h(h;\omega,\mu_\theta)} + \lambda \cdot \underbrace{d_{<\omega, h^->}\left(\tilde{\mu}_0^{r\mathcal{J}\circ f}, \tilde{\mu}_\theta^{r\mathcal{J}\circ f}\right)}_{\mathcal{L}^\omega(\omega;h,\mu\theta)},$$

where $^-$ stands for the stop–gradient operator, meaning that we won't backpropagate the affected variables.

Here $\mathcal{V}$ is a "vanilla–loss", which will be chosen to be the hinge loss:

$$\mathcal{V}(f; \mu_\theta) = \mathbb{E}_{x\sim\mu_0}[\max(0, 1 + f(x))] + \mathbb{E}_{x\sim\mu_\theta}[\max(0, 1 - f(x))].$$

In practice as well, we don't use the truncations of μ_0 and μ_θ in our code.

3.4 Architecture of the Discriminator

As explicitly mentioned inside this project's guidelines, the architecture of the Generator will be fixed throughout this note. However, we would like to change the Generator architecture to get a better computation of the inner maximum.
Inspired by the architecture of the Generator proposed in [3], we decided to use a mirrored architecture for the Discriminator (see fig. 9).

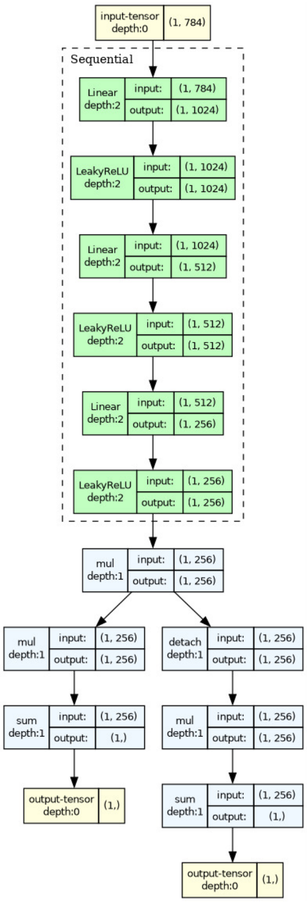

input-tensor depth:0 (1, 784)
Sequential
Linear depth:2 input: (1, 784) output: (1, 1024)
LeakyReLU depth:2 input: (1, 1024) output: (1, 1024)
Linear depth:2 input: (1, 1024) output: (1, 512)
LeakyReLU depth:2 input: (1, 512) output: (1, 512)
Linear depth:2 input: (1, 512) output: (1, 256)
LeakyReLU depth:2 input: (1, 256) output: (1, 256)
mul depth:1 input: (1, 256) output: (1, 256)
mul depth:1 input: (1, 256) output: (1, 256)
detach depth:1 input: (1, 256) output: (1, 256)
sum depth:1 input: (1, 256) output: (1,)
mul depth:1 input: (1, 256) output: (1, 256)
output-tensor depth:0 (1,)
sum depth:1 input: (1, 256) output: (1,)
output-tensor depth:0 (1,)

Figure 9: Chosen architecture for the Discriminator

3.5 Results

After some tuning of the parameter λ and of the batch size, we obtain better results with $\lambda = 1$ (as in [3]) and a batch size of 64. Evolution of the metrics Precision, Recall and FID over epochs are presented in 10.

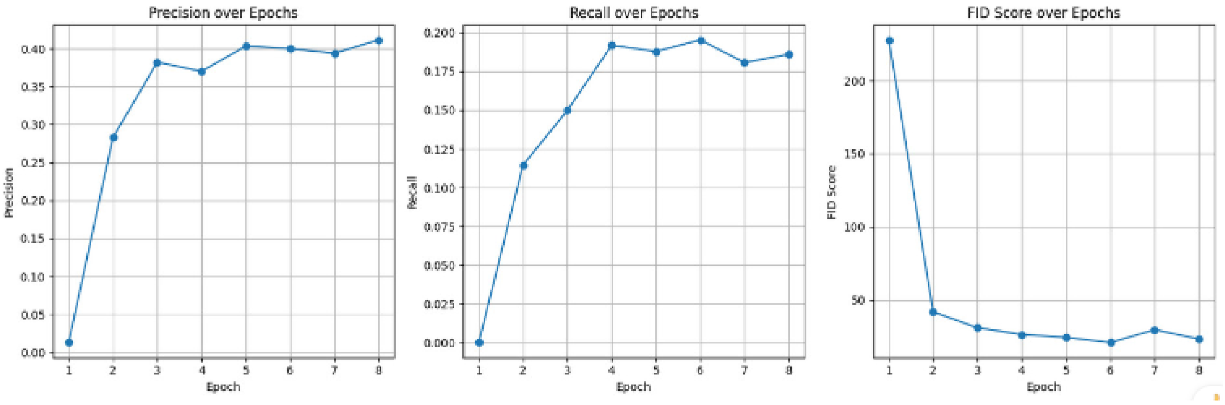


Figure 10: With a batch size of 64, we reach a FID of 19.04

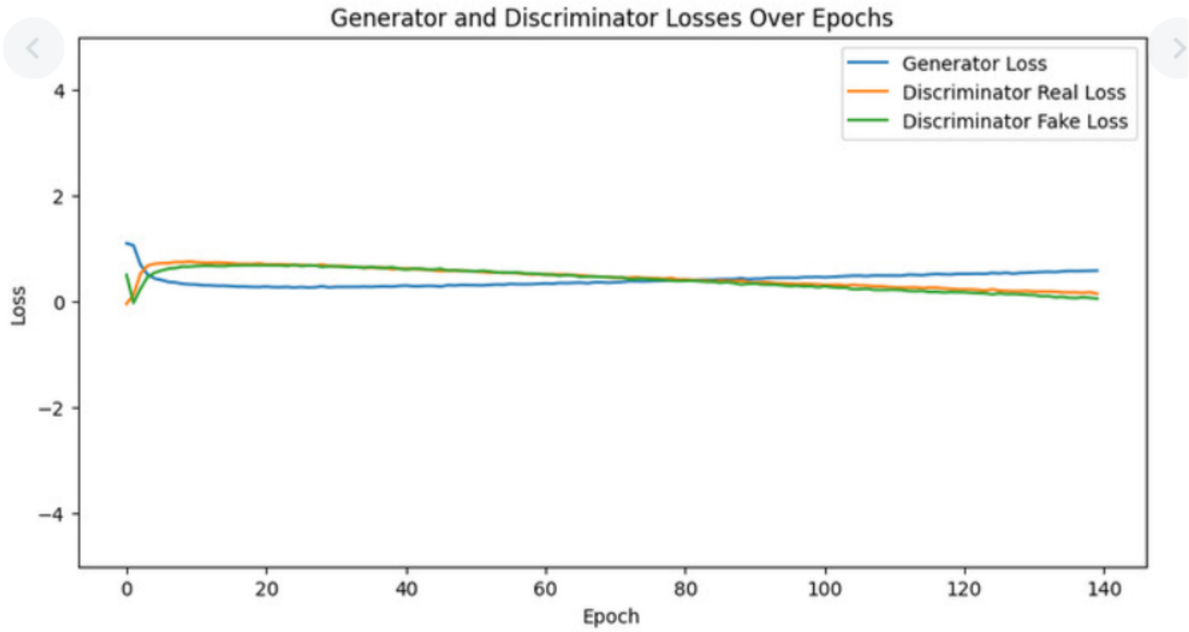


Figure 11: Generator and Discriminator Losses with a batch size of 64.

Conclusion

This research addresses fundamental limitations of traditional GANs by ensuring that the discriminator more effectively guides the generator towards the target distribution. The SAN framework provides a promising, theoretically motivated, and empirically supported method for improving GAN performance across a range of tasks.

References

[1] Arjovsky M., Chintalah S., Bottou L.. *Wasserstein GAN*, https://arxiv.org/abs/1701.07875, 2017.

[2] Santambrogio F.. *The Wassertein distances*, https://math.univ-lyon1.fr/~santambrogio/Wp.pdf, 2011.

[3] Takida Y., Imaizumi M., Shibuya T., Lai C.-H., Uesaka T., Murata N. and Mitsufuji Y.. *SAN: Inducing Metrizability of GAN with Discriminative Normalized Linear Layer*, International Conference on Learning Representations 2024, https://iclr.cc/virtual/2024/poster/18212.