NPFL114, Lecture 10



Deep Generative Models

Milan Straka

🛗 May 6, 2019





EUROPEAN UNION European Structural and Investment Fund Operational Programme Research, Development and Education Charles University in Prague Faculty of Mathematics and Physics Institute of Formal and Applied Linguistics



unless otherwise stated

Generative Models

Ú F_AL

Generative models are given a set \mathcal{X} of realizations of a random variable \mathbf{x} and their goal is to estimate $P(\mathbf{x})$.

Usually the goal is to be able to sample from $P(\mathbf{x})$, but sometimes an explicit calculation of $P(\mathbf{x})$ is also possible.

Deep Generative Models





Figure 1 of paper "Auto-Encoding Variational Bayes", https://arxiv.org/abs/1312.6114.

One possible approach to estimate $P(m{x})$ is to assume that the random variable $m{x}$ depends on a latent variable **z**:

$$P(\boldsymbol{x}) = P(\boldsymbol{z})P(\boldsymbol{x}|\boldsymbol{z}).$$

We use neural networks to estimate the conditional probability with $P_{\theta}(\boldsymbol{x}|\boldsymbol{z})$.

AutoEncoders





- unsupervised feature extraction
- input compression for z < x
- when $oldsymbol{x}+oldsymbol{arepsilon}$ is used as input, autoencoders can perform denoising

NPFL114, Lecture 10

GAN CGAN

DCGAN

*GAN

WGAN

We assume $P(\mathbf{z})$ is fixed and independent on \mathbf{x} .

We approximate $P(\boldsymbol{x}|\boldsymbol{z})$ using $P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. However, in order to train an autoencoder, we need to know the posterior $P_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$, which is usually intractable.

We therefore approximate $P_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ by a trainable $Q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})$.

VAE



Let us define variational lower bound or evidence lower bound (ELBO), denoted $\mathcal{L}(\theta, \varphi; \mathbf{x})$, as

$$\mathcal{L}(oldsymbol{ heta},oldsymbol{arphi};\mathbf{x}) = \log P_{oldsymbol{ heta}}(oldsymbol{x}) - D_{ ext{KL}}(Q_{oldsymbol{arphi}}(oldsymbol{z}|oldsymbol{x}))|P_{oldsymbol{ heta}}(oldsymbol{z}|oldsymbol{x})).$$

Because KL-divergence is non-negative, $\mathcal{L}(\boldsymbol{ heta}, \boldsymbol{arphi}; \mathbf{x}) \leq \log P_{\boldsymbol{ heta}}(\boldsymbol{x}).$

By using simple properties of conditional and joint probability, we get that

$$egin{aligned} \mathcal{L}(oldsymbol{ heta},oldsymbol{arphi};\mathbf{x}) &= \mathbb{E}_{Q_arphi(oldsymbol{z}|oldsymbol{x})}[\log P_{oldsymbol{ heta}}(oldsymbol{x}) + \log P_{oldsymbol{ heta}}(oldsymbol{z}|oldsymbol{x})] \ &= \mathbb{E}_{Q_arphi(oldsymbol{z}|oldsymbol{x})}[\log P_{oldsymbol{ heta}}(oldsymbol{x}|oldsymbol{z}) - \log Q_arphi(oldsymbol{z}|oldsymbol{x})] \ &= \mathbb{E}_{Q_arphi(oldsymbol{z}|oldsymbol{x})}[\log P_{oldsymbol{ heta}}(oldsymbol{x}|oldsymbol{z}) + \log P(oldsymbol{z}) - \log Q_arphi(oldsymbol{z}|oldsymbol{x})] \ &= \mathbb{E}_{Q_arphi(oldsymbol{z}|oldsymbol{x})}[\log P_{oldsymbol{ heta}}(oldsymbol{x}|oldsymbol{z})] - D_{\mathrm{KL}}(Q_arphi(oldsymbol{z}|oldsymbol{x}))|P(oldsymbol{z})). \end{aligned}$$

VAE

*GAN

WGAN



$\mathcal{L}(oldsymbol{ heta},oldsymbol{arphi};\mathbf{x}) = \mathbb{E}_{Q_{oldsymbol{arphi}}(oldsymbol{z}|oldsymbol{x})}[\log P_{oldsymbol{ heta}}(oldsymbol{x}|oldsymbol{z})] - D_{ ext{KL}}(Q_{oldsymbol{arphi}}(oldsymbol{z}|oldsymbol{x}))|P(oldsymbol{z}))$

We train a VAE by maximizing $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}; \mathbf{x})$, taking a single point estimate of the expectation and using a prior $P(\boldsymbol{z}) = \mathcal{N}(0, 1)$.

Note that the loss has 2 intuitive components:

- reconstruction loss: Starting with $m{x}$, and passing though Q and then again through P should arrive back at $m{x}$.
- latent loss: The distribution of $Q_{\varphi}(\boldsymbol{z}|\boldsymbol{x})$ should be as close to the prior $P(\boldsymbol{z}) = \mathcal{N}(0, 1)$, which is independent on \boldsymbol{x} .

NPFL114, Lecture 10 Autoencoders VAE ReparametrizationTrick GAN



8/33

In order to derivate through $m{z} \sim Q_{m{arphi}}(m{z}|m{x})$, note that if

 $oldsymbol{z}\sim\mathcal{N}(oldsymbol{\mu},oldsymbol{\sigma}^2),$

we can write \boldsymbol{z} as

$$oldsymbol{z} \sim oldsymbol{\mu} + oldsymbol{\sigma} \cdot \mathcal{N}(0,1).$$

Such formulation then allows differentiating z with respect to μ and σ and is called a *reparametrization trick* (Kingma and Welling, 2013).







(a) Learned Frey Face manifold

(b) Learned MNIST manifold

Figure 4 of paper "Auto-Encoding Variational Bayes", https://arxiv.org/abs/1312.6114.

DCGAN

*GAN

WGAN

NPFL114, Lecture 10

ReparametrizationTrick

ck GAN CGAN



2 + 20431950

(a) 2-D latent space

(b) 5-D latent space

(c) 10-D latent space

(d) 20-D latent space

*GAN

Figure 5 of paper "Auto-Encoding Variational Bayes", https://arxiv.org/abs/1312.6114.

NPFL114, Lecture 10

Autoencoders

VAE

ReparametrizationTrick

GAN CGAN I

I DCGAN WGAN

VAE – Too High Latent Loss





NPFL114, Lecture 10

VAE – **Too High Reconstruction Loss**



中国市场的。 中国市场的 中国市场的 中国市场的。 中国市场的 中国市场的 中国市场的 中国市场的 中国市场的 中国市场的 中国市场的	清你得你好我的你怎么你你你你你!!!!	
中国 中国 中国 中国 中国 中国 中国 中国 中国 中国	现现多少的现在分词 医白喉的 医白喉的 医白喉的	
中国 中国 中国 中国 中国 中国 中国 中国 中国 中国	やうでないないでいいいいいのうのしょう	
中国 中国 中国 中国 中国 中国 中国 中国 中国 中国	法国的财产的 化合体管理 化合体管理 化合体管理	
中国的国际,在于国际的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际的。 中国的国际会会会。 中国的国际会会会。 中国的国际会会。 中国的国际会会。 中国的国际会会。 中国的国际会会。 中国的国际会会。 中国的国际会会会。	· · · · · · · · · · · · · · · · · · ·	-
中国 中国 中国 中国 中国 中国 中国 中国 中国 中国	经分额的股份额的进行可以把你知道的帮助。	
中午市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市市	当然没能已经完全的现在分词的变量。	
○公律部公司不可以帮助。 ○公律部公司法律部公司法律部公司法律部公司法律部会会法律部委任任任任任任任任任任任任任任任任任任任任任任任任任任任任任任任任任任任任	金 医 化 医 化 日 日 日 日 日 日 日 日 日 日 日 日 日 日 日 日	
○共中国公司公司公司公司公司公司公司公司公司公司公司公司公司公司公司公司公司公司公司	深了现在这些的,你会说我的你。"	
· · · · · · · · · · · · · ·	至主致也承受留害以完成医医供保卫学师医的	100
会议部的政策的现在分词的法律的政策的法律的法律的法律的法律的法律的法律的法律的法律的法律的法律的法律的法律的法律的	医白喉的足球的 医马耳氏 医马耳氏 医马马氏 医马马氏	
会议和部队的知道要不能受到有限的部分分别的 必要和你的现在分词是是有限的。 我们们是我们们们就是我们们们就会 我们们们就是我们们就是我们们就是我们们就是我们们就是我们们们就是我们们们就是我们们就是我们们就是我们们就是我们们就是我们们就是我们们就会好的你们就会们们就会好你们就会你们们们不会。" 我们们们们们们们们们们们们们们们的,你们们们们们们们不是	ちころをもうもうものないのもろうがらの	-
会议和部队的知道要许能感到有限的情况的。 我们们要就是有限的情况。 我们们有这些问题,我们们是我们们的我们们不会不可能不会不可能不会不可能。 我们们是我们们就是我们们会会们们们就是我们们不会不可能。 我们们们们们就是我们们的我们们就是我们们不会不可能会们的。 我们们们们们们们们的?"	ふぶきつうもうがいりゅうちょうもうぶい	
金头和街路的到外都在的东西有限的时候的外方。 不可以会会的包括中国的时候就是有的人们的一个。 我们们是我们们们的你们的你们的你们们不会不会不会没有不会。 我们们们我们们们会会会们的你们的你?" 我们们们们们的我们们的你们们的你?"	日子 医子宫 医子宫 医子宫 医子宫 医白白白	
金头和街路的到来到你的话题有限的情况的过去式和过去分词 计数据数据 医脊髓脊髓管 医外外的 化化化合金 化化合金 化化化合金 化化化合金 化化化合金 化化化合金 化化化合金 化化化合金 化化化合金 化化化合金 化化化合金 化化合金 化的 化化合金 化化合金	周田 医子宫 武臣 经公 医子宫 经保留 经保险	
ල්ස්සාලින් කිරීම් සිම්සාම්ස්ත්රී කිරීම් සිම්සාම්ස්ත්රීම් සිම්සාම්සා සිම්සාම්ස් සිම්සාම්ස්ත්රීම් කිරීම් රම්දී ස්රේම්සා සිම්සාම්ස් සිම්සාම්ස් සිම්සාම්ස් සිම්සාම් සිම්සාම්ස් සිම්සාම්ස් සිම්සාම්ස් සිම්සාම්ස් සිම්සාම්ස් සිම්සාම්ස් සිම්සාම්ස්	我你你想你说我我?你你你说你你吗? "	
௹௸௸ௐௐௐ௺௺௺௺௺௶௺௺௺௺௺௺௺௺௺ ஂ௸ௐ௸௺௶௵௶௵ௐௐ௺௺௵௵௵௵௵ ௐௐௐௐ௺௶௺௺௺௺௺௺௺௺௺௺௺௺௺௺௺௶	<u>ROSESEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE</u>	
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	医马马勒法国先生百名国家子名国马马男人名	į
金米县 医牙的 医外的 医胆管的 医牙的 医肉肉 医肉肉	5 <b>3 3 3 8 9 9 9 11 3 6 5 3 3 4 4 9 2 5 5</b> 8	
	\$\$\$\$\$\$\$ <b>\$</b> \$ <b>\$</b> \$ <b>\$</b> \$ <b>\$</b> \$ <b>\$\$\$\$\$\$</b>	

NPFL114, Lecture 10



We have a *generator*, which given  $oldsymbol{z} \sim P(\mathbf{z})$  generates data  $oldsymbol{x}$ .

We denote the generator as  $G(\boldsymbol{z}; \boldsymbol{\theta}_g)$ .

Then we have a *discriminator*, which given data x generates a probability whether x comes from real data or is generated by a generator.

We denote the discrimininator as  $D(\boldsymbol{x}; \boldsymbol{\theta}_d)$ .

The discriminator and generator play the following game:

$$\min_{G} \max_{D} \mathbb{E}_{oldsymbol{x} \sim P_{ ext{data}}} [\log D(oldsymbol{x})] + \mathbb{E}_{oldsymbol{z} \sim P(oldsymbol{z})} [\log(1 - D(G(oldsymbol{z})))].$$

NPFL114, Lecture 10





Figure 1 of paper "Generative Adversarial Nets", https://arxiv.org/abs/1406.2661.

*GAN

NPFL114, Lecture 10

Autoencoders VAE

ReparametrizationTrick

GAN CGAN

DCGAN WGAN



Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used k = 1, the least expensive option, in our experiments.

for number of training iterations do

for k steps do

- Sample minibatch of m noise samples  $\{z^{(1)}, \ldots, z^{(m)}\}$  from noise prior  $p_g(z)$ .
- Sample minibatch of m examples  $\{x^{(1)}, \ldots, x^{(m)}\}$  from data generating distribution  $p_{\text{data}}(x)$ .
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m \left[ \log D\left( \boldsymbol{x}^{(i)} \right) + \log \left( 1 - D\left( G\left( \boldsymbol{z}^{(i)} \right) \right) \right) \right].$$

end for

- Sample minibatch of m noise samples  $\{z^{(1)}, \ldots, z^{(m)}\}$  from noise prior  $p_g(z)$ .
- Update the generator by descending its stochastic gradient:

$$abla_{\theta_g} rac{1}{m} \sum_{i=1}^m \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

#### end for

Autoencoders

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

Algorithm 1 of paper "Generative Adversarial Nets", https://arxiv.org/abs/1406.2661.

WGAN





Figure 2 of paper "Generative Adversarial Nets", https://arxiv.org/abs/1406.2661.

DCGAN

NPFL114, Lecture 10

Autoencoders VAE

ReparametrizationTrick

GAN

CGAN

16/33

*GAN

WGAN

**Conditional GAN** 





Figure 1 of paper "Conditional Generative Adversarial Nets", https://arxiv.org/abs/1411.1784.

NPFL114, Lecture 10

Autoencoders VAE

ReparametrizationTrick

GAN CGAN

DCGAN

WGAN

*GAN





Figure 1 of paper "An Online Learning Approach to Generative Adversarial Networks", https://arxiv.org/abs/1706.03269.

NPFL114, Lecture 10

Autoencoders VAE

ReparametrizationTrick

GAN CGAN

DCGAN

WGAN

*GAN

1



3



Figure 1 of paper "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", https://arxiv.org/abs/1511.06434.

NPFL114, Lecture 10

Autoencoders VAE

ReparametrizationTrick

GAN CGAN

DCGAN WGAN

*GAN





Figure 2 of paper "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", https://arxiv.org/abs/1511.06434.

NPFL114, Lecture 10

Autoencoders

VAE

ReparametrizationTrick

GAN CGAN

DCGAN

WGAN





Figure 3 of paper "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", https://arxiv.org/abs/1511.06434.

NPFL114, Lecture 10

Autoencoders

VAE

ReparametrizationTrick

GAN (

CGAN

DCGAN

WGAN





Figure 4 of paper "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", https://arxiv.org/abs/1511.06434.







NPFL114, Lecture 10

Autoencoders VAE ReparametrizationTrick

GAN

CGAN

DCGAN

WGAN

*GAN





NPFL114, Lecture 10

Autoencoders VAE

 ${\sf Reparametrization} {\sf Trick}$ 

GAN C

CGAN DC

DCGAN WGAN

*GAN



Figure 8 of paper "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", https://arxiv.org/abs/1511.06434.

NPFL114, Lecture 10

Autoencoders

VAE

 ${\sf Reparametrization} {\sf Trick}$ 

GAN CGAN

DCGAN WGAN



# **GANs are Problematic to Train**





Figure 2 of paper "Unrolled Generative Adversarial Networks", https://arxiv.org/abs/1611.02163.

- Feature matching
- Minibatch discrimination
- Historical averaging
- Label smoothing

NPFL114, Lecture 10

Autoencoders VAE

ReparametrizationTrick

GAN CGAN

DCGAN WGAN

# **Minibatch Discrimination**





27/33

NPFL114, Lecture 10

Autoencoders VAE ReparametrizationTrick

GAN

CGAN

DCGAN WGAN



Instead of minimizing JS divergence

$$JS(p,q) = KL(p||q) + KL(q||p),$$

Wasserstein GAN minimizes Earth-Mover distance

$$W(p,q) = \inf_{\gamma \in \Pi(p,q)} \mathbb{E}_{(x,y) \sim \gamma}ig[||x-y||ig].$$

The joint distribution  $\gamma \in \Pi(p,q)$  indicates how much "mass" must be transported from x to y, and EM is the "cost" of the optimal transport plan.

 NPFL114, Lecture 10
 Autoencoders
 VAE
 ReparametrizationTrick
 GAN
 CGAN

WGAN



Using a dual version of the Earth-Mover definition, we arrive at

$$W(p,q) = \sup_{f, ||f||_L \leq 1} \mathbb{E}_{x \sim p}ig[f(x)ig] - \mathbb{E}_{y \sim q}ig[f(x)ig].$$



NPFL114, Lecture 10

Autoencoders VAE ReparametrizationTrick GAN CGAN DCGAN WGAN *GAN



Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values  $\alpha = 0.00005$ , c = 0.01, m = 64,  $n_{\text{critic}} = 5$ .

**Require:** :  $\alpha$ , the learning rate. c, the clipping parameter. m, the batch size.  $n_{\text{critic}}$ , the number of iterations of the critic per generator iteration.

**Require:** :  $w_0$ , initial critic parameters.  $\theta_0$ , initial generator's parameters.

1: while  $\theta$  has not converged **do** 

2: **for** 
$$t = 0, ..., n_{\text{critic}}$$
 **do**  
3: Sample  $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_{r}$  a batch from the real data.  
4: Sample  $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$  a batch of prior samples.  
5:  $g_{w} \leftarrow \nabla_{w} \left[\frac{1}{m} \sum_{i=1}^{m} f_{w}(x^{(i)}) - \frac{1}{m} \sum_{i=1}^{m} f_{w}(g_{\theta}(z^{(i)}))\right]$   
6:  $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_{w})$   
7:  $w \leftarrow \text{clip}(w, -c, c)$   
8: **end for**  
9: Sample  $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$  a batch of prior samples.  
10:  $g_{\theta} \leftarrow -\nabla_{\theta} \frac{1}{m} \sum_{i=1}^{m} f_{w}(g_{\theta}(z^{(i)}))$   
11:  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_{\theta})$   
12: **end while**

Algorithm 1 of paper "Wasserstein GAN", https://arxiv.org/abs/1701.07875.

DCGAN

ReparametrizationTrick

NPFL114, Lecture 10



Figure 5: Algorithms trained with a DCGAN generator. Left: WGAN algorithm. Right: standard GAN formulation. Both algorithms produce high quality samples.



Figure 6: Algorithms trained with a generator without batch normalization and constant number of filters at every layer (as opposed to duplicating them every time as in [18]). Aside from taking out batch normalization, the number of parameters is therefore reduced by a bit more than an order of magnitude. Left: WGAN algorithm. Right: standard GAN formulation. As we can see the standard GAN failed to learn while the WGAN still was able to produce samples.

GAN

CGAN

VAE

ReparametrizationTrick

Autoencoders

Figures 5 and 6 of paper "Wasserstein GAN", https://arxiv.org/abs/1701.07875.

DCGAN

WGAN





Figure 7: Algorithms trained with an MLP generator with 4 layers and 512 units with ReLU nonlinearities. The number of parameters is similar to that of a DCGAN, but it lacks a strong inductive bias for image generation. Left: WGAN algorithm. Right: standard GAN formulation. The WGAN method still was able to produce samples, lower quality than the DCGAN, and of higher quality than the MLP of the standard GAN. Note the significant degree of mode collapse in the GAN MLP.

Figure 7 of paper "Wasserstein GAN", https://arxiv.org/abs/1701.07875.

*GAN

NPFL114, Lecture 10

GAN CGAN

# **Development of GANs**

Generative Adversarial Networks are still in active development:

- Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen: Progressive Growing of GANs for Improved Quality, Stability, and Variation <a href="https://arxiv.org/abs/1710.10196">https://arxiv.org/abs/1710.10196</a>
- Takeru Miyato, Toshiki Kataoka, Masanori Koyama, Yuichi Yoshida: Spectral Normalization for Generative Adversarial Networks <u>https://arxiv.org/abs/1802.05957</u>
- Zhiming Zhou, Yuxuan Song, Lantao Yu, Hongwei Wang, Jiadong Liang, Weinan Zhang, Zhihua Zhang, Yong Yu: Understanding the Effectiveness of Lipschitz-Continuity in Generative Adversarial Nets <u>https://arxiv.org/abs/1807.00751</u>
- Andrew Brock, Jeff Donahue, Karen Simonyan: Large Scale GAN Training for High Fidelity Natural Image Synthesis <u>https://arxiv.org/abs/1809.11096</u>
- Tero Karras, Samuli Laine, Timo Aila: A Style-Based Generator Architecture for Generative Adversarial Networks <u>https://arxiv.org/abs/1812.04948</u>

Alternative approaches are also explored: Diederik P. Kingma, Prafulla Dhariwal: **Glow: Generative Flow with Invertible 1x1 Convolutions** <u>https://arxiv.org/abs/1807.03039</u>